Wide Weighted Attention Multi-Scale Network for Accurate MR Image Super-Resolution

Haoqian Wang, Xiaowan Hu*, Xiaole Zhao, and Yulun Zhang

***Abstract*—High-quality magnetic resonance (MR) images afford more detailed information for reliable diagnoses and quantitative image analyses. Given low-resolution (LR) images, the deep convolutional neural network (CNN) has shown its promising ability for image super-resolution (SR). The LR MR images usually share some visual characteristics: structural textures of different sizes, edges with high correlation, and less informative background. However, multi-scale structural features are informative for image reconstruction, while the background is more smooth. Most previous CNN-based SR methods use a single receptive field and equally treat the spatial pixels (including the background). It neglects to sense the entire space and get diversified features from the input, which is critical for high-quality MR image SR. We propose a wide weighted attention multi-scale network (W^2AMSN) for accurate MR image SR to address these problems. On the one hand, the features of varying sizes can be extracted by the wide multi-scale branches. On the other hand, we design a non-reduction attention mechanism to recalibrate feature responses adaptively. Such attention preserves continuous cross-channel interaction and focuses on more informative regions. Meanwhile, the learnable weighted factors fuse extracted features selectively. The encapsulated wide weighted attention multi-scale block (W^2AMSB) is integrated through a recurrent framework and global attention mechanism. Extensive experiments and diversified ablation studies show the effectiveness of our proposed W^2AMSN, which surpasses state-of-the-art methods on most popular MR image SR benchmarks quantitatively and qualitatively. And our method still offers superior accuracy and adaptability on real MR images.**

***Index Terms*—Magnetic Resonance, Super-Resolution, Multi-Scale, Non-reduction Attention Mechanism, Weighted Fusion.**

I. Introduction

Medical imaging, as a medical aid for diagnosis and treatment, has become an indispensable step in current medical detection. As a diagnostic imaging technique, magnetic resonance imaging (MRI) has been applied to various systems of the whole body, such as the brain, soft tissues, and pelvic cavity. Simultaneously, the early detection of lesion structure of MRI can be more effective than computed tomography (CT) [1]. However, due to hardware and physical equipment limitations, magnetic resonance (MR) images often have a low spatial resolution during imaging of the operating organs and lungs, resulting in unclear lesion display. High-resolution (HR) MR images would provide more detailed structures and textures, which benefit the accurate diagnoses and quantitative image analyses. Still, the HR MRI often brings high scanning time cost and low signal-to-noise ratio [2]. Therefore, the popular single image super-resolution (SISR) method of recovering HR image output from Low-Resolution (LR) input has received widespread attention in medical image processing.

Image super-resolution (SR) is a typical ill-posed inverse problem in the field of image processing. The difficulty lies in how to recover local textures and microstructures in an image accurately. Early in natural images, there were some interpolation methods based on bicubic and reconstruction methods such as iterative back projection (IBP) [3] and projections onto convex sets (POCS) [4]. However, these non-learning methods have higher requirements for the prior distribution of the image itself, resulting in limited performance. Traditional learning-based methods such as example learning [5], [6], dictionary learning [7], [8] and other methods are difficult to learn enough information. The disadvantages of these traditional methods are more prominent in the application of medical images. Medical images often contain rich textures and details, and it is difficult to meet the constraints based on constant assumptions. Therefore, the non-data-driven traditional restoration methods have limited performance on medical images.

In recent years, the development of deep learning technology has made convolutional neural network (CNN) mainstream, and it has also brought many advanced methods for SR tasks of natural images [9]–[13]. The success of these methods confirms the powerful ability of the end-to-end learning method in image SR. Therefore, CNN-based methods have also begun to be used for MR images [14]–[16]. Zhang *et al.* achieved better SR performance with the dense residual network (RDN) [17], which further explored deeper networks. Li *et al.* proposed the multi-scale residual network (MSRN) and make full use of features of different scales to solve the problem of loss of detail features in images [18]. The multi-scale dense cross network (MDCN) detects multi-scale features and maximizes feature flow [19]. Yu *et al.* proved that models with wider features before ReLU activation have better performance and applied wide activation operation [20]. Medical images usually do not contain any color information and have low contrast. Although CNN-based methods have also begun to be used for MR images [14]–[16], the current reconstruction results are far inferior to natural images. Some recent innovative works in natural images SR provide new

This work is partially supported by the NSFC fund (61831014), the Guangdong Provincial Science and Technology Project (2017B010110005), the Shenzhen Science and Technology Project under Grant (ZDYBH201900000002, JCYJ20180226181021364).

H. Wang and X. Hu are with the International Graduate School at Shenzhen, Tsinghua University, and Shenzhen Institute of Future Media Technology, Shenzhen 518055, China. E-mail: wanghaoqian@tsinghua.edu.cn, huxw19@mails.tsinghua.edu.cn. (*Corresponding author: Xiaowan Hu)

X. Zhao is with the School of Information Science and Technology, Southwest Jiaotong University (SWJTU), Chengdu, Sichuan 611756, China. E-mail: zxlation@foxmail.com.

Y. Zhang is with the Department of Electrical and Computer Engineering, Northeastern University, Boston, MA 02115, USA. E-mail: yulun100@gmail.com

ideas for reconstruction of medical images.

However, those deep CNN-based methods either neglect the characteristics of MR images or suffer from intrinsic drawbacks of the network, which hinders obtaining excellent SR results on MR images. We briefly summarize the existing problems into the following three points:

Firstly, although a larger receptive field will capture more features, the commonly used methods with a multi-level structure, such as repeated pooling and pyramid down-sampling, will bring about more significant information loss. Only using a single-scale convolution cannot obtain the diversity features about different regions at the same depth of the network. Still, simple concatenation of multi-scale convolution treats each feature channel equally, hindering its representation ability.

Secondly, MR images have specific imaging features, such as diverse structural patterns with rich details and an extensive background region. Although the deep networks bring more effective expression [21]–[24], some low-level structural information tends to gradually disappear as the network deepens, resulting in excessive smoothness in high-frequency areas and increasing training difficulty. However, using residual learning and dense connection repeatedly brings lots of representation redundancy and extra computational costs.

Finally, there is a large background region in MR images, which is far less informative than the target structural areas. Meanwhile, the complexity of tissue texture is highly correlated with its spatial position. If treating all the spatial pixels of the MR image equally, the networks cannot distinguish which part is more important for reconstruction. Besides, over-reliance on convolution operations will focus too much on the local neighborhood and fail to perceive the global receptive field, limiting the representation ability of the network. The previously proposed channel attention mechanism cannot capture the dependencies of all channels. It will destroy channel correlation due to the reduction factor.

To address these problems and limitations, we propose a wide weighted attention multi-scale network (W^2AMSN) for accurate MR SR. We show the main network architecture and internal implementation details in Fig. 1. In summary, our model has three notable contributions:

- Based on the characteristics of degraded MR images, we assign learnable weights to branches of different receptive fields. The adaptive multi-scale features can improve the utilization of low-level and diverse information. Besides, we apply channel widening to increase the number of activated feature maps in the middle layer.
- We propose a multi-level attention mechanism with a non-reduction channel attention block (NCAB) to utilize features effectively. It allows the network to focus on the features interaction in different areas. The mechanism pays more attention to the informative regions (e.g., rich textures and varying brightness) while suppressing useless features (e.g., background and flat structure) adaptively.
- Extensive ablation study is conducted to demonstrate the effectiveness of each component in the newly proposed network. Quantitative and qualitative experimental results show the superiority of W^2AMSN over other advanced CNN-based methods on MR, which achieves the state-of-the-art performance on MR images SR.

The rest of this paper is organized as follows. Section II briefly lists some recent research work and advanced methods related to image SR task. Section III introduces our proposed method and some algorithm implementation details specifically. The experimental results and the analysis of each component are detailed in Section IV. Finally, the prospect of future work and the conclusion of this paper are in Section V and Section VI respectively.

II. Related Work

A. Image Super-Resolution with Deep Learning

With the development of deep learning techniques, some CNN-based methods are used to implement SR [25], [26]. Dong *et al.* firstly proposed super-resolution convolutional neural network (SRCNN) [9] and then accelerated it with fast super-resolution convolutional neural networks (FSRCNN) [10]. More improvements with larger depth and/or width were further explored. The non-local means (NLM) [27] method predicts pixels through global context information. The residual learning is then used to deepen the network and improve performance in the very deep super-resolution (VDSR) [28] network. The RDN [17] combines dense connection and residual learning, making full use of the middle layer features. The enhanced deep super-resolution (EDSR) [13] network removes unnecessary batch normalization layers in the residual structure, which reduces the model size and alleviates feature redundancy. Some networks realize the fusion of multi-scale features by fusing multiple receptive fields [18], [29]. To distinguish the feature importance of different channels, the residual channel attention networks (RCAN) use the attention mechanism to learn cross-channel features adaptively [30]. Hu et al. proposed the channel and spatial feature modulation (CSFM) network to combine and distinguish spatially and channel attention features adaptively [31]. The channel spilt network (CSN) integrates residual learning and dense connections on dual paths, which improves performance and also increases model complexity. The research of deep learning on medical image SR is still in its infancy. Most existing algorithms directly apply them piecewise to medical images through a two-dimensional network [32], [33]. The degraded medical image is not suitable for huge network training, so customizing an efficient network structure according to MR images' characteristics is very important.

B. Multi-Channel Convolution

The proposed deep SR networks [11]–[13], [17] have improved the accuracy of reconstruction compared to the previous shallow networks [9], [10]. The increase in depth brings better performance and brings more parameters and more difficult training, requiring tricks and calculation resources. Skip connections and layer concatenations have been proven to utilize multi-level features more effectively. Zhang et al. proposed a novel residual network of residual networks (RoR) to explore the optimization capabilities of residual networks [34]. The flattened convolution [35] turns the three-dimensional convolution into a continuous sequence to reduce the parameters significantly. And the group convolution [36]

increase the correlation of the filter. Also, MobileNet [37] introduced depthwise separable convolution instead of standard convolution. MobileNetV2 [38] put forward the reverse residual structure and expanded the features before activation for object detection and recognition. Some blocks that can replace the basic components in deep learning networks are also designed for "image-to-image" image restoration tasks [39]. Wide activation deep super-resolution (WDSR) [20] proves that retaining wider features before the ReLU activation layer is beneficial to preserve shallow information. We explore the channel widening operation for accurate MR image SR.

C. Multi-Scale Feature Extraction Block

Some experiments have proved that multi-scale information can effectively improve the performance of many tasks [18], [19], [40], [41]. Segmentation and recognition tasks usually utilize some operations (e.g., pooling layer and dilated convolution) to increase the network receptive field. However, down-sampling often causes information loss in SR tasks. The inception block [40] proposed by Szegedy *et al.* is a multi-size sparse structure to increase the aggregation of information. They increase the width by increasing the number of nodes in each layer and realize the optimum local combination in the convolution network. When objects within a particular scale are over-represented in the detection task, the imbalance on the scale will affect the final detection accuracy. The pyramid methods is proposed to capture the multi-scale features. So the image pyramid [42], [43] and feature pyramid [41], [44] are designed to extract cross-scale interactions globally. A mixture of the two [45] is also proposed to alleviate the imbalance of receptive field. The multi-scale residual network (MSRN) [18] used channel concatenation to fuse the multi-scale features for images SR, which cannot utilize valuable information adaptively. Simultaneously, as the network deepens, the problem of gradually disappearing features has not been solved well.

D. Attention Mechanism in Super-Resolution

The recent development of neural network models shows the importance of capturing the spatial correlation of data, and the feature extraction capabilities can be improved by embedding corresponding learning mechanisms [46]. Hu *et al.* [47] proposed a "squeeze-and-excitation" module to enhance learning ability by modeling channel dependencies. This method uses global pooling (GAP) and channel deformation to generate channel descriptors and recalibrate channel feature maps. Later, a non-local recurrent network (NLRN) [48] added a non-local operation to the neural network to capture the relevance between neighborhoods. Zhang *et al.* [30] combined this channel attention mechanism with the SR model to recover as much informative features as possible, which further improved the discriminative learning ability across feature channels. Dai *et al.* [49] proposed a second-order channel attention module (SOCA) for second-order feature statistics. However, due to channel reduction factors, the serial correlation between channels is destroyed in these attention mechanisms. Combining spatial attention and channel attention leads to a large increase in parameters [31]. The multi-grained attention networks (MGAN) makes full use of multi-scale and attention mechanisms to calculate multi-granular context [50]. MR images have shared visual characteristics. Perceiving both global and local attention mapping is conducive to restoring high-frequency textures and essential edges, which has excellent guiding significance in recovering physiological tissues.

III. Proposed Method

The overall architecture of our proposed W^{2}AMSN is shown in Fig. 1(a). The ultimate goal of the network is to learn an end-to-end mapping from LR to HR, and the main process is similar to other typical SR tasks. It consists of the following three parts: shallow feature extraction (SFE), wide weighted attention multi-scale feature fusion, and image reconstruction. First, the shallow features of the LR image are extracted through simple convolution layers and then output to the subsequent stacked wide weighted attention multi-scale blocks (W^{2}AMSB). In the image reconstruction stage, the wide weighted attention multi-scale features after the global attention layer are used to predict the final SR image. As shown in Fig. 1(b), the process inside the W^{2}AMSB is mainly divided into channel widening, attention multi-scale feature extraction, and weighted branch fusion. Besides, for the model to learn the shared information between the original LR and SR image, we introduced local skip connection (LSC), and global residual connection (GRC) [28].

A. Overall Network Architecture

1) Shallow Feature Extraction: As shown in Fig. 1(a), the SFE includes three convolution layers, the size of the convolution kernel is 3×3, 1×1 and 3×3 respectively. Define $\mathscr{F}_S(\cdot)$ as the corresponding fuction of SFE, the shallow features x_S extracted in the first stage can be expressed as:

$$\mathbf{x}_S = \mathscr{F}_S(\mathbf{x}), \tag{1}$$

where x represents the input LR image.

2) Wide Weighted Attention Multi-scale Feature Fusion: The wide weighted attention multi-scale feature fusion part consists of stacked W^{2}AMSBs and the concatenation layer. We define the function corresponding to the W^{2}AMSB as $\mathscr{F}_M(\cdot)$. Assume the number of W^{2}AMSB in the entire network is t, then the output of the i-th block of the network is

$$\mathbf{x}_i = \mathscr{F}_w^i(\mathbf{x}_{i-1}),\ i = 1,2,...,t, \tag{2}$$

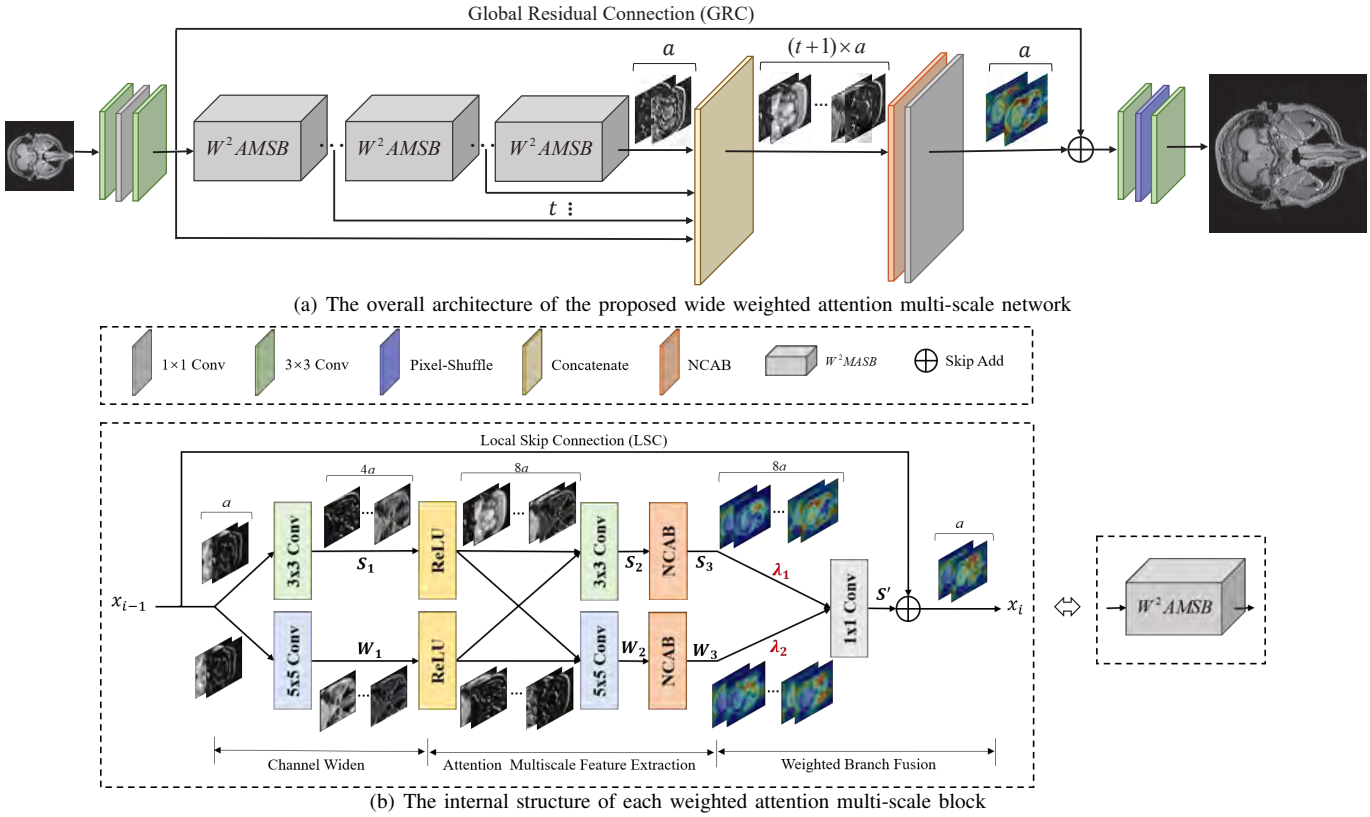
where $\mathscr{F}_w^i(\cdot)$ corresponds to the i-th W^{2}AMSB. The implementation details of $\mathscr{F}_w^i(\cdot)$ will be explained in Section III-B, III-C and III-D. The input of the first W^{2}AMSB is the output of the SFE, ie $\mathrm{x}_0 = \mathrm{x}_S$. This process is executed iteratively, and the output of the last W^{2}AMSB is as follows:

$$\mathbf{x}_t = \mathscr{F}_w^t(\mathbf{x}_{t-1}) = \mathscr{F}_w^t(\mathscr{F}_w^{t-1}(\cdots(\mathscr{F}_w^1(\mathbf{x}_0))\cdots)). \tag{3}$$

It should be noted that each output of W^{2}AMSB is directly served as the input for the next block without any transformation, which is also used as a component of the fused multi-scale features, defined as x_c. This combination helps the flow of information in the network, which can be denoted as:

$$\mathbf{x}_c = [\mathbf{x}_0, \mathbf{x}_1, ..., \mathbf{x}_t], \tag{4}$$

where $[...]$ means the concatenation.



(a) The overall architecture of the proposed wide weighted attention multi-scale network

(b) The internal structure of each weighted attention multi-scale block

Fig. 1. The architecture of the proposed W^{2}AMSN and the internal implementation details of the W^{2}AMSB. (a) The overall structure of the proposed network, which consists of convolution layers, stacked W^{2}AMSBs, Concatenate, and Attention layer. Notice that GRC and LSC are used to improve network information flow and stabilize training. (b) The W^{2}AMSB is composed of three parts: channel widening, multi-scale attention feature extraction, and weighted branch fusion. Images on the directed line visualize the feature maps in some intermediate layers, where a, $4a$, and $8a$ represent the number of channels.

3) Image Reconstruction: In the image reconstruction part, the process is divided into global attention remapping and upsampling reconstruction. Firstly, the fused features are remapped through the global attention layer to capture the dependencies of all channels and retain features selectively. The internal implementation details are shown in Fig. 2. Specifically, the features respectively input the 1×1 convolution layer through the upper and lower branches. Then, in the upper branch, the global pooling layer and the softmax function are used to obtain channel attention weights.

It should be pointed out that we use non-reduction channel attention block (NCAB) to ensure the integrity of the channel correlation. The dimension of the attention weight w_a is consistent with the number of channels of the input feature map. Subsequently, the output weights are aggregated with another branch, and residual connections are also introduced at the end of the layer to learn attention mapping better. The function is defined as $\mathscr{F}_A(\cdot)$, and the output is expressed as:

$$w_a = Softmax[G(C_1^u(\mathbf{x}_c))], \tag{5}$$

$$\mathscr{F}_A(\mathbf{x}) = w_a \otimes C_1^l(\mathbf{x}) + \mathbf{x}, \tag{6}$$

$$\mathbf{x}_a = \mathscr{F}_A(\mathbf{x}_c), \tag{7}$$

where x_a is the fusion features after the global attention layer, which will be used as input for the SR image prediction. The $G(\cdot)$ denotes the global average pooling. The $C_1^l(\cdot)$ and $C_1^u(\cdot)$ are 1×1 convolution layers with channel number C, and $\otimes$ means the Hadamard product. Normalization of the attention weights will weaken the output response, so we use element-wise add to combine the obtained attention feature map with the input feature x_c. Compared with the discrete Sigmoid function, the Softmax function is a bit more "soft" and retains more dense related information between channels.

The original LR image is undoubtedly highly similar to the HR image that needs to be restored, which indicates that the two have a lot of shared information. We introduced the GRC as a shortcut map to learn the residual information between the original input x_S and output x_a. Here we use a 1×1 convolutional layer defined as C_1^g to adjust the number of channels of the attention feature x_a. The upsampling reconstruction consists of two 3×3 convolution layers and a pixel-shuffle layer, being defined as function $\mathscr{F}_R(\cdot)$. Finally, the reconstructed image can be obtained as follows:

$$y = \mathscr{F}_R(C_1^g(\mathbf{x}_a) + \mathbf{x}_S). \tag{8}$$

We did not upsample the input LR image to the same size as HR in advance but used the pixel-shuffle [51] to reconstruct SR images at the end of the network. This module is not limited to exponential magnification based on two. It can be converted to any upscaling factor, which only needs to adjust the structure to achieve a slow increase in resolution.

B. Channel Widening

Our proposed W^{2}AMSB consists of three parts, and the first part is the channel widening. For MR images, shallow low-level visual information is critical for more accurate pixel-level prediction. If an extended activation is performed before the ReLU layer, more knowledge can pass to the subsequent

TABLE I
PERFORMANCE COMPARISON OF BLOCKS REPRESENTING DIFFERENT DOMAINS IN THE ORTHOGONAL ATTENTION MODULE.

Method	Multi-scale	Channel Widening	PSNR/SSIM
Baseline			40.29/0.9765
+CW		✓	40.57/0.9831
+MS	✓		41.21/0.9864
+MS/CW	✓	✓	**41.59/0.9901**

network for dense pixel value prediction. To widen the feature channel without increasing the number of parameters too much, we use a multi-channel convolution layer.

As shown in Fig. 1(b), the input feature tensor of the i-th W^{2}AMSB is represented as x_{i-1}, and we suppose the number of channels is a. The feature maps are passed to the upper and lower branches with the convolution kernel size 3 and 5, respectively. We expand the feature map channel by increasing the number of convolution kernels by four times, which can achieve wider activation to keep more low-level features and structural information. The process can be expressed as:

$$\begin{aligned} S_1 &= w_3^{4a} \times \mathbf{x}_{i-1} + b_1, \\ W_1 &= w_5^{4a} \times \mathbf{x}_{i-1} + b_1, \end{aligned} \tag{9}$$

where w_3^{4a} and w_5^{4a} represent the weights for different convolution kernels while the superscripts $4a$ represent the number of output channels, and the b_1 illustrates bias of the current layer. In Section IV-D, we specifically analyzed the impact of channel widening on network performance. As can be seen from Tab. I, more reservation and forward transmission of low-level features of MR images after channel widening can effectively improve the accuracy of image SR.

C. Attention Multi-scale Feature Extraction

Recent work proved that blindly pursuing the depth of the network cannot effectively improve the network performance, so some networks with the structure of extracting multi-scale features are designed to explore information from the spatial domain further. Inspired by this, for MR images, we propose a method to extract the multi-scale features of attention interest, which can capture rich details of different sizes adaptively and efficiently focus on more informative regions.

For the wide activation features extracted from branches at various scales, we first use the ReLU layer for nonlinear activation. The outputs from two branches are concatenated for information interaction, which is beneficial to capture the contextual correlation of multi-scale features. The $\sigma(\cdot)$ stands for the ReLU function. As shown in Fig. 1(b), the number of feature channels after branch aggregation changes from $4a$ to $8a$, and the intermediate features transmission is performed through convolution layers with respective sizes. The outputs are defined as S_2 and W_2, which can be formulated as:

$$\begin{aligned} S_2 &= w_3^{8a} \times [\sigma(S_1), \sigma(W_1)] + b_2, \\ W_2 &= w_5^{8a} \times [\sigma(W_1), \sigma(S_1)] + b_2, \end{aligned} \tag{10}$$

where w_3^{8a} and w_5^{4a} represent the weights of convolution layers while the superscripts $8a$ represent the number of output channels, and the b_2 depicts the bias of the current layer. As shown in Tab. I, using multi-scale interactive convolution features can effectively improve network performance.

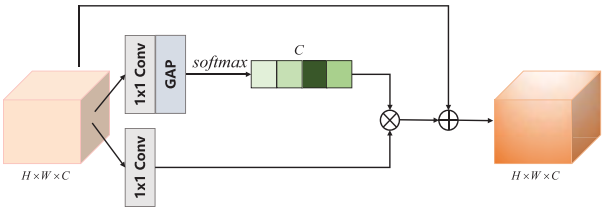


Fig. 2. The internal implementation details of NCAB in Fig. 1(b). The block has no channel reduction operation. And the dimension of attention vector is C, which is the same as the number of channels of the input feature map.

Besides, both S_2 and W_2 contain features of different sizes and regions extracted by the corresponding branch, which include some high-frequency textures and edges. The function $\mathscr{F}_A(\cdot)$ correspond to the attention layer shown in the Fig. 2. Non-reduction attention maintains continuous correlation between channels and focuses on more significant details while suppressing some low-frequency and useless information, and the output can be defined as:

$$\begin{aligned} S_3 &= \mathscr{F}_A(S_2), \\ W_3 &= \mathscr{F}_A(W_2). \end{aligned} \tag{11}$$

D. Weighted Branch Fusion

Skipping connections between residual layers will cause representation redundancy. The fusion method selects the activated features adaptively according to the importance of each branch. In the i-th W^{2}AMSB, we introduced two learnable parameters λ_1^i and λ_2^i to recalibrate the response of each branch. Meanwhile, we use the 1×1 convolution layer followed by the weighted branch concatenation to integrate features and channel reduction, which facilitates adapting to iterative learning and saves computation cost greatly. The channels of the input S_3 and W_3 is $8a$, and the output after weighted branch fusion can be expressed as:

$$S' = w_1^a \times \left[\lambda_1^i S_3, \lambda_2^i W_3\right] + b_3. \tag{12}$$

The weights of 1×1 convolution are denoted as w_1^a while the superscripts a represent the number of output channels. The b_3 represents the bias of the current layer. In Section IV-D2, we compared the performance of different scale branches while also proving the superiority of weighted fusion in fully exploiting model capabilities additionally.

The cross-layer shortcut connection can not only promote the information integration between layers, but also stabilize the training. Therefore, in order to prevent network degradation and retain original information flow effectively, we adopt locak skip connection (LSC) to each W^{2}AMSB as below:

$$\mathrm{x}_i = S' + \mathbf{x}_{i-1}, \tag{13}$$

where x_i and x_{i-1} represent the input and output of the i-th W^{2}AMSB. The shortcut connection here can be considered as a local residual learning realized by the element-wise addition.

It is worth mentioning that LSC and the GRC in the overall network constitute a multi-level residual mechanism [52], which is proved to stabilize training and improve model performance. The cross-layer connection between farther apart layers is more conducive to retaining prior information in LR images. Experiments have shown that local skip connection can alleviate the model training instability problem caused by the degradation of MR training samples.

TABLE II
MODEL SIZE COMPARISON ON T2 FOR SR ×2.

Methods	Parameters	Mult-Adds	PSNR/SSIM
SRCNN [9]	222.79K	3.26G	37.12/0.9761
VDSR [28]	1.40M	20.19G	37.67/0.9783
RDN [17]	22.12M	318.43G	37.95/0.9795
EDSR [13]	3.14M	722.61G	37.56/0.9774
RCAN [30]	15.44M	220.34G	37.88/0.9793
CSFM [31]	24.87M	397.64G	38.06/0.9802
W^{2}AMSN-S	5.93M	85.33G	<u>38.31/0.9817</u>
W^{2}AMSN	11.67M	167.91G	**38.67/0.9823**

E. Training Objective

The goal of our proposed W^{2}AMSN is to learn the end-to-end mapping function $\mathscr{F}_{W^2AMSN}$ between the LR image and the HR image. The model parameters are determined by minimizing the loss between the reconstructed image and the ground truth. Given a training dataset $\{I_i^{LR}, I_i^{HR}\}_i^N$, where N is the total number of the training set. The most widely used loss functions in SR are aiming to minimize mean-square error (MSE) and L2 loss. Although these methods can maximize the peak signal-to-noise ratio (PSNR), they tend to make the high-frequency details in images smooth excessively, which will seriously damage the visual reconstruction result. Many superior loss functions have been proposed in the super-resolution work of natural images to improve network performance. However, for medical images, perceptual loss [53] and the generative adversarial learning [54] will bring distortions in texture and structure, which poses risks for the subsequent precise diagnosis and analysis. To make a fair comparison with existing methods, we finally chose the L_1 loss function to guide model optimization:

$$L(\theta) = \frac{1}{|N|}\sum_{i=1}^{|N|}\left\|I_i^{HR} - \mathscr{F}_{\mathrm{W^2AMSN}}\left(I_i^{LR};\theta\right)\right\|_1, \qquad (14)$$

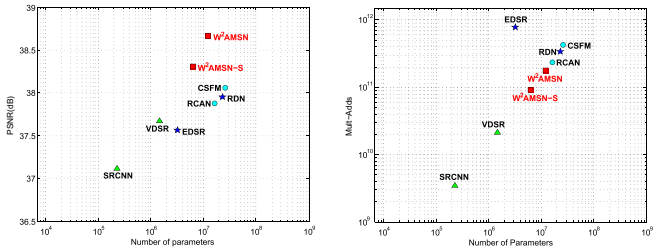
where θ denotes the parameters of our model, I_i^{HR} is the ground truth corresponding to I_i^{LR}. When using degraded MR training samples, the training process limited by L_1 loss converges faster and more stable in the experiment.

IV. EXPERIMENT

In this section, we first briefly introduce the datasets used in this work and the model implementation details. Then several state-of-the-art SR methods are compared with the proposed model. Next, we perform a series of ablation experiments to investigate and analyze the structure of our W^{2}AMSN model. We adopt PSNR, and structural similarity index metric (SSIM) [33] as the quantitative evaluation metrics.

A. Datesets

The datasets used in this paper are the same as and derived initially from the IXI dataset download from http://brain-development.org/ixi-dataset/, which consists of three types of MR images (578 PD volumes, 581 T1 volumes, and 578 T2 volumes). We divide the dataset into three parts in proportion, each of which has 500, 70, and 6 MR volumes for model training, testing, and quick validation, respectively. The size of each 3D volume is cut to $240\times240\times96$ (height$\times$ width$\times$ depth) for three different upscaling factors ($\times2$, $\times3$ and $\times4$), where 96 indicates the number of slices in the MR volume. Note that the datasets used in the experiment contain two kinds of image degradation, but only the typical bicubic degradation is studied detailedly in this paper due to limited space.



(a) Model size. (b) Model comprehensive analysis.

Fig. 3. Model analysis and comparison between several leading networks on T2 dataset for SR ×2. (a) Model performance and size of different models. (b) The comprehensive analysis of the model size and interface speed.

B. Implementation Details

We specify the implementation details of our proposed W^{2}AMSN.The configuration of our final model is shown in Fig. 1, we insert 20 Wide Weighted Attention Multi-scale Block (W^{2}AMSB, $t=20$). and the number of input channels of each W2AMSB is 32 (a=32). Therefore, the number of output feature channels of the middle layer inside the block will also be widened (4a=128, 8a=256) correspondingly. For convolution layers with kernel size 3×3 and 5×5, the zero-padding strategy is used to keep the size fixed. We initialize the learnable weight coefficients of multi-scale branches to $\{\lambda_1=1,\lambda_2=1\}$. In each training batch, 96 LR patches with a size of 32×32 are extracted as inputs. Our model is trained by ADAM optimizer [9] with $\beta_1=0.9$, $\beta_2=0.999$, and $\varepsilon=10^{-8}$. The learning rate is initialized as 10^{-4} and then decreases to half every 200 epochs. We implement all models with the Pytorch framework and train them on NVIDIA GeForce GTX 1080 Ti GPU for a thousand epochs.

C. Comparison with Other Methods

In this subsection, for the purpose of illustrating the effectiveness of the proposed W^{2}AMSN on MR image SR tasks,we compare the proposed method with several advanced techniques, inlcuding NLM [27], SRCNN [9], VDSR [28], RDN [17], EDSR [13], RCAN [30] and CSN [52].

1) Quantitative Comparison: As shown in Tab. III, we exhibit the quantitative results of the compared methods on the IXI dataset. W^{2}AMSN represents the results directly obtained by our model, and W^{2}AMSN+ indicates that geometric self-ensemble [13] is applied. It can be seen that the proposed W^{2}AMSN model outperforms other state-of-the-art methods by a large margin, giving the best SR performance on all types of MR images (PD, T1, and T2) and all SR scaling factors ($\times2$, $\times3$, and $\times4$), even without geometric self-ensemble.

Specifically, we analyze the size and inference speed of the model and compared it with other advanced image SR methods on T1 dataset for SR ×2. The quantitative comparison results are reported in Tab. II. Parameters represent the number of parameters in the model. The parameters determine the size of the model and also affect the memory usage during model inference. Mult-Adds represents two arithmetic operations in model calculations, including multiplication and addition operations. It is a quantitative metric for evaluating model calculation costs and inference speed. Different attention mechanisms focus on more informative areas in different ways. Channel attention focuses on "what" is a meaningful input image, while spatial awareness focuses on "where" is the most informative

TABLE III
QUANTITATIVE COMPARISON BETWEEN DIFFERENT SR METHODS. THE MAXIMAL PSNR (DB) AND SSIM VALUES OF EACH COMPARISON CELL ARE MARKED IN **BOLD**, AND THE SECOND ONES ARE MARKED IN *underline* (PSNR / SSIM).

Data	Scale	Bicubic	NLM [27]	SRCNN [9]	VDSR [28]	RDN [17]	EDSR [13]	RCAN [30]	CSN [52]	W^{2}AMSN	W^{2}AMSN+
PD	×2	35.04/0.9664	37.26/0.9773	38.96/0.9861	39.97/0.9861	40.31/0.9870	39.87/0.9857	40.57/0.9871	41.28/0.9895	41.59/0.9901	**41.72/0.9903**
	×3	31.20/0.9230	32.81/0.9436	33.60/0.9516	34.66/0.9599	35.08/0.9628	34.39/0.9678	35.06/0.9682	35.87/0.9693	36.22/0.9717	**36.42/0.9726**
	×4	29.13/0.8799	30.27/0.9044	31.10/0.9181	32.09/0.9311	32.73/0.9387	31.80/0.9284	32.58/0.9367	33.40/0.9486	33.72/0.9524	**33.98/0.9543**
T1	×2	33.80/0.9525	35.80/0.9685	37.12/0.9761	37.67/0.9783	37.95/0.9795	37.56/0.9774	37.88/0.9793	38.27/0.9810	38.67/0.9823	**38.75/0.9825**
	×3	30.15/0.8900	31.74/0.9216	32.17/0.9276	32.91/0.9378	33.31/0.9430	32.76/0.9347	33.18/0.9396	33.53/0.9464	34.00/0.9509	**34.12/0.9518**
	×4	28.28/0.8312	29.31/0.8655	29.90/0.8796	30.57/0.8932	31.05/0.9042	30.46/0.8902	30.85/0.8965	31.23/0.9093	31.84/0.9196	**31.99/0.9215**
T2	×2	33.44/0.9589	35.58/0.9722	37.32/0.9796	38.65/0.9836	38.75/0.9838	38.28/0.9824	39.24/0.9841	39.71/0.9863	40.22/0.9873	**40.35/0.9875**
	×3	29.80/0.9093	31.28/0.9330	32.20/0.9440	33.47/0.9559	33.91/0.9591	33.15/0.9528	33.85/0.9574	34.64/0.9647	35.02/0.9672	**35.21/0.9682**
	×4	27.86/0.8611	28.85/0.8875	29.69/0.9052	30.79/0.9240	31.45/0.9324	30.52/0.9198	31.60/0.9298	32.05/0.9413	32.37/0.9453	**32.62/0.9473**

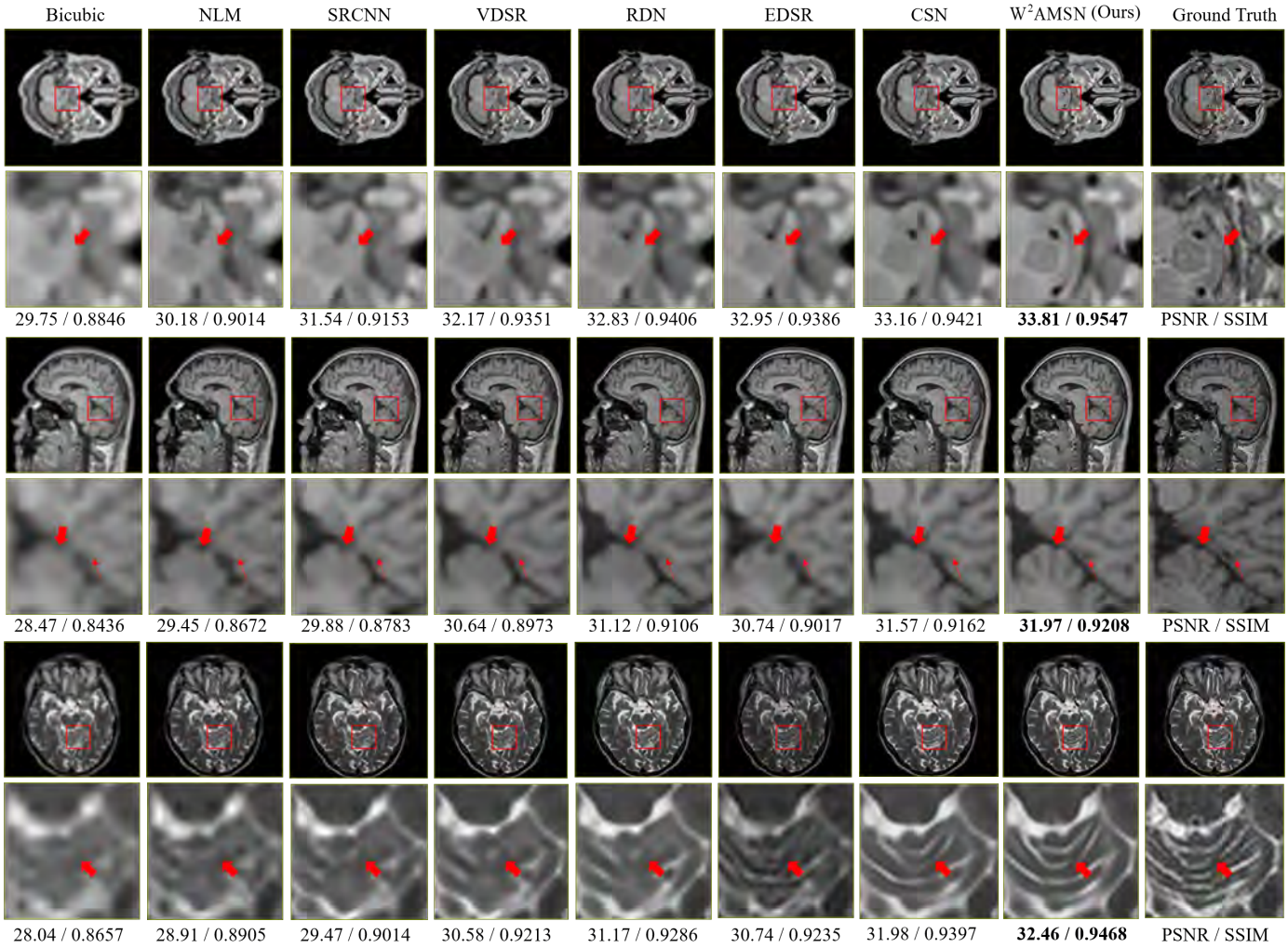
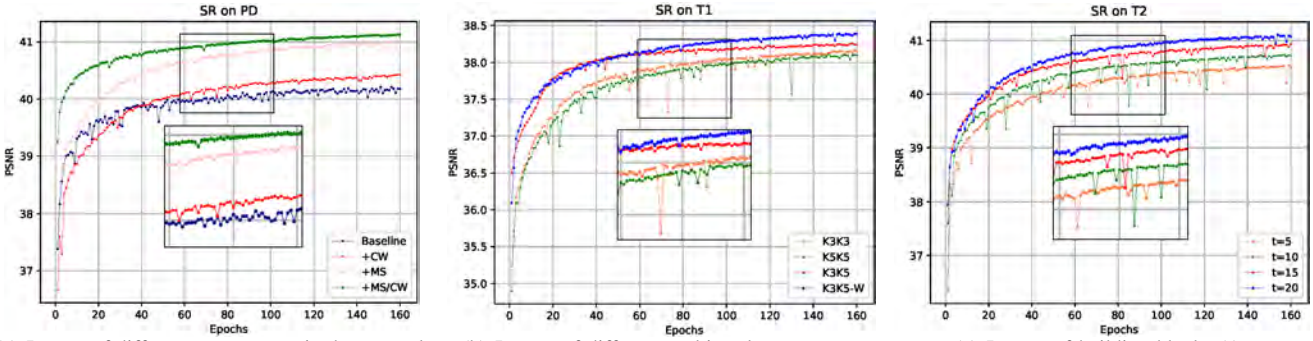


Fig. 4. The visual comparison between several advanced CNN-based SISR methods on a PD image (top) , T1 image (middle), and T2 image (bottom) with scaling factor SR×4. The maximal PSNR (dB) and SSIM values for each displayed image are marked in bold.

part. RCAN [30] is a network that uses channel-wise attention and CSFM [31] is a network that uses spatial attention. It can be seen that the spatial attention used in CSFM calculates the global correlation, which will significantly increase the model complexity and parameters. Our method utilizes moderate-scale parameters and supports fast model inference. To compare with the lightweight model, we built a small W^{2}AMSN-S model with only 8 W^{2}AMSBs, and the simplified model achieves comparable SR performance.

As shown in Fig. 3, we compare the performance, size, and interface speed of different models comprehensively to evaluate the model efficiency. In Fig. 3(a), the horizontal axis is the number of parameters, and the vertical axis is the PSNR value. The W^{2}AMSN model we proposed achieves the best performance, and the lightweight W^{2}AMSN-S model with fewer parameters also achieves significant PSNR gains. We consider both model size and interface speed to evaluate model efficiency in Fig. 3(b). It can be seen that our W^{2}AMSN has a lower parameter amount and a higher inference speed. Our method has a better trade-off between model complexity and performance while ensuring the highest SR performance, which indicates that our model is not only a high-precision MR image SR method but also a highly practical one.

2) Visual Comparison: Fig. 4 shows the visual result of the compared methods listed in Tab. III, on the PD (top), T1 (middle), and T2 (bottom) images with SR×4 respectively. As can be seen, the proposed W^{2}AMSN model displays significantly visible superiority over other methods. For ex-

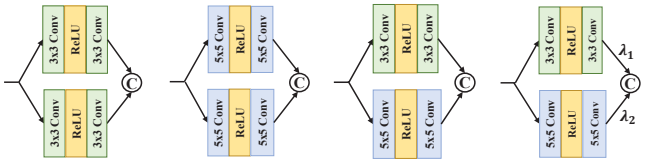


(a) Impact of different components in the network (b) Impact of different multi-scale structures (c) Impact of building blocks (t)

Fig. 5. Validation PSNR curves for different ablation experiments. (a) The comparison between the impact of using only one or both of MS and CW on network performance carried out on PD for SR×2. (b) The performance comparison between the different multi-scale branch structures shown in Fig. 6 on T1 for SR ×2. (c) The performance comparison between the models with different number of W^{2}AMSB on T2 for SR ×2.

TABLE IV
QUANTITATIVE RESULTS OF ABLATION STUDY ON DIFFERENT COMBINATIONS OF THE NUMBER OF CHANNELS ON PD FOR SR ×2.

CW	Scale Factor	Input-Middle-Output	PSNR/SSIM
0	1	32-32-32	38.74/0.9783
0	1	64-64-64	39.19/0.9854
1	2	32-64-32	40.28/0.9883
1	2	64-128-64	41.53/0.9895
1	4	32-128-32	41.59/0.9901
1	4	64-256-64	**41.63/0.9902**



(a) K3K3 (b) K5K5 (c) K3K5 (d) K3K5-W

Fig. 6. Different structures of multi-scale convolution combinations for feature extraction.(a) K3K3 without weighted fusion (b) K5K5 without weighted fusion. (c) K3K5 without weighted fusion. (d) K3K5 with weighted fusion.

ample, in the PD image, there is a black line indicated by the red arrow. This structure is almost completely lost in the results of Bicubic, NLM [27], SRCNN [9], VDSR [28], and even RDN [17]. Although it can be observed in the results of EDSR [13] and CSN [52], our model presents a clearer indication and better approximation to the ground truth. A similar comparison can also be seen from the results of the T2 images. In the area marked by the red arrow, the result of CSN is slightly better than other methods that do not reconstruct the texture. However, it should be noted that in the ground truth, the two thin white lines are disconnected, but the reconstruction of CSN connects the two lines smoothly, which is a wrong distortion. In the same position, we can see that our method reconstructs the truncated texture well, which illustrates the superiority of the proposed W^{2}AMSN in restoring important structural details in MR images.

D. Ablation Study

1) Channel Widening and Multi-scale Branch: To verify the impact of multi-scale (MS) and channel widening (CW) mechanisms on network performance, we built a "baseline", in which all CW convolutions and MS branches are replaced by unexpanded filters and single-scale branch, respectively. We conduct the ablation test using only one or both of MS and CW carried out on PD for SR×2 specifically. The notations and quantitative results are shown in Fig. 5(a) and Tab. I. It can be seen that the simple addition of CW operation has a slight improvement in the SR performance, which proves that the wide activation can retain more low-level features in MR images. The MS mechanism boosts model performance effectively, achieving a performance of 0.9dB compared with baseline. Using CM and MS simultaneously can contribute most (1.3dB) to the improvement of PSNR/SSIM values and make the model training more stable.

As shown in Tab. IV, we analyzed the effects of the different settings of the number of channels on the PSNR/SSIM quantitatively. In the first column of the Tab. IV, "0" represents that no channel widening operation is performed. In contrast, "1" represents the widening of the output channel in the middle layer, and we set different channel expansion scale factors. From the reconstruction evaluation results, it can be seen that the network performance increases as the number of input channels increases. It is a common observation where 64 channels would achieve better performance than 32 channels in image SR. However, even if the number of input channels is 32, we can obtain higher performance by doubling the middle layer channels to 64 and reducing them to 32. Such an observation proves that the channel widening and compression is more effective than directly increasing the channel number. Also, for different inputs, the wider the input feature map, the greater the expansion factor, and the better the network performance. More importantly, the channel reduction before the output can retain the advantage of wide activation without adding too many parameters. Although the best SR performance is achieved when the number of input channels is 64 and the expansion factor is 4, the final network still chooses the setting rule of 32-128-32, considering the trade-off of parameters and training speed.

2) Weighted Multi-Scale Convolution: Our proposed MFF can be configured with a different branch structure. To future explore the capability of the module specifically, we design four different combinations of convolution receptive field and fusion methods (see Fig. 6). It should be noted that only the modified convolution layer and other necessary connections are plotted for simplified representation. The K3K3 in Fig. 6(a) represent the structure that the upper and lower branches both have 3×3 convolution kernels while the K5K5 structure in Fig. 6(b) change the size from 3 to 5. The K3K5 structure in Fig. 6(c) represents the upper and lower branches that have different convolution receptive fields. Besides, for investigating

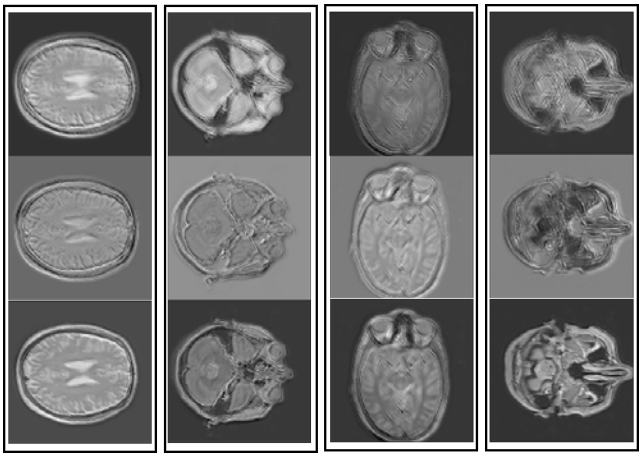


Fig. 7. The visual comparison between output feature maps of convolution layers at different locations on 3×3 convolution layer (top), 5×5 convolution layer (middle), and multi-scale feature fusion layer (bottom).

TABLE V
PERFORMANCE COMPARISON OF DIFFERENT ATTENTION MECHANISMS.

Method	Baseline	+ CAB	+ NCAB
PSNR/SSIM	31.34/0.9294	32.05/0.9374	**32.37/0.9453**

the weighted fusion impact, we add weighted fusion operation to K3K5 shown in Fig. 6(d), which is denoted as K3K5-W. The performance of these four compared structures on the validation set (T1, ×2) is shown in Fig. 5(b). We can observe that K3K5 has significantly better performance than K3K3 and K5K5, which shows the superiority of the mixed scales. The *blue* curve (K3K5-W) is the result after adding the learnable weight factors, and it achieves a gain of about 0.15dB compared with the unweighted *red* curve (K3K5). Such comparisons indicate that weighted fusion is indeed conducive to boost performance.

To better explain the superiority of mixed-scale convolution, we analyze the middle layer of the first W^{2}AMSB module and visualize the feature maps at three locations, including two sizes of convolution layers and the concatenation layer after fusion. As shown in the Fig. 7, it can be seen that the layer with a convolution kernel size of 3 pays more attention to extracting local details while the layer of 5 has a better response to structural textures. Furthermore, the integrated global features combine the advantages of the two scales. Consequently, the edges and detail textures are more precise.

3) Network Depth: It is reasonable that more parameters will bring better networkability of model fitting, but it also means that the network training will require more resources and a longer time. In this part, we analyze the model performance with different values of t (the number of W^{2}AMSB), which is the main parameter that determines the depth of the overall network. Also, there are five fixed convolution layers at the beginning and end of the proposed network. Thus, the depth of the overall W^{2}AMSN is given by:

$$D = 5t + u + 5, \tag{15}$$

where u is the depth of the pixel-shuffle layer, which depends on the value of upscaling factor [13]. The ablation experiment is based on the T2 image dataset with an upscaling factor of 2 ($u = 1$). As shown in Fig. 5(c), when $t = 5$, the performance of the network has exceeded the baseline. The small network can be used to achieve a better trade-off between model size and performance. The PSNR increases as t increases. But the convergence speed will not slow down as the network deepens, and the training process is relatively stable. It indicates that the idea of multi-scale feature weighted fusion helps to offset the limited reception field of a relatively small network.

TABLE VI
PERFORMANCE COMPARISON OF ATTENTION COMBINATIONS AT DIFFERENT LOCATIONS IN THE NETWORK.

Method	Block Attention	Network Attention	PSNR/SSIM
A			31.34/0.9294
B	✓		32.12/0.9416
C		✓	31.98/0.9368
D (Ours)	✓	✓	**32.37/0.9453**

4) Attention Mechanism: As shown in Fig. 1, we perform the attention layer to two positions of the network, one is at the multi-scale attention feature extraction stage of each W^{2}AMSB, and the other is at the end of the overall W^{2}AMSN to guide the image reconstruction. We compare the performance between the proposed NCAB and the channel attention block (CAB) proposed in RCAN. The baseline module is built by removing the attention layer. The results in Tab. V show the advantages of NCA, which has achieved a higher improvement (1dB) compared to the baseline.

We train the model that combines different attentions on the T2 dataset with SR ×4. The performance comparison results after 400 iterations are reported in Tab. VI. The block and network attention represent the attention layer at each W^{2}AMSB and the end of the network, respectively. Method A is the baseline without the attention mechanism. The feedforward network could less recognize these informative patterns and textures without any attention block. The results in the first three rows prove the effectiveness of the non-reduction attention mechanism in both block and network, as each of them brings improvement over the backbone. Obviously, the network with multi-level attention is able to obtain the best performance and increase the highest PSNR from 31.34 dB to 32.37 dB, thus achieving a significant improvement.

5) Performance on in-vivo Images: We also perform SR experiments on two in-vivo T1 images collected from Alltech Medical Systems Co., LTD. These images were from real MRI scanners without any additional preprocessing. In this case, the ground truth HR image is not available, and the image degradation is also unknown. We visually compare the reconstructed T1 images processed by different models with SR ×4, including NLM [27], SRCNN [9], VDSR [28], RDN [17], EDSR [13], and CSN [52]. As shown in Fig.8, our W^{2}AMSN model can recover more sharp edges and finer details in real LR MR images compared with other state-of-the-art methods. Clearer and richer structural textures are beneficial to supplement image information, which is of great significance for MR imaging in practical clinical applications.

6) Attention Visualization: The class activation map (CAM) [55] is originally used for image classification, representing the weighted linear sum of the existence of visual patterns in different spatial positions. The CAMs help show which areas the model focuses on by mapping the response of the feature map to the original image. The feature maps from the last convolutional layer in the SR network are globally

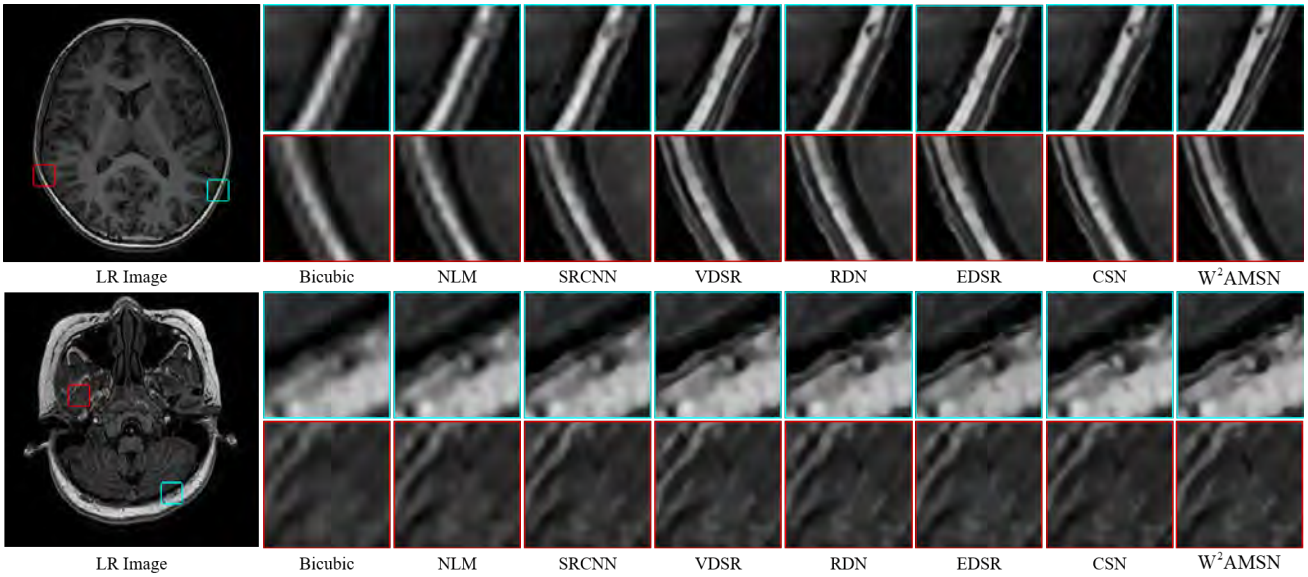
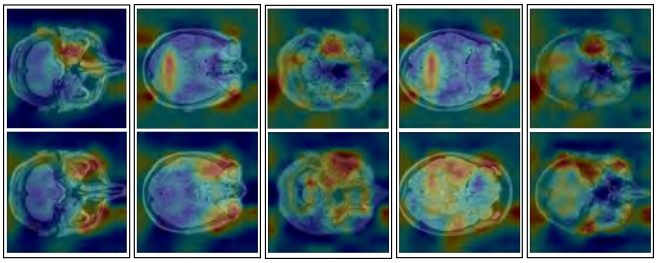
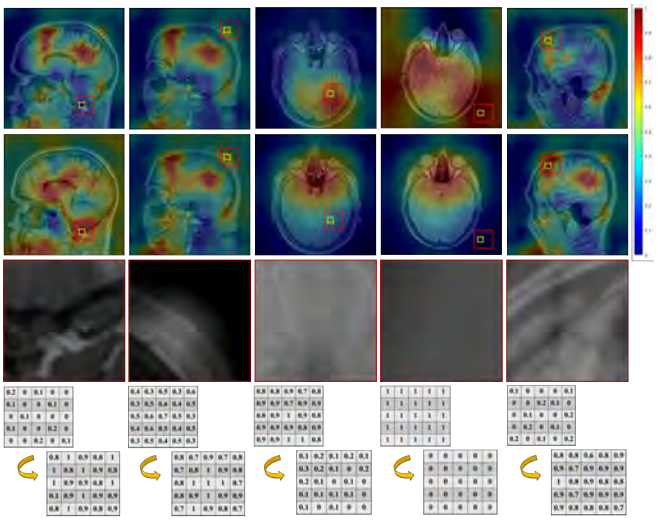


Fig. 8. The visual comparison on the SR performance between several state-of-the-art SR methods on in-vivo T1 MR images with SR $\times 4$. In this case, the ground truth HR images are not available. We cropped and enlarged the blue and red boxes in the original LR images to get a clearer visual experience.



(a) The CAM visualization before and after the 8-th W^{2}AMSB.



(b) The CAM visualization before and after the global attention.

Fig. 9. The visualization of CAM before and after the attention layer in two different locations. Noticed that the significant discriminative regions are marked as highlighted, where red means significant, orange, green, and blue indicate that the importance gradually decreases in order.

pooled to obtain the corresponding weight and remapped to the original feature map. To get a deeper understanding of the importance of modeling channel correlation in MR images, we visualized the CAM diagrams before and after the attention layers in two different locations, shown in Fig. 9. Besides, we add detailed comparison and quantitative analysis of local activation regions in Fig. 9(b). CAM shows the degree of saliency of the area in a special highlight form. Red regions

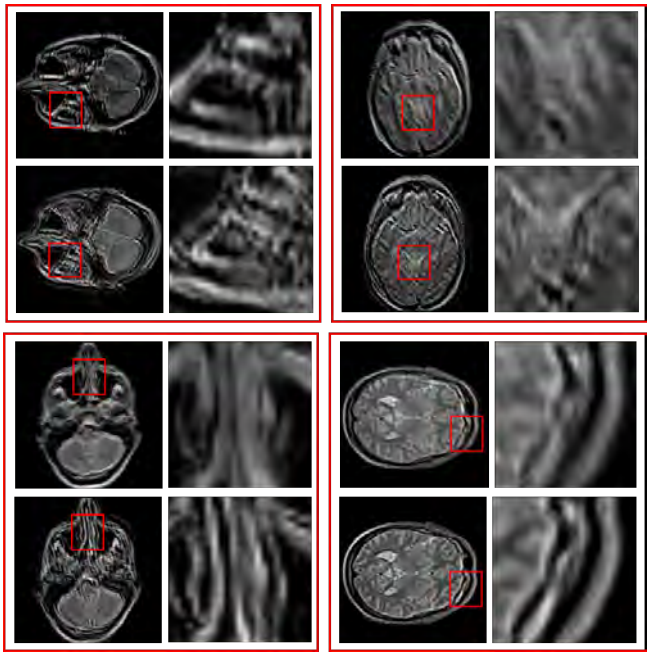


Fig. 10. Visual comparison on the SR performance of whether branch weighting factors are added. Note that the two rows in each of the four red bounding boxes represent different results of the same image (PD $\times$ 2), which use fixed coefficients (top) or learnable weighting factors $\{\lambda_1, \lambda_2\}$ (bottom).

mean very important, and orange, green, and blue indicate that the importance gradually decreases in order.

Firstly, we can see from the Fig. 9(a) that the feature response after the attention layer inside the W^{2}AMSB has extracted more repeating structural patterns and detailed high-frequency textures. It is worth mentioning that the attention layer can explicitly model the connection of contextual information. So it can capture a broader range of useful features and suppress less useful ones. The same is the attention layer at the end of the W^{2}AMSN. In Fig. 9(b), we enlarge and analyze the areas where the activation weight changes significantly. We visualize the area within the red bounding box in the original image to reflect local features. Then, we select the room with a size of 5×5 (marked with the yellow bounding

box) and analyze the activation weight quantitatively. The results show that for the richly structured position, the red highlight information area is expanded. On the contrary, the importance of some less relevant regions is reduced. This observation indicates that the global attention mechanism is more conducive to capturing salient regions on a larger scale. It is more concentrated in high-frequency and structure-rich locations while weakening the attention of some flat areas.

7) Adaptive Weighted Fusion: This part of the experiment is mainly analysis the difference in visual reconstruction results with or without the learnable weight coefficients of each branch in W^{2}AMSB. In the unweighted case, we initialize $\lambda_1 = \lambda_2 = 1$, which are constant and no guidance for the fusion of features from different branches. From the visual comparison of each sub-picture in Fig. 10, we can see that weight factors help improve visualization performance. In the marked position of red bounding boxes, more rich details are reconstructed, especially in some high-contrast locations, such as edges and tissues. The apparent differences in proton density and the apparent signal intensity help the doctor quickly distinguish different pathological tissues and make a qualitative diagnosis. Therefore, the feature visualization results indicate that weighted branch concatenation is beneficial to enhance the capacity of the model.

V. Discussion and Future Work

A. Image Texture Restoration

Structural edges and textures in MR images are usually important information used to distinguish different tissues and lesions. However, in deep super-resolution networks, some low-level structural information will gradually disappear as the network deepens, resulting in excessive smoothness in high-frequency regions. Therefore, how to retain more structural features is the focus of our next work. At present, in natural images, generative adversarial networks (GAN) [54] have superior performance in texture restoration, but the learning method of generating pixels is risky in medical images. We hope to preserve more useful features in future work by preprocessing the LR images with structural enhancement.

B. 3D Image Super-Resolution

The current work on MR image super-resolution is mainly for 2D format. Still, since many types of medical images are in 3D format, the spatial correlation information will be lost in splitting channels. There are already some methods to prove the scalability of CNN networks on 3D images [14], [15], [56], [57]. Our next work is to explore the performance of our W^{2}AMSN on 3D MR images. Making certain modifications to the network structure to use the 3D structure information rationally makes it possible to enhance the SR performance without introducing too many parameters.

C. More Efficient and Lightweight Networks

In the current image super-resolution methods using deep learning, performance improvement is often accompanied by the rapid growth of the network depth and the parameters. Those larger networks have high requirements on the memory and processor performance of the device. Although our network will not introduce explosive parameter growth in the deepening process, 3D image training will lead to some problems (e.g., model storage and model prediction speed). SqueezeNet [58], MobileNet [37] and ShuffleNet [59] obtained lightweight models by changing the convolution method. We will devote ourselves to designing more efficient computing methods for model compression to reduce network parameters without losing network performance in future work.

D. Real MR Image Super-Resolution

The synthetic MR datasets are based on the known downsampling method, and the low-quality real MR image contains a variety of unknown complex degradations. Therefore, the network performance based on the synthetic MR images is easily limited by the type of training dataset. Although the proposed method shows good adaptability on real in-vivo T1 images, there is still much room for improvement in the network mobility in the real world. In the future, we consider collecting LR-HR image pairs from real MR scanners to fine-tune the existing network, which can benefit the clinical application of deep learning-based MR SR methods.

VI. Conclusion

The main obstacle to applying deep models to MR image SR tasks is that the models cannot learn efficient expressions from limited data. The low utilization of the extracted features cannot guide the reconstruction of the fine structure. Multi-scale feature fusion can extract high-level semantics and low-level textures. A robust attention mechanism can further increase the model's ability to adapt to dense channel correlation selectively. Based on MR images sharing specific visual characteristics, we propose a wide weighted attention multi-scale model in this paper. Unlike simply concatenating the features of each scale, we design an adaptive non-reduction attention mechanism and learnable weighted feature fusion in each block. Considering the repeating structure and simple distribution of MR images, we use channel widening to improve the feature representation ability efficiently. Mixed self-attention includes local and global attention layers, allowing the model to accurately restore high-frequency distinctive structural textures in MR images. Furthermore, multi-level residual learning and residual scaling are introduced when stacking W^{2}AMSB, which can stabilize network training. Extensive experiments and ablation studies verify the advantages of the proposed W^{2}AMSN over other state-of-the-art methods, quantitatively and qualitatively. The superior reconstruction performance on real in-vivo MR images proves that our W^{2}AMSN has high clinical applicability and domain adaptability.

References

[1] E. Carmi, S. Liu, N. Alon, A. Fiat, and D. Fiat, "Resolution enhancement in mri," *Magn. Reson. Imag.*, vol. 24, no. 2, pp. 133–154, 2006.

[2] E. Plenge, D. H. Poot, M. Bernsen, G. Kotek, G. Houston, P. Wielopolski, L. van der Weerd, W. J. Niessen, and E. Meijering, "Super-resolution methods in mri: can they improve the trade-off between resolution, signal-to-noise ratio, and acquisition time?" *Magn. Reson. Med.*, vol. 68, no. 6, pp. 1983–1993, 2012.

[3] M. Irani and S. Peleg, "Improving resolution by image registration," *CVGIP: Graphical models and image processing*, vol. 53, no. 3, pp. 231–239, 1991.

[4] H. Stark and P. Oskoui, "High-resolution image recovery from image-plane arrays, using convex projections," *JOSA A*, vol. 6, no. 11, pp. 1715–1726, 1989.

[5] W. T. Freeman, T. R. Jones, and E. C. Pasztor, "Example-based super-resolution," *IEEE Comput. Graph. Appl. Mag.*, vol. 22, no. 2, pp. 56–65, 2002.

[6] C. Kim, K. Choi, and J. B. Ra, "Example-based super-resolution via structure analysis of patches," *IEEE Signal Process. Lett.*, vol. 20, no. 4, pp. 407–410, 2013.

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

[8] J. Yang, Z. Wang, Z. Lin, S. Cohen, and T. Huang, "Coupled dictionary training for image super-resolution," *IEEE Trans. Image Process.*, vol. 21, no. 8, pp. 3467–3478, 2012.

[9] 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, 2015.

[10] C. Dong, C. C. Loy, and X. Tang, "Accelerating the super-resolution convolutional neural network," in *Proc. Eur. Conf. Comput. Vis.* Springer, 2016, pp. 391–407.

[11] J. Kim, J. Kwon Lee, and K. Mu Lee, "Deeply-recursive convolutional network for image super-resolution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2016, pp. 1637–1645.

[12] Y. Tai, J. Yang, and X. Liu, "Image super-resolution via deep recursive residual network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2017, pp. 3147–3155.

[13] B. Lim, S. Son, H. Kim, S. Nah, and K. Mu Lee, "Enhanced deep residual networks for single image super-resolution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog. Workshop*, 2017, pp. 136–144.

[14] C.-H. Pham, A. Ducournau, R. Fablet, and F. Rousseau, "Brain mri super-resolution using deep 3d convolutional networks," in *Proc. IEEE Int. Symp. Biomed. Imag.*, 2017, pp. 197–200.

[15] Y. Chen, Y. Xie, Z. Zhou, F. Shi, A. G. Christodoulou, and D. Li, "Brain mri super resolution using 3d deep densely connected neural networks," in *Proc. IEEE Int. Symp. Biomed. Imag.*, 2018, pp. 739–742.

[16] C. Zhao, A. Carass, B. E. Dewey, and J. L. Prince, "Self super-resolution for magnetic resonance images using deep networks," in *Proc. IEEE Int. Symp. Biomed. Imag.*, 2018, pp. 365–368.

[17] Y. Zhang, Y. Tian, Y. Kong, B. Zhong, and Y. Fu, "Residual dense network for image super-resolution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2018, pp. 2472–2481.

[18] J. Li, F. Fang, K. Mei, and G. Zhang, "Multi-scale residual network for image super-resolution," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 517–532.

[19] 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.*, 2020.

[20] J. Yu, Y. Fan, J. Yang, N. Xu, Z. Wang, X. Wang, and T. Huang, "Wide activation for efficient and accurate image super-resolution," *arXiv preprint arXiv:1808.08718*, 2018.

[21] N. Cohen, O. Sharir, and A. Shashua, "On the expressive power of deep learning: A tensor analysis," in *Conference on Learning Theory*, 2016, pp. 698–728.

[22] R. Eldan and O. Shamir, "The power of depth for feedforward neural networks," in *Conference on learning theory*, 2016, pp. 907–940.

[23] S. Liang and R. Srikant, "Why deep neural networks for function approximation?" *arXiv preprint arXiv:1610.04161*, 2016.

[24] F. Scarselli and A. C. Tsoi, "Universal approximation using feedforward neural networks: A survey of some existing methods, and some new results," *Neural networks*, vol. 11, no. 1, pp. 15–37, 1998.

[25] X. Yang, S. Zhant, C. Hu, Z. Liang, and D. Xie, "Super-resolution of medical image using representation learning," in *International Conference on Wireless Communications & Signal Processing*. IEEE, 2016, pp. 1–6.

[26] J. Park, D. Hwang, K. Y. Kim, S. K. Kang, Y. K. Kim, and J. S. Lee, "Computed tomography super-resolution using deep convolutional neural network," *Physics in Medicine & Biology*, vol. 63, no. 14, p. 145011, 2018.

[27] J. V. Manjón, P. Coupé, A. Buades, V. Fonov, D. L. Collins, and M. Robles, "Non-local mri upsampling," *Medical image analysis*, vol. 14, no. 6, pp. 784–792, 2010.

[28] J. Kim, J. Kwon Lee, and K. Mu Lee, "Accurate image super-resolution using very deep convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2016, pp. 1646–1654.

[29] Z. He, Y. Cao, L. Du, B. Xu, J. Yang, Y. Cao, S. Tang, and Y. Zhuang, "Mrfn: Multi-receptive-field network for fast and accurate single image super-resolution," *IEEE Trans. Multimedia*, vol. 22, no. 4, pp. 1042–1054, 2019.

[30] 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.*, 2018, pp. 286–301.

[31] 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.*, 2019.

[32] O. Oktay, W. Bai, M. Lee, R. Guerrero, K. Kamnitsas, J. Caballero, A. de Marvao, S. Cook, D. O'Regan, and D. Rueckert, "Multi-input cardiac image super-resolution using convolutional neural networks," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2016, pp. 246–254.

[33] 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, 2004.

[34] K. Zhang, M. Sun, T. X. Han, X. Yuan, L. Guo, and T. Liu, "Residual networks of residual networks: Multilevel residual networks," *IEEE Trans. Circuits Syst. Video Technol.*, 2017.

[35] J. Jin, A. Dundar, and E. Culurciello, "Flattened convolutional neural networks for feedforward acceleration," *arXiv preprint arXiv:1412.5474*, 2014.

[36] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2017, pp. 1492–1500.

[37] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," *arXiv preprint arXiv:1704.04861*, 2017.

[38] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2018, pp. 4510–4520.

[39] A. Lahiri, S. Bairagya, S. Bera, S. Haldar, and P. K. Biswas, "Lightweight modules for efficient deep learning based image restoration," *IEEE Trans. Circuits Syst. Video Technol.*, 2020.

[40] 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 Recog.*, 2015, pp. 1–9.

[41] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2017, pp. 2117–2125.

[42] B. Singh and L. S. Davis, "An analysis of scale invariance in object detection snip," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2018, pp. 3578–3587.

[43] B. Singh, M. Najibi, and L. S. Davis, "Sniper: Efficient multi-scale training," in *Proc. Adv. Neural Inf. Process. Syst.*, 2018, pp. 9310–9320.

[44] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "Ssd: Single shot multibox detector," in *Proc. Eur. Conf. Comput. Vis.* Springer, 2016, pp. 21–37.

[45] Y. Li, Y. Chen, N. Wang, and Z. Zhang, "Scale-aware trident networks for object detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 6054–6063.

[46] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 11, pp. 1254–1259, 1998.

[47] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2018, pp. 7132–7141.

[48] D. Liu, B. Wen, Y. Fan, C. C. Loy, and T. S. Huang, "Non-local recurrent network for image restoration," in *Proc. Adv. Neural Inf. Process. Syst.*, 2018, pp. 1673–1682.

[49] T. Dai, J. Cai, Y. Zhang, S.-T. Xia, and L. Zhang, "Second-order attention network for single image super-resolution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2019, pp. 11 065–11 074.

[50] H. Wu, Z. Zou, J. Gui, W.-J. Zeng, J. Ye, J. Zhang, H. Liu, and Z. Wei, "Multi-grained attention networks for single image super-resolution," *IEEE Trans. Circuits Syst. Video Technol.*, 2020.

[51] 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 Recog.*, 2016, pp. 1874–1883.

[52] X. Zhao, Y. Zhang, T. Zhang, and X. Zou, "Channel splitting network for single mr image super-resolution," *IEEE Trans. Image Process.*, vol. 28, no. 11, pp. 5649–5662, 2019.

[53] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *Proc. Eur. Conf. Comput. Vis.* Springer, 2016, pp. 694–711.
[54] 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.
[55] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2921–2929.
[56] J. Hu, X. Wu, and J. Zhou, "Single image super resolution of 3d mri using local regression and intermodality priors," in *Eighth International Conference on Digital Image Processing (ICDIP 2016)*, vol. 10033. International Society for Optics and Photonics, 2016, p. 100334C.
[57] Y. Chen, F. Shi, A. G. Christodoulou, Y. Xie, Z. Zhou, and D. Li, "Efficient and accurate mri super-resolution using a generative adversarial network and 3d multi-level densely connected network," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2018, pp. 91–99.
[58] F. N. Iandola, S. Han, M. W. Moskewicz, K. Ashraf, W. J. Dally, and K. Keutzer, "Squeezenet: Alexnet-level accuracy with 50x fewer parameters and¡ 0.5 mb model size," *arXiv preprint arXiv:1602.07360*, 2016.
[59] X. Zhang, X. Zhou, M. Lin, and J. Sun, "Shufflenet: An extremely efficient convolutional neural network for mobile devices," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2018, pp. 6848–6856.

Haoqian Wang (M'13) received the B.S. and M.E. degrees from Heilongjiang University, Harbin, China, in 1999 and 2002, respectively, and the Ph.D. degree from the Harbin Institute of Technology, Harbin, in 2005. He was a Post-Doctoral Fellow with Tsinghua University, Beijing, China, from 2005 to 2007. He has been a Faculty Member with the Graduate School at Shenzhen, Tsinghua University, Shenzhen, China, since 2008, where he has also been an Associate Professor since 2011, and the director of Shenzhen Institute of Future Media Technology. His current research interests include generative adversarial networks, video communication, and signal processing.

Xiaowan Hu received B.E. degree from School of Electronic Science and Technology, Northwestern Polytechnical University, China, in 2019. She is currently pursuing an M.E. degree from the Department of Automation, Tsinghua University, China. She was awarded National Scholarship many times, was an outstanding graduate of Northwestern Polytechnical University, and won several national awards in China Robot Competition. Her research interests include image restoration and deep learning.

Xiaole Zhao received his B.S. and M.S. degrees in 2013 and 2016 from Southwest University of Science and Technology (SWUST), Mianyang, China, and his Ph.D. degree from University of Electronic Science and Technology of China (UESTC), Chengdu, China. He is currently working as an assistant professor at Southwest Jiaotong University (SWJTU), Chengdu, China. His research interests include computer vision and image processing, machine learning, medical image analysis, and deep learning techniques.

Yulun Zhang received B.E. degree from School of Electronic Engineering, Xidian University, China, in 2013 and M.E. degree from Department of Automation, Tsinghua University, China, in 2017. He is currently pursuing a Ph.D. degree with the Department of ECE, Northeastern University, USA. He received the Best Student Paper Award at IEEE International Conference on Visual Communication and Image Processing (VCIP) in 2015. He also won the Best Paper Award at IEEE International Conference on Computer Vision (ICCV) RLQ Workshop in 2019. His research interests include image restoration and deep learning.