Multi-Residual Feature Fusion Network for lightweight Single Image Super-Resolution

Jiayi Qin[1] and Zheng He[1] and Binyu Yan[1] and Gwanggil Jeon[2] and Xiaomin Yang[1]
College of Electronics and Information Engineering, Sichuan University, Chengdu, Sichuan, China[1]
School of Aeronautics & Astronautics, Sichuan University, Chengdu, Sichuan, China[1]
Department of Embedded Systems Engineering, Incheon National University, Incheon, 22012, Korea[2]
qinjiayi@stu.scu.edu.cn, hezheng@stu.scu.edu.cn, gjeon@inu.ac.kr, yby@scu.edu.cn, arielyang@scu.edu.cn

***Abstract*—Recently, single image super-resolution (SISR) methods using deep convolution neural networks (CNNs) have achieved remarkable performance. Especially, lightweight networks have received unprecedented attention because of their broad application prospects. However, existing methods for lightweight SR lack the adequate utilization of hierarchical features, which weakens the representation ability of the network. To alleviate this issue, we propose an effective and accurate multi-residual feature fusion network (MRFFN) for SISR. Specifically, we design a multi-residual block (MRB) to boost the representation ability of the network. By adopting the multi-residual learning (MRL) strategy, MRB can efficiently improve reconstruction results while halving the parameters, compared with the ordinary residual block (RB). To use the hierarchical features sufficiently, we construct a multi-residual fusion block (MRFB) by cascading the MRBs. Finally, we build our MRFFN by densely stacking MRFBs and introduce double-residual learning (DRL) strategy into the network at the global level. Extensive experiments demonstrate that the MRFFN is superior to the state-of-the-art SISR models while taking up less computing resources.**

***Index Terms*—Single Image Super-Resolution, Convolution Neural Network, Lightweight Network, Multi-Residual Feature Fusion.**

I. Introduction

Single image super-resolution (SISR) is a classical low-level computer vision task, which aims at reconstructing a high-resolution (HR) output image from its degraded low-resolution (LR) observation. However, SISR is inherently ill-posed since numerous HR images can be mapped to an identical LR observation. To solve this problem, numerous SISR methods, including interpolation-based strategies [1], reconstruct-based methods [2], and learning-based models [3], [4], [5], [6], [7], [8] have been proposed one after another.

In recent years, deep convolution neural networks (CNNs) have been proposed to boost the feature representation ability for accurate SISR predictions, which achieved outstanding reconstruction performance. As a pioneer, Dong et al. constructed SRCNN to establish a non-linear mapping from LR images to HR counterparts and obtain promising results. From then on, a flurry of researchers who studied CNN-based methods [9], [10], [11], [12], [13] have dedicated searching a proper mapping function from an interpolated input to its HR output. Moreover, to enhance the representation ability of networks, many existing methods [5], [11], [13], [14] attempted to enlarge their receptive fields directly by deepening or widening the networks. Although CNN-based SR models have made significant progress, the expensive computational consumption makes it difficult to adopt deep CNNs in practical applications.

To this end, numerous lightweight methods [10], [15], [16] were proposed to construct more efficient and accurate networks, which can meet the requirement of real-world applications. DRRN [10], DRCN [6], and CARN [15] adopted the recursive learning mechanism to form the lightweight networks while consuming fewer parameters. Hui et al. introduced the information distillation strategy and its variant to construct IDN [16] and IMDN [17], which effectively combined the informative features and enlarged the receptive fields for hierarchical features extraction. Although remarkable progress has been made in the above-mentioned lightweight networks, they still have the following limitations: (1) Most lightweight SR modules seldom achieve a superior balance between the number of parameters and the reconstruction performance; (2) Most of the lightweight SR methods do not make full use of the hierarchical features for image reconstruction, thereby hindering the network representation ability.

To address the above issues, we proposed a multi-residual block (MRB) to make outstanding balance and adopted a multi-residual learning (MRL) strategy to use the hierarchical features sufficiently. On one hand, we introduced a mixed attention module (MAM) to build the MRB, which can achieve a satisfying trade-off between performances and parameters. Extensive experiments have demonstrated that the MRB can obtain comparable performance while maintaining a reasonable number of parameters. On the other hand, we designed a multi-residual fusion block (MRFB) to fully use the hierarchical features, which adopted the multi-residual learning (MRL) strategy. Besides, by employing the double-residual learning (DRL) strategy, our multi-residual feature fusion network (MRFFN) showed better reconstruction results.

Our contributions of the proposed method can be summarized in the following two folds:

1. We employ the double-residual learning (DRL) strategy at the global level to form our multi-residual feature fusion network (MRFFN) (see Fig. 1) for lightweight SISR. Experimental results demonstrate that the proposed network can achieve a favorable trade-off between the network parameters and reconstruction performances. Meanwhile, compared with the other state-of-the-art methods, our MRFFN has remarkable

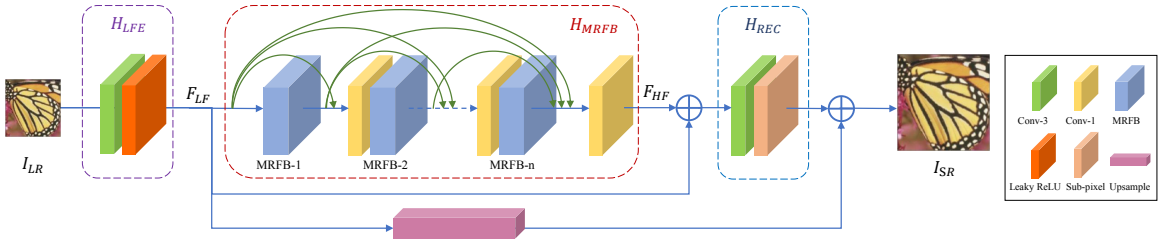

Fig. 1. Schematic diagram of multi-residual feature fusion network (MRFFN) framework.

performance with lower computational complexity.

2. We propose a mixed attention module (MAM) to build the multi-residual block (MRB), which achieves competitive results while utilizing fewer parameters. In addition, we adopt the multi-residual learning (MRL) strategy, which is beneficial to image reconstruction.

II. Related Work

In this section, we will briefly introduce the related technologies and methods in two aspects: single image super-resolution and attention mechanism.

A. Single image super-resolution

With the durative and rapid development of deep learning, plenty of methods based on CNNs have been proposed for SISR and achieved remarkable reconstruction performance. Dong et al. [18] creatively introduced a three-layer convolutional network named SRCNN, which was the preliminary work for image SR. To further accelerate the SRCNN model, they also designed an FSRCNN [18] framework, which upscaled the input at the end of the network. Afterward, Kim et al. explored the effectiveness of the network depth for achieving impressive performance. Based on residual learning, they designed deep SR models VDSR [4] and DRCN [6] for powerful feature expression. Tong et al. developed SRDenseNet [7] that employed dense connections to promote the flow of information. Furthermore, a deep and wide network EDSR [13] was proposed for better-recovering quality, which optimized the SRResNet [19] by removing the redundant modules and stacking residual blocks. Other deep CNN-based networks, like Memnet [20] and RDN [9], focused on the hierarchical features extracted from different receptive fields and further increased the depth of the framework. In addition, by fully utilizing the similarity of feature maps in spatial and channel dimensions, NLRN [21] and RCAN [5] were formed and outperformed other methods. Then, a very deep feedback network (SRFBN) [22] was presented to enhance the representative ability in computer vision tasks. Very recently, a PSNR-oriented method, namely EBRN [11] achieved visual and quantity improvements due to the powerful representation brought by the residual module. Meanwhile, Liu et al. produced the RFANet [12] to solve the case that it is insufficient to use the hierarchical features.

Even though the significant performance came from deep layers, they have expensive costs in both the computational resources and storage consumption. To alleviate this issue, numerous lightweight networks were proposed to maintain lower computation complexity for real-world applications. Hui et al. employed a state-based recursive strategy to construct an information distillation network (IDN) [16], which achieved better accuracy at a moderate size. Similarly, they developed a lightweight network with the multi-distillation named IMDN [17], which was superior to most existing SR methods on public benchmark datasets.

B. Attention mechanism

Attention mechanism, a weight distribution strategy, concentrates on the informative features and weights them according to various computer vision tasks. In recent years, attention mechanisms were also widely utilized in different SISR tasks such as image recovery, object detection, and facial recognition. For better refining the features of inner-channel, Zhang et al. [5] proposed a residual channel attention network (RCAN) to obtain the outstanding performance gain. Then, Hu et al. designed a compact, lightweight, and efficient squeeze-and-extraction (SE) which weighted channels discriminately. However, SENet [23] only explored the first-order statistic, which hindered the network's discriminative ability. To this end, Dai et al. [24] designed a second-order network (SAN) to study second-order statistics of features. Combining the advantages of spatial attention strategies and channel attention mechanisms, CBAM [25] inferred attention maps from the channel and spatial dimensions. Although the attention mechanism has made remarkable progress in SISR, there is still room for improvement in the representation accuracy. Inspired by RCAN [5] and CCNet [26], we proposed a mixed attention module, which combines the two attention mechanisms mentioned above (see Fig. 4).

III. Proposed Method

In this section, we will introduce our lightweight multi-residual feature fusion network (MRFFN) (as shown in Fig. 1). In detail, the backbone network consists of several multi-residual blocks (MRFBs). To fully utilize the intermediate features at the global level, we adopt a double-residual learning (DRL) strategy and densely stack MRFBs. As for each MRFB (see Fig. 2), it contains multiple MRBs to achieve better performance. Similarly, we adopt a multi-residual learning (MRL) strategy, which can also use the hierarchical features

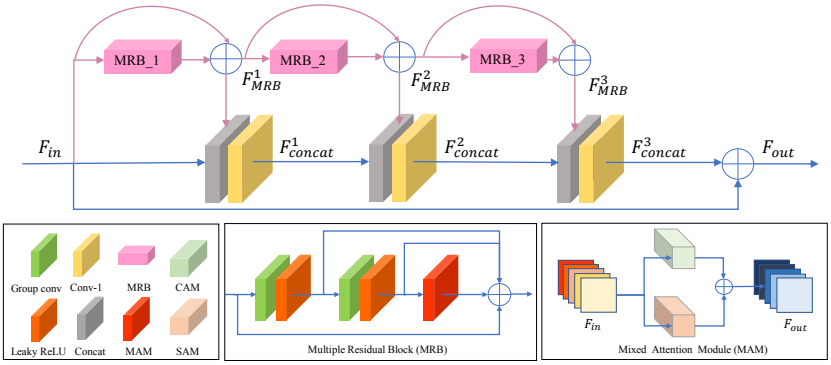

Fig. 2. Schematic diagram of multi-residual fusion block (MRFB).

at the local level. In addition, inside each MRB, we take advantage of group convolution and mixed attention module (MAM) to refine features efficiently.

A. Network structure

MRFFN is mainly composed of three parts (as shown in Fig. 1): a low-level feature extraction module, a non-linear mapping module, and a reconstruction module. We define the input LR image and the SR output as I_{LR} and I_{SR} accordingly. The real-world HR images trained in pairs with I_{LR} are then defined as I_{HR}. In the low-level feature extraction module, a 3×3 convolution (represented as $Conv3$) is employed to extract the shallow features of the LR image.

$$F_{LF} = H_{LFE}(I_{LR}), \tag{1}$$

where F_{LF} represents the output of the I_{LR} after low-level feature extraction, the $H_{LFE}(\cdot)$ denotes the low-level feature extraction operation. Then, the extracted shallow features are put into multiple densely connected MRFBs and processed step by step, then the feature F_{MRFB_n} can be extracted.

$$F_{MRFB_n} = H_{MRFB}(Conv1([F_1, ..., F_{MRFB_{(n-1)}}])), \tag{2}$$

here, F_{MRFB_n} represents the features extracted from the n-th ($n \geq 2$) MRFB. Accordingly, $H_{MRFB}(\cdot)$ denotes the process of each MRFB. In addition, $[\cdot]$ means concatenation operation, and $Conv1(\cdot)$ means 1×1 convolution. Assuming that there are N MRFBs in the network, we can get the high-level feature F_{HF} after the gradual processing of these blocks.

$$F_{HF} = F_{LF} + Conv1([F_1, ..., F_{MRFB_N}]), \tag{3}$$

where "+" stands for a global residual learning operation, which can more accurately establish the connection between low-level features and high-level features. Finally, we perform another global residual learning operation on the extracted high-level features F_{HF} and the input of MRFFN to obtain the output super-resolution image I_{SR}.

$$I_{SR} = H_{REC}(F_{HF}) + H_{UP}(F_{LF}), \tag{4}$$

where $H_{REC}(\cdot)$ and $H_{UP}(\cdot)$ indicate the reconstruction part and bilinear interpolation respectively. In the proposed framework, we choose the L_1 loss to train the network. For S training samples, L_1 loss function can be defined as follows:

$$L(\theta) = \frac{1}{S}\left\| H_{MRFFN}(I^i_{LR}) - I^i_{HR} \right\|_1, \tag{5}$$

Here, I^i_{LR} and I^i_{HR} denotes the ith image pairs in the training dataset. Besides, θ is the parameter set of our proposed network. Then, $H_{MRFFN}(\cdot)$ represents the function of MRFFN.

B. Multi-residual fusion block (MRFB).

The multi-residual fusion block (MRFB) is shown in Fig. 2. Given the input F_{in}, we can get the F_{out} of MRFB after a series of processing. The information flows through the proposed multi-residual block (MRB), we can naturally obtain the extracted feature.

First, we refine the hierarchical feature from each MRB. The following is the process of MRL strategy:

$$\begin{aligned} F^1_{MRB} &= H_{MRB\_1}(F_{in}) + F_{in}, \\ F^2_{MRB} &= H_{MRB\_2}(F^1_{MRB}) + F^1_{MRB}, \\ F^3_{MRB} &= H_{MRB\_3}(F^1_{MRB}) + F^2_{MRB}, \end{aligned} \tag{6}$$

where F^i_{MRB} represents the features extracted by the i-th MRB ($1 \leq i \leq 3$), and $H_{MRB\_i}(\cdot)$ refers to the feature extraction operation of the i-th MRB.

Second, we concatenate the residual outputs adopting the MRL strategy of every MRB to further improve the reconstruction ability. The extracted features we fuse in the k-th layer can be defined as F^k_{concat}:

$$\begin{aligned} F^1_{concat} &= Conv1([F_{in}, F^1_{MRB}]), \\ F^k_{concat} &= Conv1([F^{(k-1)}_{concat}, F^k_{MRB}]), \end{aligned} \tag{7}$$

where F^k_{MRB} means the residual features extracted by k-th ($1 < k \leq 3$) MRB.

Finally, given there are K MRBs in each MRFB, F^K_{concat} indicates the fused features after the process of K MRBs. After K-$level$ ($K = 3$) residual feature fusion, the hierarchical residual features are refined. Meanwhile, we can get F_{out} by performing residual learning with the F_{in}.

$$F_{out} = F^K_{concat} + F_{in}, \tag{8}$$

here, F^K_{concat} can also be denoted as F^3_{concat} (see Fig. 2), which represents the output features of hierarchical residual feature fusion.

C. Multi-residual block (MRB).

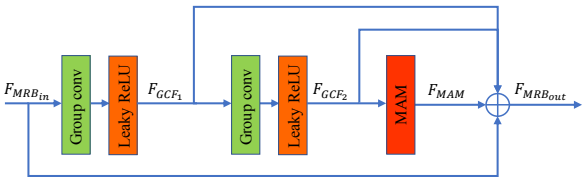

Fig. 3. Schematic diagram of multi-residual block (MRB).

The structure of multi-residual block (MRB) is given in Fig. 3. Given the input feature of MRB are denoted as $F_{MRB_{in}}$. By the process of the group convolution layers, and the utilization of activation function to fit the non-linear mapping, we can get the feature F_{GCF_t}:

$$\begin{aligned} F_{GCF_1} &= f_{GCF}(F_{MRB_{in}}), \\ F_{GCF_2} &= f_{GCF}(F_{GCF_1}), \end{aligned} \tag{9}$$

where F_{GCF_t} represents the feature obtained after the t-th group convolution layer. $f_{GCF}(\cdot)$ is the operation combining group convolution with activation function, i.e., Leaky ReLU [36].

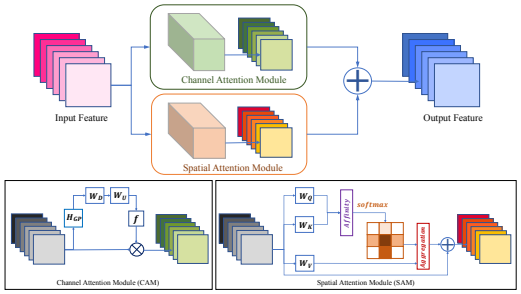

Fig. 4. Schematic diagram of mixed attention module (MAM). SAM is from [26], CAM is from [5].

Next, we define the function of the mixed attention module as $f_{MAM}(\cdot)$.

$$f_{MAM}(x) = H_{CAM}(x) \oplus H_{SAM}(x), \tag{10}$$

given an input x, $H_{CAM}(\cdot)$ and $H_{SAM}(\cdot)$ respectively correspond to the operation of the mixed attention module (MAM) (the module can refer to Fig. 4). Here, $H_{CAM}(\cdot)$ can be defined as following:

$$H_{CAM}(x) = x \otimes f_{sigmoid}(W_U(W_D(H_{GP}(x)))), \tag{11}$$

here, $\otimes$ denotes the operation of weighting different channels and H_{GP} represents the global pooling operation. Meanwhile, $f_{sigmoid}(\cdot)$ stands for the sigmoid activation function. Besides, $W_U(\cdot)$ and $W_D(\cdot)$ correspond to two fully-connection layers (more details can refer to [5]).

Then, we can express the operation of SAM with the following formula:

$$H_{SAM}(x) = x \oplus H_{Agg}(W_V(x), f_{softmax}(H_{Aff}(H_{QK}(x)))), \tag{12}$$

where $H_{Agg(\cdot)}$, $H_{Aff}(\cdot)$, and $H_{QK}(\cdot)$ are a series of operations given in Fig. 4, which can also refer to [26]. In addition, operator $\oplus$ indicates the element-wise addition operation and $W_V(\cdot)$ stands for a 1×1 convolution. Then, $f_{softmax}(\cdot)$ represents the softmax activation function. Moreover, the $H_{QK}(\cdot)$ can be described as follow:

$$H_{QK}(x) = W_Q(x) \times W_K(x)^{\mathrm{T}}, \tag{13}$$

where $W_Q(\cdot)$ and $W_K(\cdot)$ respectively represent two different 1×1 convolution layers. Then, "$\times$" denotes the matrix multiplication.

Finally, The above feature extraction results F_{GCF_1}, F_{GCF_2}, and F_{MAM} are combined for residual learning, we can get the output $F_{MRB_{out}}$.

$$F_{MRB_{out}} = F_{GCF_1} + F_{GCF_2} + F_{MAM} + F_{MRB_{in}}, \tag{14}$$

where F_{GCF_1} and F_{GCF_2} are features that respectively extracted from the first and second group convolution layers, while F_{MAM} denotes the output features of MAM. Besides, $F_{MRB_{in}}$ is the input of MRB.

IV. Experiment

A. Datasets and metrics

Based on previous work [15], [27], [28], we utilize the DIV2K [29] as the training dataset, which contains 800 high-resolution RGB images [30]and is widely employed in recent SISR methods. For evaluation, we adopt the most widely used five standard benchmark datasets, namely Set5 [31], Set14 [32], B100 [33], Urban100 [34] and Manga109 [35].

We select the peak signal-to-noise ratio (PSNR) and structural similarity (SSIM) [36] as the evaluation indexes of image reconstruction quality, which are obtained from the Y channel of YCbCr space. In addition, we adopt Mult-Adds to evaluate the computational complexity of a CNN model, which represents the number of composite multiply-accumulate operations for the single images. As with [15], we also assume the HR image size is 1280×720 to calculate Mult-Adds. By performing bicubic interpolation on Matlab, HR images of different scaling factors ($\times 2$, $\times 3$ and $\times 4$) can be obtained.

B. Implementation details

In the network structure, Leaky ReLU [37] is employed as the activation function followed by all of the group convolution layers. During training, the augmented dataset consists of images randomly flipped horizontally or vertically and 90° rotation. In each training batch, 16 RGB HR images are randomly cropped and fed to our MRFFN. The size of the LR patch images is determined by the corresponding factors. We select the normal initialization and apply the Adam [38] optimizer to optimizes MRFFN. For convenience, the initial learning rate of the Adam optimizer is set to 5×10^{-4} and decreases half for every 2×10^2 epochs. All experiments in this article are carried out under the PyTorch framework on NVIDIA 1080 Ti GPUs.

TABLE I
THE EFFECT OF MULTI-RESIDUAL BLOCK (MRB) ON PERFORMANCE AND EFFECTIVENESS. EVALUATION IN PSNR AND SSIM FOR ×4 SR. SRRESNET* IS REIMPLEMENTED RESULTS WITH DIV2K DATASET.

Methods	Params	Mult-Adds	Set5 PSNR/SSIM	Set14 PSNR/SSIM	B100 PSNR/SSIM	Urban100 PSNR/SSIM	Manga109 PSNR/SSIM
SRResNet* [19] _RB	1,510K	125.6G	32.03 / 0.8942	28.47 / 0.7787	27.49 / 0.7335	25.80 / 0.7770	30.09 / 0.9032
SRResNet* [19] _MRB	1,047K	98.8G	32.12 / 0.8938	28.52 / 0.7792	27.52 / 0.7344	25.99 / 0.7821	30.14 / 0.9048
MRFFN_RB	1,074K	61.6G	32.21 / 0.8951	28.62 / 0.7824	27.59 / 0.7362	26.11 / 0.7864	30.51 / 0.9088
MRFFN_MRB	699K	40.04G	32.29 / 0.8960	28.70 / 0.7828	27.60 / 0.7366	26.29 / 0.7903	30.61 / 0.7903

TABLE II
COMPARISON OF DIFFERENT NUMBERS OF MRFBS. WE OBSERVE THE BEST PSNR (DB) VALUES ON SET5 (×4) AND THE MULTI-ADDS OF SCALING FACTOR ×4.

#MRFBs	Params	Mul-adds	Set5 PSNR/SSIM	Set14 PSNR/SSIM	B100 PSNR/SSIM	Urban100 PSNR/SSIM	Manga109 PSNR/SSIM
B=2	356K	20.21G	32.01 / 0.8940	28.49 / 0.7785	27.49 / 0.7331	25.87 / 0.7776	30.11 / 0.9032
B=3	526K	29.83G	32.14 / 0.8937	28.53 / 0.7792	27.51 / 0.7339	26.03 / 0.7827	30.26 / 0.9050
B=4	699K	40.04G	32.23 / 0.8949	28.62 / 0.7807	27.57 / 0.7357	26.14 / 0.7862	30.39 / 0.9071
B=5	877K	49.76G	32.26 / 0.8955	28.61 / 0.7814	27.58 / 0.7363	26.27 / 0.7899	30.58 / 0.9088
B=6	1,059K	60.08G	32.28 / 0.8954	28.66 / 0.7820	27.59 / 0.7363	26.24 / 0.7897	30.56 / 0.9093

C. *Effectiveness of Multi-residual block (MRB)*

The structure of MRB is in Fig. 3, which consists of group convolution layers, Leaky ReLU, and mixed attention module (MAM). To verify the effectiveness of the proposed MRBs, we launch a comparative experiment as shown in Table I. To make a fair comparison, we complete comparative experiments between MRBs and ordinary RBs on SRResNet. The experimental results show that our MRBs have superiority compared to RBs, in the case of fewer parameters. Furthermore, we conduct comparative experiments on our MRFFN, which confirms that the MRB outperforms the RB.

In Table I, we give the results of comparative experiments under the structures of MRFFN and SRResNet. In SRResNet, by replacing the ordinary RBs with MRBs, a significant improvement is achieved in PSNR on five benchmark datasets. By replacing RB with MRB, we can achieve similar performance while reducing the parameters by nearly 500K.

D. *Ablation study*

According to the ideas of the network and the main contributions of this article, we conduct the following ablation experiments: the discussion of the numbers of MRFBs, the effectiveness of multi-residual learning (MRL) strategy, the effectiveness of double-residual learning (DRL) strategy, and the effectiveness of the mixed attention module (MAM). The corresponding experimental outcomes are described below.

Discussion on the numbers of MRFBs. Based on the previous researches, we find that deepening or widening can improve the construction ability but will bring about the increase of parameters. To construct a lightweight network for SISR, we must consider the parameter firstly. To better balancing the computation complexity and performance, we conduct the experiment that compares the PSNR of MRFFN with various numbers of MRFBs. According to other methods, we ingeniously set the numbers of MRFBs to 2, 3, 4, 5, and 6.

As shown in Table II, the deepening of the network brings the improvement of network performance. Considering the computational consumption of the MRFFN, with the increase of MRFB, the amounts of parameters and multi-Adds operations show a positive correlation growth trend. From the perspective of the reconstruction effect reflected by PSNR and SSIM, the deepening of the network will bring performance improvement (for example, the two indicators on Set5 show a trend of gradual increase). In summary, we can conclude that when the value of B is 4, a better balance can be obtained between the parameters and performance. Therefore, the numbers of MRFBs in our network is 4.

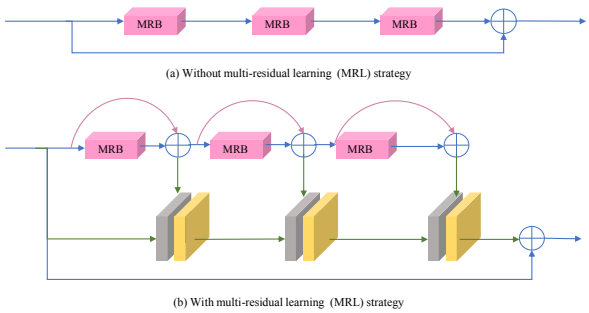

Fig. 5. Explore the effectiveness of multi-residual learning (MRL) strategy. (a) is the module without using MRL strategy, (b) is the module with employing MRL strategy.

TABLE III
VALIDITY OF MULTI-RESIDUAL LEARNING (MRL) STRATEGY. EVALUATION IN PSNR AND SSIM ON FIVE BENCHMARK DATASETS FOR ×4 SR.

Model	Set5 PSNR/SSIM	Set14 PSNR/SSIM	B100 PSNR/SSIM	Urban100 PSNR/SSIM	Manga109 PSNR/SSIM
w/o MRL	32.15 / 0.8942	28.57 / 0.7806	27.56 / 0.7352	26.09 / 0.7848	30.31 / 0.9059
w/ MRL	32.23 / 0.8949	28.62 / 0.7807	27.57 / 0.7357	26.14 / 0.7862	30.39 / 0.9071

Discussion on the effectiveness of multi-residual learning strategy. As shown in Fig. 5, we can refer to the structures of methods with or without multi-residual learning (MRL) strategy. In Table III, we can find that the network which adopts the MRL strategy will achieve better performances, compared with the network without MRL strategy. Experiments show that the MRL strategy can effectively enhance the reconstruction performance of the network, which has a significant improvement in PSNR and SSIM. What's more, the introduction of the MRL can further improve the characterization ability of the network

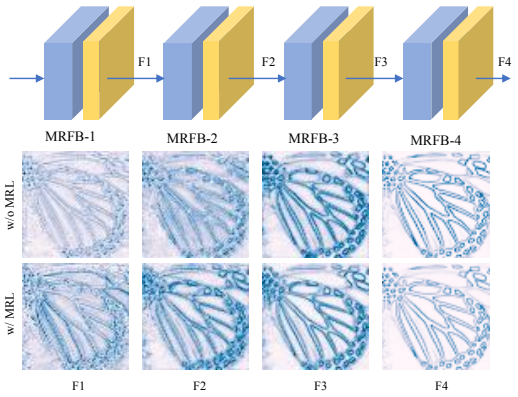

Fig. 6. Schematic diagram of the structure with or without multi-residual learning (MRL) strategy. Fi represents the average feature map extracted by the ith MRFB. At the bottom, the visualization results corresponding to the hierarchical features of different MRFBs are given. w/o MRL strategy is the result feature maps without MRL and w/ MRL represents the images adopt MRL strategy.

and achieve a more accurate reconstruction effect.

Fig. 6 shows the comparison of the visualization results before and after the MRL strategy is adopted. Through observation, we can find that using the MRL strategy can better restore the detailed information of the image and maintain a more complex spatial structure. The employ of MRL strategy further preserves the details of the spatial relationship to form the feature maps, which is of great help to image reconstruction.

TABLE IV
INVESTIGATION OF THE DOUBLE-RESIDUAL LEARNING (DRL) STRATEGY.

Model	Set5 PSNR/SSIM	Set14 PSNR/SSIM	B100 PSNR/SSIM	Urban100 PSNR/SSIM	Manga109 PSNR/SSIM
w/o DRL	32.20 / 0.8946	28.59 / 0.7807	27.55 / 0.7356	26.11 / 0.7859	30.38 / 0.9070
w/ DRL	32.24 / 0.8955	28.62 / 0.7817	27.55 / 0.7358	26.16 / 0.7875	30.49 / 0.9081

Discussion on double-residual learning strategy. Our MRFFN adopts the strategy of double-residual learning (DRL) at the global level to strengthen the representation ability of the network, to achieve more accurate image reconstruction. Different from the general residual learning, DRL effectively utilizes the shallow and deep features to jointly learn the non-linear mapping of our network. Regarding the proof of the validity of the DRL strategy, we can observe Table IV. In detail, on datasets Set5, Urban100 and Manga109, the PSNR value of each network has been increased by 0.04 dB, 0.05 dB, and 0.11 dB respectively.

Discussion on mixed attention module. The mixed attention module is demonstrated in Fig. 4, which is made up of the channel attention (CA) module and spatial attention (SA) module. For enhancing the representation ability of the MRFFN, we conduct a pixel-wise addition operation towards the two attention modules. The final results are given in Table V, which fully demonstrate the superiority of our MAM block, compared to a separate CA module or SA module. By adopting

TABLE V
INVESTIGATION OF THE MIXED ATTENTION MODULE (MAM). CA DENOTES CHANNEL ATTENTION MECHANISM, SA REPRESENTS SPATIAL ATTENTION MECHANISM.

Methods	Attention		Set5	Set14	B100	Urban100	Manga109
	CA	SA	PSNR/SSIM	PSNR/SSIM	PSNR/SSIM	PSNR/SSIM	PSNR/SSIM
w/o MAM			32.05 / 0.8929	28.51 / 0.7794	27.51 / 0.7336	25.87 / 0.7785	30.19 / 0.9045
w/ CA	✓		32.12 / 0.8940	28.52 / 0.7793	27.53 / 0.7339	25.94 / 0.7813	30.24 / 0.9052
w/ SA		✓	32.18 / 0.8946	28.59 / 0.7807	27.56 / 0.7353	26.13 / 0.7860	30.39 / 0.9067
w/ MAM	✓	✓	32.23 / 0.8949	28.62 / 0.7807	27.57 / 0.7357	26.14 / 0.7862	30.39 / 0.9071

MAM in our network, we can find that the representation performance measured by PSNR can increase by 0.18 dB on Set5, compared with the structure without MAM. In addition, by comparing the network that only utilizes the CA or SA mechanism and the one that uses a hybrid module, we can conclude that MAM is indeed effective.

E. Comparison with the state-of-the-arts

In this section, to demonstrate our MRFFN can perform well on five publicly available SR benchmark datasets, we conduct extensive experiments and compare them with 11 state-of-the-art lightweight methods. We deliberately select the following eleven data sets for testing: SRCNN [3], FSRCNN [18], VDSR [4], DRCN [6], LapSRN [8], DRRN [10], MemNet [20], CARN [15], IMDN [17], IDN*[1] [16], and our MRFFN. In this experiment, we measure all methods with the PSNR and SSIM values on five benchmark datasets, which shows the proposed method achieves the best performance and outperforms IMDN by a considerable margin.

As shown in Table VI, our proposed MRFFN has achieved a comprehensive surpass in all scales ($\times 2$, $\times 3$, or $\times 4$). In particular, compared with IMDN, our network uses similar parameters with excellent performance and has completed all-around transcendence. Let's take the $\times 4$ model as an example: our network parameters are slightly less than IMDN's, while the PSNR is improved by nearly 0.2 dB on the Manga109. In short, according to the results in the table, we can confirm that MRFFN can achieve the ideal image reconstruction effect within a reasonable range of parameters.

Visual comparisons of MRFFN with other lightweight methods on Urban100 and Manga109 datasets for $\times 4$ are shown in Fig. 7. To prove that our network can process different images, we select images with different styles and characteristics for comparison. First, we use two images on Urban100 to verify the effectiveness of MRFFN: "Img_073" shows that the network with the ability to recover images with square patterns is significantly better than other methods; meanwhile, "Img_092" confirms the reconstruction accuracy of MRFFN for stripe pattern recovery, which outperform other networks far beyond. Then, we select two images from Manga109 to further verify MRFFN's superiority: "Love Hina_vol14" shows the excellent results of the network's restoration on curve details, and "Shimattelkouze_vol26" also illustrates the network's powerful ability to restore texture information. In

[1]IDN* represents the result of retraining the IDN network on TensorFlow with the DIV2K data set.

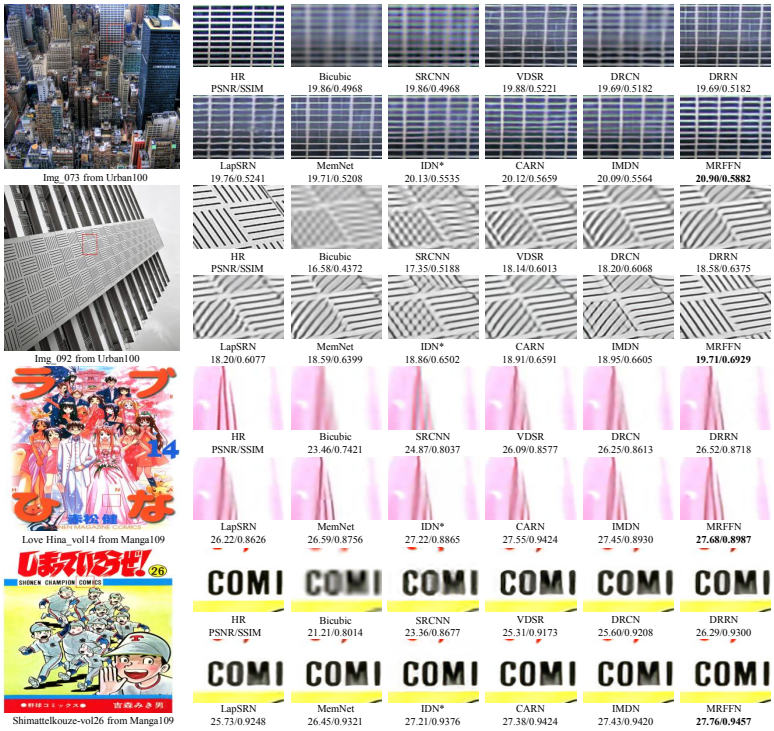

Fig. 7. Visual comparisons of MRFFN with other state-of-the-art lightweight methods on Urban100 and Manga109 datasets for $\times 4$ SR. The best results are highlighted.

summary, our network is more dominant than other frameworks in terms of performance and resource consumption.

V. Conclusions

In this paper, we proposed a multi-residual feature fusion network (MRFFN) for lightweight SISR. To enhance the reconstruction ability of the framework, we designed a multi-residual block (MRB), which contained a mixed attention module (MAM). Meanwhile, we proposed a multi-residual fusion block (MRFB) to boost network performance due to the introduction of multi-residual learning (MRL) strategy. Furthermore, the double-residual learning (DRL) was developed to construct our MRFFN, which achieved remarkable performance while increasing relatively few parameters. Extensive experiments demonstrated that the proposed method was superior to the state-of-the-art on five public benchmark datasets. In the future, we will continue leveraging the advantages of attention mechanisms and residual features, which can significantly boost the network representation ability.

Acknowledgements

The research in our paper is sponsored by the funding from Sichuan University under grant 2020SCUNG205.

References

[1] Zhang, Lei, and Xiaolin Wu. “An Edge-Guided Image Interpolation Algorithm via Directional Filtering and Data Fusion.” IEEE Transactions on Image Processing, vol. 15, no. 8, 2006, pp. 2226–2238.

[2] Zhang, Kaibing, et al. “Single Image Super-Resolution With Non-Local Means and Steering Kernel Regression.” IEEE Transactions on Image Processing, vol. 21, no. 11, 2012, pp. 4544–4556.

[3] Dong, Chao, et al. “Learning a Deep Convolutional Network for Image Super-Resolution.” European Conference on Computer Vision, 2014, pp. 184–199.

[4] Kim, Jiwon, et al. “Accurate Image Super-Resolution Using Very Deep Convolutional Networks.” 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 1646–1654.

[5] Zhang, Yulun, et al. “Image Super-Resolution Using Very Deep Residual Channel Attention Networks.” Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 294–310.

[6] Kim, Jiwon, et al. “Deeply-Recursive Convolutional Network for Image Super-Resolution.” 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 1637–1645.

[7] Tong, Tong, et al. “Image Super-Resolution Using Dense Skip Connections.” 2017 IEEE International Conference on Computer Vision (ICCV), 2017, pp. 4809–4817.

[8] Lai, Wei-Sheng, et al. “Deep Laplacian Pyramid Networks for Fast and Accurate Super-Resolution.” 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 5835–5843.

[9] Zhang, Yulun, et al. “Residual Dense Network for Image Super-Resolution.” 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018, pp. 2472–2481.

[10] Tai, Ying, et al. “Image Super-Resolution via Deep Recursive Residual Network.” 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 2790–2798.

[11] Qiu, Yajun, et al. “Embedded Block Residual Network: A Recursive Restoration Model for Single-Image Super-Resolution.” 2019 IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 4180–4189.

[12] Liu, Jie, et al. “Residual Feature Aggregation Network for Image Super-Resolution.” 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 2359–2368.

[13] Lim, Bee, et al. “Enhanced Deep Residual Networks for Single Image

TABLE VI
AVERAGE PSNR/SSIM FOR SCALE FACTORS ×2, ×3 AND ×4 ON SET5, SET14, B100, URBAN100, AND MANGA109. THE BEST AND THE SECOND RESULTS ARE HIGHLIGHTED WITH RED AND BLUE COLORS, RESPECTIVELY.

Method	Scale	Params	Mult-Adds	Set5 PSNR/SSIM	Set14 PSNR/SSIM	B100 PSNR/SSIM	Urban100 PSNR/SSIM	Manga109 PSNR/SSIM
Bicubic	×2	-	-	33.66 / 0.9299	30.24 / 0.8668	29.56 / 0.8431	26.88 / 0.8403	30.80 / 0.9339
SRCNN [3]		57K	52.7G	36.66 / 0.9524	32.45 / 0.9067	31.36 / 0.8879	29.50 / 0.8946	35.60 / 0.9633
FSRCNN [18]		12K	6.0G	37.00 / 0.9558	32.63 / 0.9088	31.53 / 0.8920	29.88 / 0.9020	36.67 / 0.9710
VDSR [4]		665K	612.6G	37.53 / 0.9587	33.03 / 0.9124	31.90 / 0.8960	30.76 / 0.9140	37.22 / 0.9750
DRCN [6]		1,774K	17,974.3G	37.63 / 0.9588	33.04 / 0.9124	31.85 / 0.8942	30.75 / 0.9133	37.55 / 0.9732
LapSRN [8]		813K	29.9G	37.52 / 0.9591	33.08 / 0.9130	31.80 / 0.8950	30.41 / 0.9101	37.27 / 0.9740
DRRN [10]		297K	6,796G	37.74 / 0.9591	33.23 / 0.9136	32.05 / 0.8973	31.23 / 0.9188	37.88 / 0.9749
MemNet [20]		677K	2,662.4G	37.78 / 0.9597	33.28 / 0.9124	32.08 / 0.8978	31.31 / 0.9195	37.72 / 0.9740
IDN* [16]		579K	124.6G	37.85 / 0.9598	33.58 / 0.9178	32.11 / 0.8989	31.95 / 0.9266	38.23 / 0.9758
CARN [15]		1,592K	222.8G	37.76 / 0.9590	33.52 / 0.9166	32.09 / 0.8978	31.92 / 0.9266	38.36 / 0.9765
IMDN [17]		694K	158.8G	38.00 / 0.9605	33.63 / 0.9177	32.19 / 0.8996	32.17 / 0.9283	38.88 / 0.9784
MRFFN(ours)		679K	136.8G	38.11 / 0.9608	33.83 / 0.9195	32.25 / 0.9004	32.57 / 0.9321	39.06 / 0.9776
Bicubic	×3	-	-	30.39 / 0.8682	27.55 / 0.7742	27.21 / 0.7385	24.46 / 0.7349	26.95 / 0.8566
SRCNN [3]		57K	52.7G	32.75 / 0.9090	29.30 / 0.8215	28.41 / 0.7863	26.24 / 0.7989	30.48 / 0.9117
FSRCNN [18]		12K	5.0G	33.18 / 0.9140	29.37 / 0.8240	28.53 / 0.7910	26.34 / 0.8080	31.10 / 0.9210
VDSR [4]		665K	612.6G	33.66 / 0.9213	29.77 / 0.8314	28.82 / 0.7976	27.14 / 0.8279	32.01 / 0.9340
DRCN [6]		1,774K	17,974.3G	33.82 / 0.9226	29.76 / 0.8311	28.80 / 0.7963	27.15 / 0.8276	32.24 / 0.9343
DRRN [10]		297K	6,796.9G	34.03 / 0.9244	29.96 / 0.8349	28.95 / 0.8004	27.53 / 0.8378	32.71 / 0.9379
MemNet [20]		677K	2,662.4G	34.09 / 0.9248	30.00 / 0.8350	28.96 / 0.8001	27.56 / 0.8376	32.51 / 0.9369
IDN* [16]		588K	56.3G	34.24 / 0.9260	30.27 / 0.8408	29.03 / 0.8038	27.99 / 0.8489	33.30 / 0.9421
CARN [15]		1,592K	118.8G	34.29 / 0.9255	30.29 / 0.8407	29.06 / 0.8034	28.06 / 0.8493	33.50 / 0.9440
IMDN [17]		703K	71.5G	34.36 / 0.9270	30.32 / 0.8417	29.09 / 0.8046	28.17 / 0.8519	33.61 / 0.9445
MRFFN(ours)		687K	68.8G	34.50 / 0.9279	30.41 / 0.8428	29.14 / 0.8062	28.41 / 0.8567	33.85 / 0.9460
Bicubic	×4	-	-	28.42 / 0.8104	26.00 / 0.7027	25.96 / 0.6675	23.14 / 0.6577	24.89 / 0.7866
SRCNN [3]		57K	52.7G	30.48 / 0.8628	27.50 / 0.7513	26.90 / 0.7010	24.52 / 0.7221	27.58 / 0.8555
FSRCNN [18]		12K	4.6G	30.72 / 0.8660	27.61 / 0.7550	26.98 / 0.7150	24.62 / 0.7280	27.90 / 0.8610
VDSR [4]		665K	612.6G	31.35 / 0.8838	28.01 / 0.7674	27.29 / 0.7251	25.18 / 0.7524	28.83 / 0.8870
DRCN [6]		1,774K	17,974.3G	31.53 / 0.8854	28.02 / 0.7670	27.23 / 0.7233	25.14 / 0.7510	28.93 / 0.8854
LapSRN [8]		813K	149.4G	31.54 / 0.8850	28.19 / 0.7720	27.32 / 0.7270	25.21 / 0.7560	29.09 / 0.8900
DRRN [10]		297K	6,796.9G	31.68 / 0.8888	28.21 / 0.7720	27.38 / 0.7284	25.44 / 0.7638	29.45 / 0.8946
MemNet [20]		677K	2,662.4G	31.74 / 0.8893	28.26 / 0.7723	27.40 / 0.7281	25.50 / 0.7630	29.42 / 0.8942
IDN* [16]		600K	32.3G	31.99 / 0.8928	28.52 / 0.7794	27.52 / 0.7339	25.92 / 0.7801	30.22 / 0.9032
CARN [15]		1,592K	90.9G	32.13 / 0.8937	28.60 / 0.7806	27.58 / 0.7349	26.07 / 0.7837	30.47 / 0.9084
IMDN [17]		715K	40.9G	32.21 / 0.8948	28.58 / 0.7811	27.56 / 0.7353	26.04 / 0.7838	30.45 / 0.9075
MRFFN(ours)		699K	40.0G	32.29 / 0.8960	28.71 / 0.7830	27.61 / 0.7371	26.35 / 0.7918	30.66 / 0.9101

Super-Resolution." 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2017, pp. 1132–1140.

[14] Dai, Tao, et al. "Second-Order Attention Network for Single Image Super-Resolution." 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 11065–11074.

[15] Ahn, Namhyuk, et al. "Fast, Accurate, and, Lightweight Super-Resolution with Cascading Residual Network." Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 256–272.

[16] Hui, Zheng, et al. "Fast and Accurate Single Image Super-Resolution via Information Distillation Network." 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018, pp. 723–731.

[17] Hui, Zheng, et al. "Lightweight Image Super-Resolution with Information Multi-Distillation Network." Proceedings of the 27th ACM International Conference on Multimedia, 2019, pp. 2024–2032.

[18] Dong, Chao, et al. "Image Super-Resolution Using Deep Convolutional Networks." IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 38, no. 2, 2016, pp. 295–307.

[19] Lai, Wei-Sheng, et al. "Fast and Accurate Image Super-Resolution with Deep Laplacian Pyramid Networks." IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 41, no. 11, 2019, pp. 2599–2613.

[20] Ledig, Christian, et al. "Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network." 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 105–114.

[21] Volos, Haris, et al. "Mnemosyne: Lightweight Persistent Memory." Proceedings of the Sixteenth International Conference on Architectural Support for Programming Languages and Operating Systems, vol. 39, no. 1, 2011, pp. 91–104.

[22] Liu, Ding, et al. "Non-Local Recurrent Network for Image Restoration." 32nd Conference on Neural Information Processing Systems, NeurIPS 2018, vol. 31, 2018, pp. 1673–1682.

[23] Li, Zhen, et al. "Feedback Network for Image Super-Resolution." 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 3867–3876.

[24] Hu, Jie, et al. "Squeeze-and-Excitation Networks." 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, vol. 42, no. 8, 2018, pp. 2011–2023.

[25] Woo, Sanghyun, et al. "CBAM: Convolutional Block Attention Module." Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 3–19.

[26] Huang, Zilong, et al. "CCNet: Criss-Cross Attention for Semantic Segmentation." 2019 IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 603–612.

[27] Zhang, Kai, et al. "Learning a Single Convolutional Super-Resolution Network for Multiple Degradations." 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018, pp. 3262–3271.

[28] Zhang, He, and Vishal M. Patel. "Density-Aware Single Image De-Raining Using a Multi-Stream Dense Network." 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018, pp. 695–704.

[29] Timofte, Radu, et al. "NTIRE 2017 Challenge on Single Image Super-Resolution: Methods and Results." 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2017, pp. 1110–1121.

[30] Agustsson, Eirikur, and Radu Timofte. "NTIRE 2017 Challenge on Single Image Super-Resolution: Dataset and Study." 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2017, pp. 1122–1131.

[31] Bevilacqua, Marco, et al. "Low-Complexity Single-Image Super-Resolution Based on Nonnegative Neighbor Embedding." British Machine Vision Conference 2012, 2012, pp. 1–10.

[32] Zeyde, Roman, et al. "On Single Image Scale-up Using Sparse-Representations." Proceedings of the 7th International Conference on Curves and Surfaces, 2010, pp. 711–730.

[33] Han, Heonjong, et al. "TRRUST v2: An Expanded Reference Database of Human and Mouse Transcriptional Regulatory Interactions." Nucleic Acids Research, vol. 46, 2018.

[34] Huang, Jia-Bin, et al. "Single Image Super-Resolution from Transformed Self-Exemplars." 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 5197–5206.

[35] Matsui, Yusuke, et al. "Sketch-Based Manga Retrieval Using Manga109 Dataset." Multimedia Tools and Applications, vol. 76, no. 20, 2017, pp. 21811–21838.

[36] Wang, Zhou, et al. "Image Quality Assessment: From Error Visibility to Structural Similarity." IEEE Transactions on Image Processing, vol. 13, no. 4, 2004, pp. 600–612.

[37] He, Kaiming, et al. "Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification." 2015 IEEE International Conference on Computer Vision (ICCV), 2015, pp. 1026–1034.

[38] Kingma, Diederik P., and Jimmy Lei Ba. "Adam: A Method for Stochastic Optimization." International Conference on Learning Representations (ICLR) 2015.