Signal, Image and Video Processing (2022) 16:937–945
https://doi.org/10.1007/s11760-021-02038-6

ORIGINAL PAPER

Single-image super-resolution via selective multi-scale network

Zewei He[1,2] · Binjie Ding[1,2] · Guizhong Fu[3] · Yanpeng Cao[1,2] · Jiangxin Yang[1,2] · Yanlong Cao[1,2]

Received: 2 April 2021 / Revised: 5 September 2021 / Accepted: 17 September 2021 / Published online: 20 October 2021

Abstract

In this paper, we aim to improve the performance of single-image super-resolution (SISR) by designing a more effective feature extraction module and a better fusion scheme for integrating hierarchical features. Firstly, we propose a selective multi-scale module (SMsM) to adaptively aggregate multi-scale features via self-learned weights and thus extract more distinctive representation. Then, we design an attentive global feature fusion (AGFF) scheme to reduce the redundant information inside the extracted hierarchical features by employing a gate mechanism (in the form of group convolution) and adaptively re-calibrate the features with channel-wise attention weights before fusion. Stacked SMsMs and AGFF compose a novel network which is termed selective multi-scale network (SMsN). Extensive experimental results demonstrate that our SMsN model outperforms some state-of-the-art SISR methods in terms of accuracy and efficiency.

Keywords Super-resolution · Convolutional neural network · Selective multi-scale network · Feature fusion

1 Introduction

Recently, single-image super-resolution (SISR), which aims to recover the high-resolution (HR) images from the corresponding low-resolution (LR) images, has become an extremely popular research topic among computer vision and robotics research communities [5,10,28]. The basic hypothesis to solve this challenging problem is that a mapping from LR to HR images can be learned from many training pairs.

Since a simple three-layer network SRCNN is firstly presented to learn the nonlinear mapping function [4], many sophisticated network architectures have been designed to improve the performance [2,17,18,29,30]. Most existing CNN-based models in SISR domain focus on designing deeper or more complex networks [12,32,36]. In this paper, our motivation is to explore effective techniques to extract and integrate hierarchical features to improve SISR performance, achieving higher restoration accuracy using fewer parameters.

Specifically, we propose a selective multi-scale network (SMsN) to adaptively aggregate multi-scale features and ease the training difficulty. Figure 1 shows the proposed architecture of SMsN. First, we design a module based on the selective kernel module (SKM) [24]. We optimize the standard SKM by replacing the $Softmax$ function with $Sigmoid$ to expand the solution search space. Such selective multi-scale module (SMsM) allows our proposed network to learn more distinctive features for the subsequent SISR task. We also propose an attentive global feature fusion (AGFF) scheme to take into account both low-level and high-level features. Different from global feature fusion (GFF), we embed a gate mechanism in the form of group convolution to filter out the redundant information inside the hierarchical features. Before fusion, the reduced features are adaptively re-calibrated with channel-wise attention weights. Thus, it can reduce the training difficulty of the network and enhance super-resolution results. In summary, the contributions of this paper are organized as follows:

This work was supported in part by the National Key Research and Development Program of China (2020YFB1711400) and the National Natural Science Foundation of China (52075485).

✉ Yanlong Cao
sdcaoyl@zju.edu.cn

Zewei He
zeweihe@zju.edu.cn

[1] State Key Laboratory of Fluid Power and Mechatronic Systems, School of Mechanical Engineering, Zhejiang University, Hangzhou 310027, China

[2] Key Laboratory of Advanced Manufacturing Technology of Zhejiang Province, School of Mechanical Engineering, Zhejiang University, Hangzhou 310027, China

[3] School of Mechanical Engineering, Suzhou University of Science and Technology, Suzhou 215009, China

- Inspired by the selective kernel module (SKM) in the image classification domain, we design a selective multi-scale module (SMsM) to learn channel-wise weights for adaptively fusing multi-scale features extracted from multiple branches with different kernels. The $Softmax$ operation is replaced with $Sigmoid$ to expand the search space for more distinctive features.
- To better utilize the hierarchical features of stacked SMsMs, we propose an enhanced global feature fusion (GFF) scheme, called attentive global feature fusion (AGFF). Different from GFF, we employ a gate mechanism based on group convolution to relieve the redundancy problem of hierarchical features and embed the channel attention mechanism to re-calibrate the features before fusion.
- Based on the above proposed techniques, we present a compact but powerful selective multi-scale network (SMsN) for high-quality SISR. Comparing with state-of-the-art methods, SMsN model achieves comparable image restoration performance by using fewer parameters and floating-point operations (FLOPs).

2 Related work

Recently, deep learning has achieved dominant advantages against conventional SISR methods. We briefly review some CNN-based SISR methods in this section.

Dong et al. [4,5] firstly proposed the three-layer model (i.e., SRCNN) to learn the end-to-end mapping function between pre-upscaled LR and HR images. After this pioneer work, substantial methods have been presented to chase more accurate recovery results by exploring deeper and more complex structures. Kim et al. designed two deep networks (VDSR [17] and DRCN [18]), employing residual learning and recursive convolution layers, respectively, to ease the training difficulty and meanwhile promote the performance. Tai et al. developed 52-layer DRRN [29] with recursive residual blocks and 80-layer MemNet [30] with memory blocks. He et al. [9] proposed MRFN with multi-receptive-field modules and optimized the model with a novel weighted Huber loss. Though containing only 22 layers, MRFN outperformed DRRN and MemNet. Later, SRResNet [22] stacked multiple original residual blocks [8] to boost the SISR reconstruction accuracy. EDSR [26] employed enhanced residual blocks and made a significant performance improvement in terms of PSNR and SSIM, and won the NTIRE 2017 competition [31]. SRDenseNet [32] utilized the dense block [13] as the basic module and achieved good performance. Zhang et al. [36] proposed RDN, which combined residual block and dense block to extract more abundant features for SISR, achieving higher values of metrics. Li et al. [25] developed SRFBN to integrate a feedback mechanism to refine low-level representations using high-level information. Liu et al. [27] designed a novel residual feature aggregation (RFA) framework to fully utilize the hierarchical features on the residual branches.

More recently, attention mechanism has been involved in SISR and can further improve the performance. Attention mechanism drives CNN to focus on salient or important parts and thus can guide the details recovery in SISR. Jiang et al. [16] introduced SENet [11] into capsule block for SISR. Zhang et al. [35] integrated SENet [11] into residual block and employed residual-in-residual structure to form RCAN, which pushed the state-of-the-art SISR performance forward. However, SENet only exploited first-order statistics of features and ignored higher-order statistics. Dai et al. [3] presented a second-order channel attention module to improve the discriminative ability of SISR network. Hu et al. [12] developed a CSAR block via combining spatial and channel-wise attention mechanisms into the residual block to adaptively modulate the feature representations.

Another direction is to develop SISR models for mobile application. Dong et al. [6] initially proposed a fast variant of SRCNN (i.e., FSRCNN), which took LR images as the network input and employed a deconvolution layer in the model tail for upscaling the spatial size. For similar purpose, Shi et al. [33] proposed the sub-pixel convolution (i.e., pixel-shuffle operation) to rearrange the tensor elements for fast and accurate upscaling of LR images. Following them, the lightweight SISR methods were developed. Representatively, Hui et al. presented information distillation network (IDN) [15] and IMDN [14] to balance performance and speed, thus improved the applicability. He et al. [2] modified the residual block and developed an energy-aware improved deep residual network (EA-IDRN) to investigate a number of design options for fast and accurate SISR. Afterward, Lai et al. developed LapSRN [20] and ms-LapSRN [21] by progressively reconstructing sub-band residual HR images at multiple pyramid levels.

3 Methodology

3.1 Network structure

Figure 1 shows the pipeline of our proposed SMsN model. Our SMsN model consists of three sub-networks: an initial feature extraction sub-network (IFENet) to learn feature maps from low-resolution input I^{LR}, a feature mapping sub-network (FMNet) to transform low-level features into high-level ones, and an image reconstruction sub-network (IRNet) to reconstruct the super-resolved high-resolution image I^{SR}.

Given a LR input image $I^{LR} \in \mathbb{R}^{3\times H\times W}$, a 3×3 convolutional layer is firstly deployed in IFENet to extract initial features $F_0 \in \mathbb{R}^{C\times H\times W}$ as

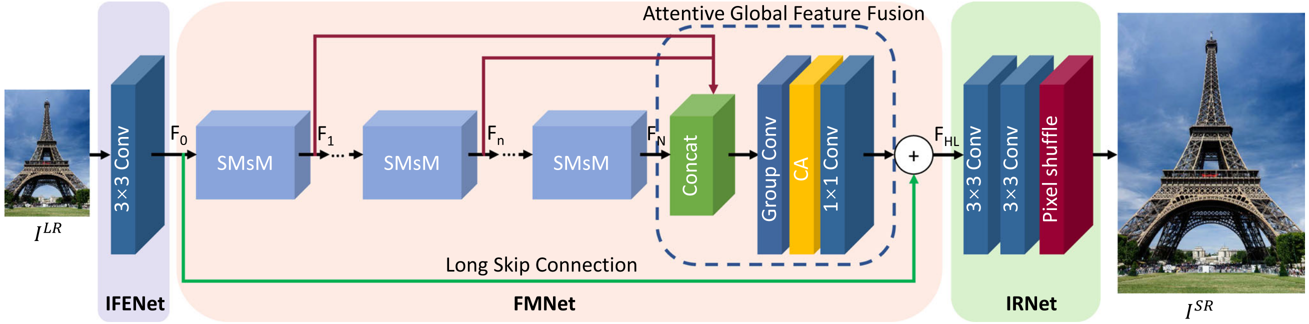

Fig. 1 Network architecture of our proposed Selective Multi-scale Network (SMsN), which consists of three sub-networks: an initial feature extraction sub-network (IFENet), a feature mapping sub-network (FMNet), and an image reconstruction sub-network (IRNet). CA denotes the channel attention mechanism, F_n denotes the output features of n-th Selective Multi-scale Module (SMsM), and F_{HL} denotes the high-level features before upsampling part. Given a low-resolution input image I^{LR}, the aim of our SMsN is to reconstruct the high-resolution image output I^{SR}

$$F_0 = \mathcal{F}_{IFENet}(I^{LR}), \tag{1}$$

where $\mathcal{F}_{IFENet}(\cdot)$ stands for the operation of IFENet. Then, the extracted F_0 is fed into FMNet for learning the high-level features $F_{HL} \in \mathbb{R}^{C\times H\times W}$, which are used for reconstructing super-resolved I^{SR}. FMNet is the core of our SMsN model, and it contains a sequence of stacked Selective Multi-scale Module (SMsM), an Attentive Global Feature Fusion (AGFF), and a long skip connection.

$$\begin{aligned} F_{HL} &= \mathcal{F}_{FMNet}(F_0) \\ &= F_0 + \mathcal{F}_{AGFF}(\mathcal{F}_{SMsM,N}(\mathcal{F}_{SMsM,N-1}(\cdots \\ &\quad (\mathcal{F}_{SMsM,1}(F_0))\cdots))), \end{aligned} \tag{2}$$

where $\mathcal{F}_{FMNet}(\cdot)$ represents the operation of FMNet. $\mathcal{F}_{SMsM,n}(\cdot)$ and $\mathcal{F}_{AGFF}(\cdot)$ denote the operation of n-th SMsM module and the operation of AGFF, respectively. For a $\times R$ upscaling SISR task, two 3×3 convolutional layers are utilized to convert the channel number of F_{HL} from C to $3 \times R \times R$ and the pixel shuffle operation [33] upscales and reconstructs the super-resolved output $I^{SR} \in \mathbb{R}^{3\times RH\times RW}$ as

$$\begin{aligned} I^{SR} &= \mathcal{F}_{IRNet}(F_{HL}) \\ &= \mathcal{F}_{\uparrow R}(\mathcal{F}_{3\times3}(\mathcal{F}_{3\times3}(F_{HL}))), \end{aligned} \tag{3}$$

where $\mathcal{F}_{IRNet}(\cdot)$ stands for the operation of IRNet. $\mathcal{F}_{3\times3}(\cdot)$ and $\mathcal{F}_{\uparrow R}$ denote the 3×3 convolution layer and the $\times R$ pixel shuffle operation, respectively. The SMsN model is optimized by minimizing the pixel-wise difference between the predicted super-resolved image I^{SR} and corresponding ground truth I^{GT}. We adopt the $L1$ loss function to drive the weights learning [26,32].

3.2 Selective multi-scale module

Extracting image features on different scales has been proved to be effective in recent SISR literature [9,23]. The most commonly used technique for feature fusion is employing a simple concatenation layer and a convolutional layer to integrate the extracted multi-scale features. The convolutional layer linearly aggregates the concatenated multi-scale features. To improve the adaption capability of feature fusion, Li et al. [24] present a *Selective Kernel module* (SKM), which utilizes self-learned selection weights to aggregate information from multiple branches, making the neurons adaptively adjust sizes of their receptive fields.

Inspired by its successful application in the image classification domain, SKM is integrated into our *Selective Multi-scale Module* (SMsM) to solve the challenging SISR problem. First, we remove some irrelevant components such as batch normalization and group convolution. Then, we replace the $Softmax$ function with $Sigmoid$ for more flexible weights learning. Also, we add a convolutional layer and a skip connection to ease the training procedure.

As illustrated in Fig. 2, given a feature map $\mathbf{X} \in \mathbb{R}^{C\times H\times W}$, two convolutional layers with different kernel sizes (i.e., 3×3 and 5×5) conduct transformations, respectively: $\widetilde{\mathcal{F}}_{3\times3}$: $\mathbf{X} \rightarrow \widetilde{\mathbf{U}} \in \mathbb{R}^{C\times H\times W}$ and $\widehat{\mathcal{F}}_{5\times5} : \mathbf{X} \rightarrow \widehat{\mathbf{U}} \in \mathbb{R}^{C\times H\times W}$. For the efficiency consideration, the 5×5 convolution is replaced by using a 3×3 dilated convolution and setting dilation value to 2. Then, $\widetilde{\mathbf{U}}$ and $\widehat{\mathbf{U}}$ are combined via element-wise summation operation:

$$\mathbf{U} = \widetilde{\mathbf{U}} + \widehat{\mathbf{U}}, \tag{4}$$

where $\mathbf{U} \in \mathbb{R}^{C\times H\times W}$ are the fused multi-scale features.

Then, we encode the global information by simply using a global average pooling (GAP) to generate initial channel-wise weights $\mathbf{s} \in \mathbb{R}^{C\times1\times1}$. The c-th element of $\mathbf{s}$ can be

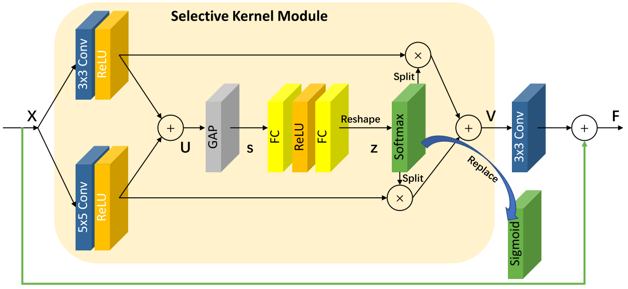

Fig. 2 Diagram of proposed Selective Multi-scale Module (SMsM). The yellow box part is the original SKM from [24]. SMsM replaces *Softmax* with *Sigmoid*, adds a 3 × 3 convolution after the **V**, and fuses with the input **X** to obtain the final output **F**

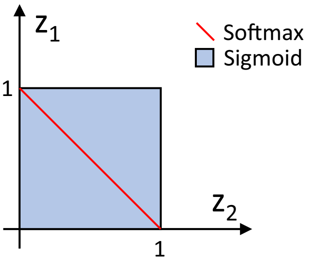

Fig. 3 Solution space illustration of *Softmax* and *Sigmoid* operations

expressed as:

$$\mathbf{s}_c = GAP(\mathbf{U}_c) = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} \mathbf{U}_c(i, j), \tag{5}$$

where $GAP(\cdot)$ denotes the global average pooling operation and $\mathbf{U}_c(i, j)$ is the value at coordinate position (i, j) of the c-th channel of $\mathbf{U}$.

We employ two fully connected (FC) layers and a ReLU activation to compute the selection weights $\mathbf{z} \in \mathbb{R}^{2C \times 1 \times 1}$:

$$\mathbf{z} = FC_2(max(0, FC_1(\mathbf{s}))), \tag{6}$$

where $max(0, x)$ denotes the ReLU activation. In our implementation, the FC layers are performed by 1×1 convolutional layers. The first FC layer reduces the dimension from C to C/r, and the second FC layer expands the dimension from C/r to $2C$. r means the reduction ratio and $r = 4$ in our implementation.

In the original SKM, the authors reshape the weights $\mathbf{z}$ to $2 \times C \times 1$, adopt a *Softmax* operation to normalize the selection weights across channels, and split the normalized weights into two vectors: $\mathbf{z}_1 \in \mathbb{R}^{C \times 1 \times 1}$ and $\mathbf{z}_2 \in \mathbb{R}^{C \times 1 \times 1}$. However, the *Softmax* operation imposes a strong constraint for $\mathbf{z}_1$ and $\mathbf{z}_2$ as $\mathbf{z}_1 + \mathbf{z}_2 = 1$. This means the search space of $\mathbf{z}_1$ and $\mathbf{z}_2$ must be on the red line segment of Fig. 3.

As the result, when $\mathbf{z}_1$ learns large values to emphasize the features extracted using 3 × 3 convolutions, then the computed weights $\mathbf{z}_2$ for larger scale features will become insignificant, and vice versa. However, imposing such a competitive mechanism in the fusion process of the extracted multi-scale features restricts the search space of $\mathbf{z}_1$ and $\mathbf{z}_2$ and might not be optimal for the SISR task. Therefore, we replace the original *Softmax* with the *Sigmoid* operation in our proposed SMsM for more flexible learning of $\mathbf{z}_1$ and $\mathbf{z}_2$ as $0 < \mathbf{z}_1, \mathbf{z}_2 < 1$. In this situation, the search space of $\mathbf{z}_1$ and $\mathbf{z}_2$ is expanded to the light blue box of Fig. 3.

By applying the computed attention weights to multi-scale features, the fused features $\mathbf{V} \in \mathbb{R}^{C \times H \times W}$ are obtained as

$$\mathbf{V} = \mathbf{z}_1 \cdot \widetilde{\mathbf{U}} + \mathbf{z}_2 \cdot \widehat{\mathbf{U}}. \tag{7}$$

For the SISR task, we add a 3 × 3 convolutional layer and fuse with the input **X** to obtain the final output $\mathbf{F} \in \mathbb{R}^{C \times H \times W}$ as

$$\mathbf{F} = \mathbf{X} + \mathcal{F}_{3 \times 3}(\mathbf{V}). \tag{8}$$

3.3 Attentive global feature fusion

With the growth of depth, network will become difficult to train due to the gradient vanishing problem. More specifically, the low-level features gradually lost/disappear in the forward pass process to deeper layers, as shown in the plain structure (Fig. 4a). However, it is reported in many previous research works that the low-level features play a non-negligible role in reconstructing the super-resolved HR image [12,30,34]. It is important to fully utilize the hierarchical features extracted in different modules.

A feasible solution to alleviate the gradient vanishing problem is to directly send the low-level features to deeper layers via skip connections. Given the multi-scale features extracted by individual SMsMs, the global feature fusion (GFF) [23,36] performs concatenation operation at the end of the network and utilize a 1 × 1 convolutional layer to adaptively select informative features as

$$\mathcal{F}_{GFF} : \mathcal{F}_{1 \times 1}([F_1 || F_2 || \cdots || F_N]), \tag{9}$$

where $[||]$ denotes the concatenation operation and F_n denotes the output features of the n-th SMsM.

However, the concatenated hierarchical features (features extracted using individual SMsMs) contain a large amount of redundant information. Therefore, it is difficult to generate distinctive features for the subsequent SISR task by utilizing a 1 × 1 convolutional layer to linearly combine features from such a huge concatenated feature bank. Moreover, the output

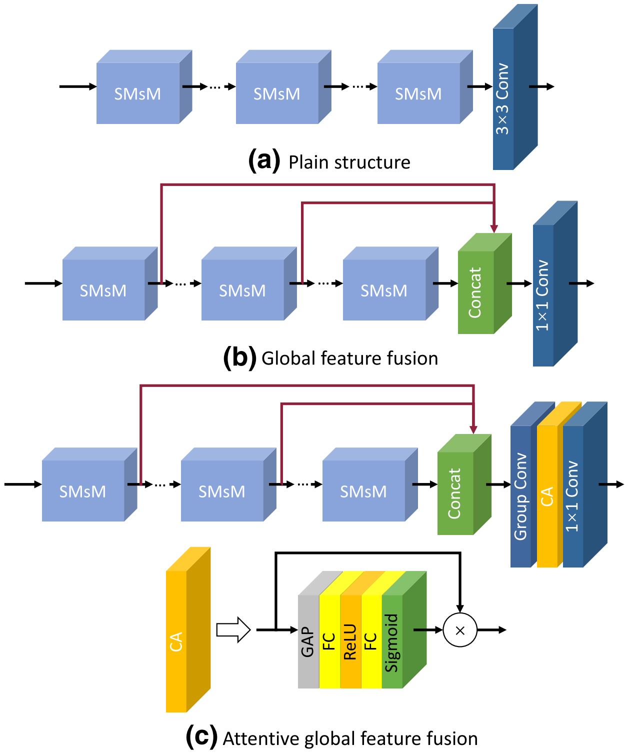

Fig. 4 Different connection schemes between SMsMs: **a** plain structure; **b** global feature fusion; **c** attentive global feature fusion. CA denotes the channel attention mechanism, GAP denotes the global average pooling operation, and FC denotes the fully connected layer

of 1 × 1 convolution is produced by a weighted summation through all channels; thus, the inter-dependencies between channels are implicitly embedded in the convolution layer and cannot adaptively change for different scenes.

We design an attentive global feature fusion (AGFF) scheme (Fig. 4c) to solve the mentioned problems. To relieve the redundancy problem, we firstly employ a group convolution by setting the group numbers to N to filter out some similar features inside a single SMsM. The channel dimension is reduced from $C*N$ to $\frac{C}{4}*N$ via this parameter-compression operation. Then, we utilize the channel attention (CA) mechanism [11,35] to re-calibrate reduced hierarchical features via explicitly learning channel-wise importance scores. The scores are feature specific, which can be adaptively adjusted based on input images. According to [35], we set the reduction ratio in the CA to 16. AGFF can be formulated as

$$\mathcal{F}_{AGFF} : \mathcal{F}_{1\times 1}(\mathcal{F}_{CA}(\mathcal{F}_{GC}([F_1||F_2||\cdots||F_N]))), \quad (10)$$

where $\mathcal{F}_{GC}(\cdot)$ and $\mathcal{F}_{CA}(\cdot)$ denote the group convolution and channel attention operation.

Table 1 Comparative results of *Ms_Baseline*, *Ms_SKM* and *SMsM*

Modules	Ms_Baseline	Ms_SKM	SMsM
Weights Function	N/A	*Softmax*	*Sigmoid*
Urban100 - ×2	32.25 dB	32.38 dB	32.46 dB
Urban100 - ×4	26.10 dB	26.20 dB	26.29 dB
# Module Param.	147,456	113,664	113,664

Tested on Urban100 for scale factors ×2 and ×4

Table 2 Comparative results of different fusion schemes.

Schemes	Plain	GFF	AGFF
Urban100 - ×2	32.46 dB	32.48 dB	32.51 dB
Urban100 - ×2	26.29 dB	26.32 dB	26.39 dB

Tested on Urban100 for scale factors ×2 and ×4

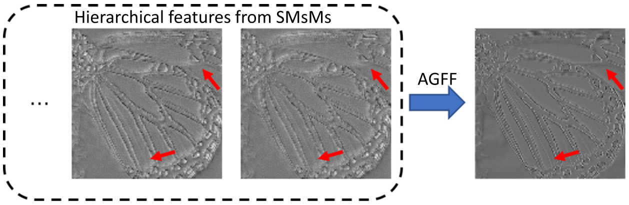

Fig. 5 Visualization of the feature maps before and after AGFF. Please zoom-in the figure for better observation

4 Experimental results and comparisons

4.1 Implementation detail

We train our SMsN model on DIVerse 2K resolution image dataset (i.e., DIV2K) [1] and evaluate on five commonly used public benchmark datasets: Set5, Set14, B100, Urban100, and Manga109. Then, we also evaluate the effectiveness of our SMsN on real-captured low-resolution images. We report two evaluation metrics (i.e., PSNR and SSIM) on the Y channel (i.e., luminance) of transformed YCbCr space and discarded pixels in the boundary areas of images according to [5].

We implemented our SMsN model (stacking $N = 64$ SMsMs) in the Pytorch platform and trained this model by optimizing $L1$ loss function on a single NVIDIA RTX2080Ti GPU. The Adam [19] solver is utilized to optimize the weights by setting $\beta_1 = 0.9$, $\beta_2 = 0.999$ and $\varepsilon = 1e^{-8}$. The initial learning rate is set to $1e^{-4}$ and halved every 200 epochs. When training the SMsN model for ×3 and ×4 SISR tasks, we initialized the weights using the parameters of the pre-trained ×2 model. The trained SMsN model and source codes will be made public.

dummy

Table 3 Benchmark results compared with the state-of-the-art SISR methods

Scale	Method	Set5 PSNR/SSIM	Set14 PSNR/SSIM	B100 PSNR/SSIM	Urban100 PSNR/SSIM	Manga109 PSNR/SSIM
×2	Bicubic	33.66/0.9299	30.24/0.8688	29.56/0.8431	26.88/0.8403	30.80/0.9339
	LFFN [34]	37.95/0.9597	–/–	32.20/0.8994	32.39/0.9299	38.73/0.9765
	IMDN [14]	38.00/0.9605	33.63/0.9177	32.19/0.8996	32.17/0.9283	38.88/0.9774
	MSRN [23]	38.07/0.9608	33.68/0.9184	32.22/0.9002	32.32/0.9304	38.64/0.9771
	D-DBPN [7]	38.09/0.9600	33.85/0.9190	32.27/0.9000	32.55/0.9324	38.89/0.9775
	EDSR [26]	38.11/0.9602	33.92/0.9195	32.32/0.9013	32.93/0.9351	39.10/0.9773
	RDN [36]	38.24/0.9614	34.01/0.9212	32.34/0.9017	32.89/0.9353	39.18/0.9780
	SRFBN [25]	38.11/0.9609	33.82/0.9196	32.29/0.9010	32.62/0.9318	39.08/0.9779
	SMsN	38.23/0.9614	34.04/0.9215	32.36/0.9019	33.07/0.9364	39.26/0.9777
	SMsN-L	**38.26/0.9615**	**34.05/0.9216**	**32.37/0.9020**	**33.13/0.9370**	**39.31/0.9781**
×3	Bicubic	30.39/0.8682	27.55/0.7742	27.21/0.7385	24.46/0.7349	26.95/0.8556
	LFFN [34]	34.43/0.9266	–/–	29.13/0.8059	28.34/0.8558	33.65/0.9445
	IMDN [14]	34.36/0.9270	30.32/0.8417	29.09/0.8046	28.17/0.8519	33.61/0.9445
	D-DBPN [7]	–/–	–/–	–/–	–/–	–/–
	MSRN [23]	34.48/0.9276	30.40/0.8436	29.13/0.8061	28.31/0.8560	33.56/0.9451
	EDSR [26]	34.65/0.9280	30.52/0.8462	29.25/0.8093	28.80/0.8653	34.17/0.9476
	RDN [36]	34.71/0.9296	30.57/0.8468	29.26/0.8093	28.80/0.8653	34.13/0.9484
	SRFBN [25]	34.70/0.9292	30.51/0.8461	29.24/0.8084	28.73/0.8641	34.18/0.9481
	SMsN	34.65/0.9295	30.58/0.8473	29.29/0.8102	28.91/0.8668	34.31/0.9488
	SMsN-L	**34.75/0.9299**	**30.62/0.8474**	**29.31/0.8105**	**28.97/0.8678**	**34.36/0.9493**
×4	Bicubic	28.42/0.8104	26.00/0.7027	25.96/0.6675	23.14/0.6577	24.89/0.7866
	LFFN [34]	32.15/0.8945	–/–	27.52/0.7377	26.24/0.7902	30.66/0.9099
	IMDN [14]	32.21/0.8948	28.58/0.7811	27.56/0.7353	26.04/0.7838	30.45/0.9075
	MSRN [23]	32.25/0.8958	28.63/0.7833	27.61/0.7377	26.20/0.7905	30.57/0.9103
	D-DBPN [7]	32.47/0.8980	28.82/0.7860	27.72/0.7400	26.38/0.7946	30.91/0.9137
	EDSR [26]	32.46/0.8968	28.80/0.7876	27.71/0.7420	26.64/0.8033	31.02/0.9148
	RDN [36]	32.47/0.8990	28.81/0.7871	27.72/0.7419	26.61/0.8028	31.00/0.9151
	SRFBN [25]	32.47/0.8983	28.81/0.7868	27.72/0.7409	26.60/0.8015	31.15/0.9160
	SMsN	32.50/0.8991	28.82/0.7871	27.76/0.7424	26.68/0.8036	31.15/0.9153
	SMsN-L	**32.51/0.8991**	**28.86/0.7877**	**27.78/0.7430**	**26.78/0.8061**	**31.22/0.9164**

Bold and underline indicate the best and the second best performance, respectively

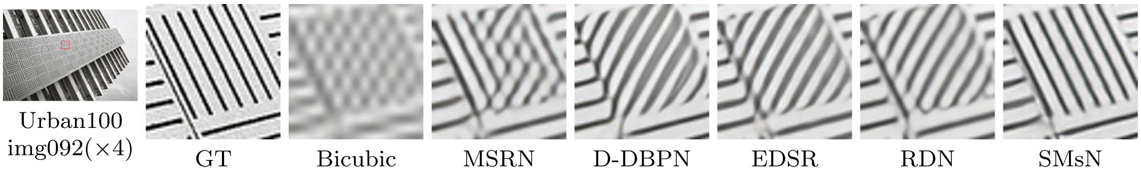

Fig. 6 Qualitative comparison of ×4 SR results for “img092” in Urban100 dataset

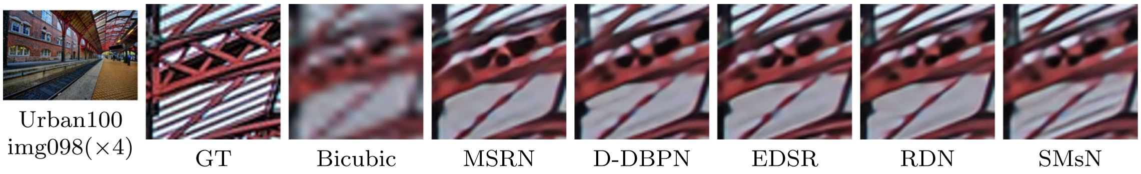

Fig. 7 Qualitative comparison of ×4 SR results for “img098” in Urban100 dataset

4.2 Ablation study

In this part, we set comprehensive experiments to evaluate the effectiveness of (1) Selective Multi-scale Module and (2) Attentive Global Feature Fusion.

SMsM: Firstly, we remove the Selective Kernel Module part in SMsM (Fig. 2) and concatenate features from 3 × 3 and 5 × 5 convolution layers. This modified feature extraction module is adopted as our baseline (*Ms_Baseline*). Also, we replace the $Sigmoid$ operation in SMsM with $Softmax$ and denote this module as *Ms_SKM*. Experimental evaluation of *Ms_Baseline*, *Ms_SKM*, and *SMsM* is conducted on the Plain structure (in Fig. 4a). For convenience, we report results by training only 200 epochs.

Comparative results using different feature extraction modules are summarized in Table 1. By utilizing SKM to fuse multi-scale features from 3 × 3 and 5 × 5 convolutions, *Ms_SKM* surpasses *Ms_Baseline* by 0.13 dB and 0.1 dB on Urban100 dataset for scale factors ×2 and ×4, respectively. Further, we notice that the proposed *SMsM* performs considerably better than *Ms_SKM* via changing the $Softmax$ operation to $Sigmoid$. Such improvement indicates that more flexible attention weights can better depict the inter-dependencies between channel-wise features.

AGFF: We also perform comparative experiments to evaluate the effectiveness of the proposed attentive global feature fusion scheme. Firstly, we stack 64 SMsMs in the manner of Plain structure (in Fig. 4a). Also, global feature fusion (GFF) (in Fig. 4b) is employed as another alternative. The comparative results are shown in Table 2. It is observed that our proposed AGFF achieves more performance gain compared with the Plain and GFF structures.

We also visualize the feature maps before and after AGFF in Fig. 5. Specifically, by following [14], we average the feature maps in the channel dimension in our implementation. It is observed that after redundancy removal and attention-based re-calibration, the features represent more distinctive details (highlighted with red arrows).

4.3 Quantitative and qualitative comparisons

We compare our SMsN model with a number of state-of-the-art SISR methods: LFFN [34], IMDN [14], MSRN [23], D-DBPN [7], EDSR [26], RDN [36] and SRFBN [25]. We also train an enhanced SMsN-L model by stacking $N = 100$ SMsM modules. It is noted that D-DBPN and SRFBN are trained using more training images than other methods (including ours).

Quantitative analysis Table 3 shows quantitative evaluation results (PSNR and SSIM values) on Set5, Set14, B100, Urban100, and Manga109 datasets for scale factors ×2, ×3, and ×4 SISR tasks. It is noted that our proposed SMsN model ranks the best performer in 24 out of the total 30 SISR tasks. Moreover, the enhanced SMsN-L model outperforms other state-of-the-art SISR methods on all datasets for all scale factors.

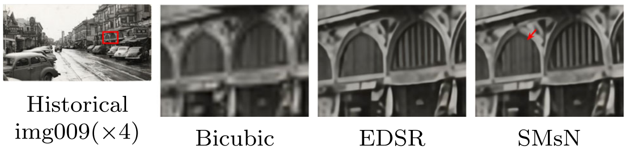

Fig. 8 Qualitative comparison of ×4 SR results on real example ("img009") from Historical dataset

Qualitative analysis Figures 6 and 7 show some qualitative comparisons of our SMsN model and other SISR methods. It is obviously observed that our SMsN model outperforms the others by recovering clearer structures and sharper contours in the super-resolved HR images. We further evaluate our SMsN model using real images (Fig. 8). Our SMsN can recover more details (highlighted with red arrow), when comparing with EDSR [26].

Model size Table 4 shows the PSNR values and model sizes of recent SISR methods. Among these methods, MSRN contains fewer parameters at the cost of an obvious performance drop. Our SMsN uses only 17.7% and 34.3% parameters of EDSR and RDN model but achieves 0.13 dB and 0.15 dB performance gain, respectively. Even better-performed SMsN-L has fewer parameters, when compared with EDSR and RDN. We also take two state-of-the-art methods, i.e., RCAN [35] and SAN [3] into consideration, and it is noted that our SMsN-L can achieve comparable performance with more efficient model. We provide the number of floating-point operations (FLOPs) and inference time of several SISR models in Table 4 as the major indicators of their computational efficiencies. # FLOPs indicates the number of operations by multi-adds, which is the number of composite multiply accumulated operations for processing a 1280 × 720 × 3 RGB image. Our SMsN and SMsN-L have fewer # FLOPs than the others besides MSRN. Note that the inference time is not strictly proportional to # FLOPs, since different operations in PyTorch have different degrees of parallelism.

5 Conclusion

We introduce a compact but powerful SISR model (i.e., SMsN) by designing more effective feature extraction module (i.e., SMsM) and hierarchical feature fusion scheme (AGFF). The proposed method first optimizes selective kernel module (SKM) [24] (replacing $Softmax$ operation with $Sigmoid$) to expand the search space for more distinctive features. Then, an attentive global feature fusion (AGFF) scheme is employed to reduce the redundant information

Table 4 Performance, # parameter, # FLOPs and inference time comparisons with some best performing SISR methods

Methods	MSRN	EDSR	RDN	RCAN	SAN	SMsN	SMsN-L
PSNR	30.57 dB	31.02 dB	31.00 dB	31.21 dB	31.18 dB	31.15 dB	31.22 dB
# Param.	6.08 M	43.09 M	22.27 M	15.59 M	15.86 M	7.63 M	11.99 M
# FLOPs	410.64 G	3216.47 G	1488.78 G	1020.28 G	1040.84 G	473.09 G	736.82 G
Inference Time	0.15 s	0.54 s	0.34 s	0.46 s	0.66 s	0.26 s	0.37 s

The PSNR, inference time are tested on Manga109 dataset with scale factor ×4, and the # FLOPs index is computed on 720P RGB images (1280 × 720 × 3)

inside the extracted hierarchical features and embed the channel attention mechanism to re-calibrate the features before fusion. Our experimental results demonstrate that our proposed SMsM and AGFF are effective for SISR, and our SMsN/SMsN-L performs favorably against state-of-the-art methods while using fewer parameters and FLOPs.

References

1. Agustsson, E., Timofte, R.: NTIRE 2017 Challenge on single image super-resolution: dataset and study. In: CVPR workshop, pp. 126–135 (2017). https://doi.org/10.1109/CVPRW.2017.150
2. Cao, Y., He, Z., Ye, Z., Li, X., Cao, Y., Yang, J.: Fast and accurate single image super-resolution via an energy-aware improved deep residual network. Signal Process. **162**, 115–125 (2019). https://doi.org/10.1016/j.sigpro.2019.03.018
3. Dai, T., Cai, J., Zhang, Y., Xia, S.T., Zhang, L.: Second-order Attention Network for Single Image Super-resolution. In: CVPR, pp. 11065–11074 (2019)
4. Dong, C., Loy, C.C., He, K., Tang, X.: Learning a deep convolutional network for image super-resolution. In: ECCV, pp. 184–199 (2014)
5. Dong, C., Loy, C.C., He, K., Tang, X.: Image Super-Resolution Using Deep Convolutional Networks. IEEE Trans. Pattern Anal. Mach. Intell. **38**(2), 295–307 (2016). https://doi.org/10.1109/TPAMI.2015.2439281
6. Dong, C., Loy, C.C., Tang, X.: Accelerating the super-resolution convolutional neural network. In: ECCV, pp. 391–407 (2016)
7. Haris, M., Shakhnarovich, G., Ukita, N.: Deep back-projection networks for super-resolution. In: CVPR, pp. 1664–1673 (2018)
8. He, K., Zhang, X., Ren, S., Sun, J.: Deep Residual Learning for Image Recognition. In: CVPR, pp. 770–778 (2016)
9. He, Z., Cao, Y., Du, L., Xu, B., Yang, J., Cao, Y., Tang, S., Zhuang, Y.: MRFN: multi-receptive-field network for fast and accurate single image super-resolution. IEEE Trans. Multimed. **22**(4), 1042–1054 (2020). https://doi.org/10.1109/TMM.2019.2937688
10. He, Z., Tang, S., Yang, J., Cao, Y., Ying Yang, M., Cao, Y.: Cascaded Deep Networks With Multiple Receptive Fields for Infrared Image Super-Resolution. IEEE Trans. Circuits Syst. Video Technol. **29**(8), 2310–2322 (2019)
11. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: CVPR, pp. 7132–7141 (2018). https://doi.org/10.1109/CVPR.2018.00745
12. Hu, Y., Li, J., Huang, Y., Gao, X.: Channel-wise and Spatial Feature Modulation Network for Single Image Super-Resolution. IEEE Transactions on Circuits and Systems for Video Technology pp. 1–1 (2019). https://doi.org/10.1109/TCSVT.2019.2915238
13. Huang, G., Liu, Z., Weinberger, K.Q., van der Maaten, L.: Densely Connected Convolutional Networks. In: CVPR, pp. 4700–4708 (2017)
14. Hui, Z., Gao, X., Yang, Y., Wang, X.: Lightweight Image Super-Resolution with Information Multi-distillation Network. In: ACMMM, pp. 2024–2032 (2019)
15. Hui, Z., Wang, X., Gao, X.: Fast and accurate single image super-resolution via information distillation network. In: CVPR, pp. 723–731 (2018)
16. Jiang, T., Zhang, Y., Wu, X., Rao, Y., Zhou, M.: Single image super-resolution via squeeze and excitation network. In: BMVC, pp. 1–11 (2018)
17. Kim, J., Lee, J.K., Lee, K.M.: Accurate Image Super-Resolution Using Very Deep Convolutional Networks. In: CVPR, pp. 1646–1654 (2016)
18. Kim, J., Lee, J.K., Lee, K.M.: Deeply-Recursive Convolutional Network for Image Super-Resolution. In: CVPR, pp. 1637–1645 (2016)
19. Kingma, D.P., Ba, J.L.: Adam: A method for stochastic optimization. In: ICLR, pp. 1–15 (2015)
20. Lai, W.S., Huang, J.B., Ahuja, N., Yang, M.H.: Deep Laplacian Pyramid Networks for Fast and Accurate Super-Resolution. In: CVPR, pp. 624–632 (2017). https://doi.org/10.1109/CVPR.2017.618
21. Lai, W.S., Huang, J.B., Ahuja, N., Yang, M.H.: Deep Laplacian Pyramid Networks for Fast and Accurate Super-Resolution. IEEE Trans. Pattern Anal. Machine Intell. (2019). https://doi.org/10.1109/TPAMI.2018.2865304
22. Ledig, C., Theis, L., Huszar, F., Caballero, J., Cunningham, A., Acosta, A., Aitken, A., Tejani, A., Totz, J., Wang, Z., Shi, W.: Photo-realistic single image super-resolution using a generative adversarial network. In: CVPR, pp. 4681–4690 (2017)
23. Li, J., Fang, F., Mei, K., Zhang, G.: Multi-scale residual network for image super-resolution. In: ECCV, p. In Press (2018)
24. Li, X., Wang, W., Hu, X., Yang, J.: Selective Kernel Networks. In: CVPR, pp. 510–519 (2019)
25. Li, Z., Yang, J., Liu, Z., Yang, X., Jeon, G., Wu, W.: Feedback Network for Image Super-Resolution. In: CVPR, pp. 3867–3876 (2019)
26. Lim, B., Son, S., Kim, H., Nah, S., Lee, K.M.: Enhanced deep residual networks for single image super-resolution. In: CVPR workshop, pp. 136–144 (2017). https://doi.org/10.1109/CVPRW.2017.151
27. Liu, J., Zhang, W., Tang, Y., Tang, J., Wu, G.: Residual feature aggregation network for image super-resolution. In: CVPR, pp. 2359–2368 (2020)
28. Pan, J., Liu, Y., Sun, D., Ren, J., Cheng, M.M., Yang, J., Tang, J.: Image formation model guided deep image super-resolution. In: AAAI, vol. 34, pp. 11807–11814 (2020)
29. Tai, Y., Yang, J., Liu, X.: Image Super-Resolution via Deep Recursive Residual Network. In: CVPR, pp. 3147–3155 (2017). https://doi.org/10.1109/CVPR.2017.298

30. Tai, Y., Yang, J., Liu, X., Xu, C.: MemNet: A Persistent Memory Network for Image Restoration. In: ICCV, pp. 4539–4547 (2017)
31. Timofte, R., Agustsson, E., Gool, L.V., Yang, M.H., Zhang, L., et al.: NTIRE 2017 Challenge on Single Image Super-Resolution: Methods and Results. In: CVPR workshop, pp. 1110–1121 (2017)
32. Tong, T., Li, G., Liu, X., Gao, Q.: Image Super-Resolution Using Dense Skip Connections. In: ICCV, pp. 4799–4807 (2017)
33. Wenzhe, S., Caballero, J., Huszar, F., Totz, J., Aitken1, A.P., Bishop, R., Rueckert, D., Wang, Z.: Real-Time Single Image and Video Super-Resolution Using an Efficient Sub-Pixel Convolutional Neural Network. In: CVPR, pp. 1874–1883 (2016). https://doi.org/10.1109/CVPR.2016.207
34. Yang, W., Wang, W., Zhang, X., Sun, S., Liao, Q.: Lightweight Feature Fusion Network for Single Image Super-Resolution. IEEE Signal Process. Lett. **26**(4), 538–542 (2019). https://doi.org/10.1109/LSP.2018.2890770
35. Zhang, Y., Li, K., Li, K., Wang, L., Zhong, B., Fu, Y.: Image Super-Resolution Using Very Deep Residual Channel Attention Networks. In: ECCV, pp. 286–301 (2018)
36. Zhang, Y., Tian, Y., Kong, Y., Zhong, B., Fu, Y.: Residual Dense Network for Image Super-Resolution. In: CVPR, pp. 2472–2481 (2018)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.