Applied Intelligence (2020) 50:1620–1632
https://doi.org/10.1007/s10489-019-01548-8

Progressive residual networks for image super-resolution

Jin Wan[1] · **Hui Yin**[1] · **Ai-Xin Chong**[2] · **Zhi-Hao Liu**[1]

Published online: 1 February 2020

Abstract

The recent advances in deep convolutional neural networks (DCNNs) have convincingly demonstrated high-capability reconstruction for single image super-resolution (SR). However, it is a big challenge for most DCNNs-based SR models when the scaling factor increases. In this paper, we propose a novel Progressive Residual Network (PRNet) to integrate hierarchical and scale features for single image SR, which works well for both small and large scaling factors. Specifically, we introduce a Progressive Residual Module (PRM) to extract local multi-scale features through dense connected up-sampling convolution layers. Meanwhile, by embedding residual learning into each module, the relative information between high-resolution and low-resolution multi-scale features is fully exploited to boost reconstruction performance. Finally, the scale-specific features are fused to the reconstruction module for restoring the high-quality image. Extensive quantitative and qualitative evaluations on benchmark datasets illustrate that our PRNet achieves superior performance and in particular obtains new state-of-the-art results for large scaling factors such as $4\times$ and $8\times$.

Keywords Image super-resolution · Progressive residual network · Multi-scale features · Residual learning · Deep convolutional neural networks

1 Introduction

Image super-resolution (SR) is one of the most important and challenging tasks in computer vision. It aims to generate a visually pleasing high-resolution (HR) image from a low-resolution (LR) input. This task widely benefits medical imaging, virtual reality, video surveillance, to name a few.

Single image SR is an ill-posed problem as the given LR image loses high-frequency information of the image. The inverse problem of single image SR becomes particularly pronounced when the scaling factor increases. The learning-based approach attempts to solve this ill-posed problem by directly or indirectly learning the mapping between LR and its HR image counterpart.

In recent years, with the help of deep convolutional neural networks (DCNNs), several frameworks for image SR, *e.g.*, [3, 10, 11, 15, 18], were proposed where performance grows rapidly. Dong et al. [3] proposed a Super-Resolution Convolution Neural Network (SRCNN) which firstly used CNN to sample images and achieved significant improvements. The performance of SRCNN was limited by its shallow structure. In [10, 11], Kim et al. increased the depth of network to 20, achieving notable improvements over SRCNN. In order to get higher performance, the network tends to be deeper and deeper. Recently, Lim et al. [18] built a very wide network EDSR and a very deep one MDSR by using simplified residual blocks, which achieved a very satisfactory performance on super-resolution tasks. After that, many super-resolution models of dense connection integration have been proposed to effectively utilize hierarchical features, including SRDenseNet [16] and RDN [32].

Although these latest models have produced promising results by learning deeper hierarchical features, there are

✉ Hui Yin
hyin@bjtu.edu.cn

Jin Wan
17120426@bjtu.edu.cn

Ai-Xin Chong
18112015@bjtu.edu.cn

Zhi-Hao Liu
16120394@bjtu.edu.cn

[1] Beijing Key Laboratory of Traffic Data Analysis and Mining, Beijing Jiaotong University, Beijing 100044, China

[2] Key Laboratory of Beijing for Railway Engineering, Beijing Jiaotong University, Beijing 100044, China

still several bottlenecks. A major issue is that the great majority network models get relatively poor results in large scaling factors such as 8×, as shown in Fig. 1. One of the main reasons is that they may not be able to effectively combine multi-scale feature information with hierarchical feature information for reconstruction.

Our findings For SR, as objects in images have different scales and perspectives, combining the hierarchical features and multi-scale features from a deeper network would give more clues for reconstruction. However, the previous SR methods only gained and incorporated more contextual knowledge through deeper networks and intricate skip connections, such as EDSR [18] and RDN [32], it is difficult to integrate complementary multi-scale feature information using a single stream structure. Further, [28, 33] found that increasing the width of a deep network may be more beneficial than increasing the depth. Therefore, in order to facilitate information integration of the image SR, multi-stream structure and network widening may be effective. Finally, the existing super-resolution multi-scale approach [17] did not fully utilize and fuse multi-scale feature information. And it was not conducive to building and training deeper network structures on limited hardware facilities.

Our Contributions Inspired by these observations and findings, we propose a progressive residual network (PRNet) for still single image SR. Our networks successfully perform on the large scaling factors, as shown in Fig. 1 (PRNet). We summarize our research and contributions as follows:

- We propose a novel framework PRNet for high-quality image SR. The network integrates multi-scale and hierarchical feature information from the original LR image by using multi-stream structure.
- We propose a Progressive Residual Module (PRM) for local multi-scale features representation in PRNet, which can not only extract the deep multi-scale features from the image by progressive structure, but also fully utilize all the different size layers within it via local dense connections. Simultaneously, residual learning is introduced to promote the flow of gradient and information. It is worth mentioning that the larger the scaling factor is, the more multi-scale feature information will be fused by adjusting the width of PRM.
- We propose an effective multi-scale features fusion architecture for image reconstruction in PRNet, which can adaptively fuse global feature information of different scales to improve the performance of image reconstruction.
- Extensive experimental results show that our model achieves state-of-the-art performance on several popular benchmarks. And the larger the scaling factor becomes, the more highlighted the superiority of PRNet will be, which is due to its specially designed structure.

The remainder of this paper is organized as follows. Section 2 reviews the related works. In Section 3, we elaborate on the proposed methods. Experimental results and analysis are reported in Section 4. Finally, Section 5 summarizes our work.

2 Related work

Single image super-resolution (SR) researches can be divided into three classes: interpolation-based [31], reconstruction-based [30] and learning-based [3, 16, 22–24, 27]. The most popular one is learning-based, including neighbor embedding [6], sparse coding [25, 29] and random forest [21]. Methods in this class learn the complex mapping relationship between LR and HR by using large training datasets. As an implementation of learning-based approaches, SRCNN [3] employed a three-layer

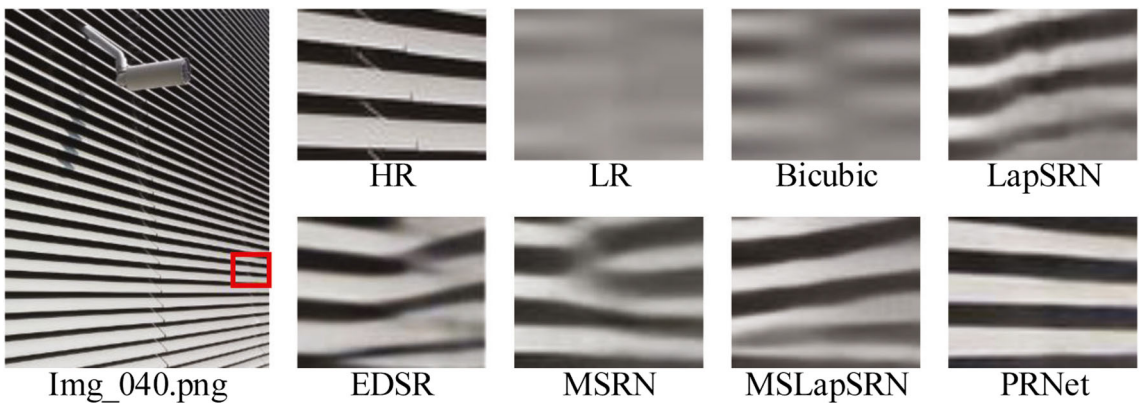

Fig. 1 Visual comparison for 8x super-resolution on Urban100 [9]. PSNR: Bicubic (15.89 dB), LapSRN [15] (18.27 dB), EDSR [18] (19.53 dB), MSRN [17] (19.41 dB), MSLapSRN [14] (18.52 dB), and PRNet (ours)(**20.16 dB**). (Zoom in for best view)

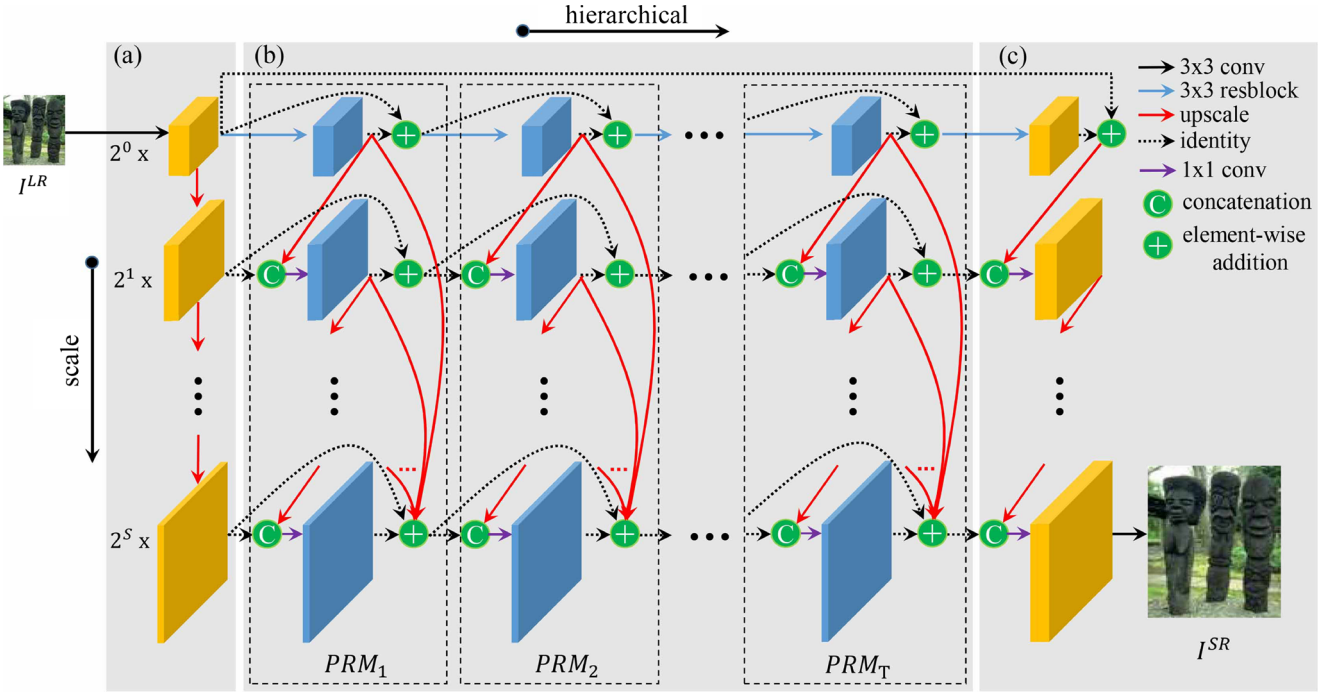

Fig. 2 Illustrating the architecture of the proposed progressive residual network. **a** Initial multi-scale features extraction net (MSFENet). **b** Progressive residual modules (PRMs). **c** Reconstruction net with multi-scale features fusion (ReconNet). The horizontal and vertical directions correspond to the hierarchical depth of the network and the scale of the feature maps, respectively. Upper-right legend: 3 × 3 resblock = a ReLU between two 3 × 3 convolutions, upscale = pixelshuffle with different up-sampling scale factors, identity = no operation

full convolutional neural network to learn end-to-end mapping between low/high resolution images, which was the first successful attempt to apply CNN to SR problems. To expand the field of perception, DRCN [11] and VDSR [10] increased the depth of network through skip connections and residual learning. Those works often used pre-processed LR images as input, which was upscaled to HR space via an up-sampling operator, such as bicubic. However, ESPCN [22] has been proved that it will increase computational complexity and produce visible artifacts.

To solve this, many deep neural network models [16, 18, 23] were introduced to take advantage of the hierarchical features of the original low-resolution (LR) image, while the original low-resolution (LR) image did not use preprocessing. Among these methods, the most well-known one is EDSR [18] which was the champion of the NTIRE2017 SR Challenge. It is based on SRResNet [16] to enhance performance by removing the normalization layer and using deeper and broader network structures.

Further, to make full use of the information across several layers or blocks, the pattern of multiple or dense skip connections between layers or modules is adopted in SR. Inspired by densely connected convolutional networks (DenseNet [8]) which achieved high performance in image classification, [13, 27] utilized the DenseNet structure as building modules to reuse learnt feature maps and introduced dense skip connections to fuse features at different levels. Compared with SRDenseNet [27], RDN [32] used a residual dense connection method that extracted and fused multi-level features from the original LR image to further improve the performance.

The aforementioned methods showed impressive performance of super-resolution, while most of them lose some useful hierarchical features and ignored the multi-scale features from the original LR image. Meanwhile, we believe that as the scaling factor is amplified, more multi-scale information should be integrated. Although Li et al. [17] introduced two fixed-size convolution kernels (3 × 3 and 5 × 5) to detect image features of different scales, it is impossible to adaptively select multi-scale features based on the different scaling factors. To resolve these cases, we propose progressive residual network (PRNet) to integrate hierarchical features and multi-scale features from all the layers in the LR space efficiently. We will detail our PRNet in the next section.

3 Progressive residual networks

In this section, we describe the design methodology of the proposed PRNet. Firstly, we introduce the network architecture of PRNet in Section 3.1. Then, Sections 3.2

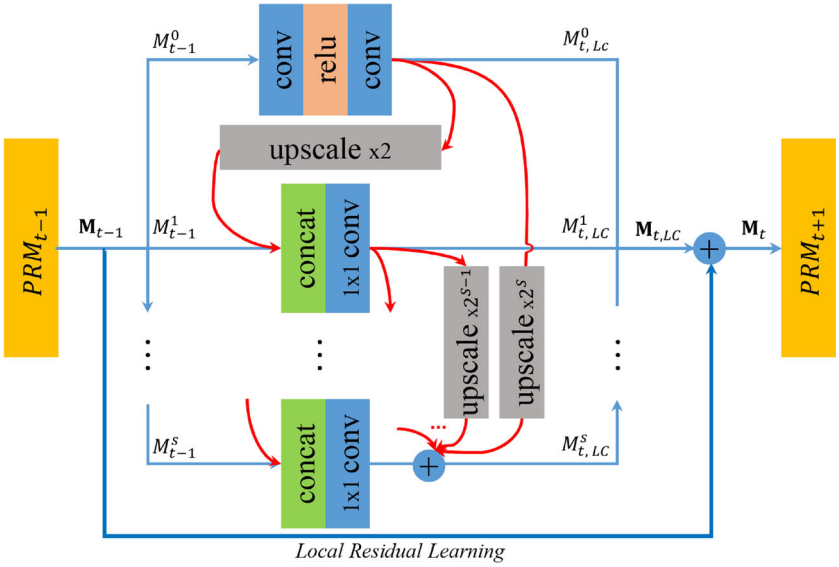

Fig. 3 Progressive residual module (PRM) architecture

3.3 detail two main parts: progressive residual module and reconstruction net with multi-scale features fusion.

3.1 Network architecture

In PRNet, the aim is to estimate a super-resolution image I^{SR} from a low-resolution input image I^{LR}. Correspondingly, we use I^{HR} to denote the high-resolution image. I^{LR} is obtained by performing the down-sampling operation. Figure 2 shows the architecture of our proposed PRNet. It consists of three parts: initial multi-scale features extraction net (MSFENet), multiple stacked progressive residual modules (PRMs) and finally a reconstruction net with multi-scale features fusion (ReconNet).

Specifically, a convolutional layer and several up-sampling layers are used in MSFENet to extract the initial multi-scale shallow features $\mathbf{M}_0 = \{M^s_0, s = 0, 1, 2, 3\}$. M^0_0 is extracted from I^{LR} and the formula is as follows:

$$M^0_0 = \mathcal{F}_{Ext}(I^{LR}) \,, \tag{1}$$

where $\mathcal{F}_{Ext}(\cdot)$ denotes 3×3 convolution operation. M^0_0 is then used for further multi-scale features extraction and global residual learning. So we can further have

$$M^s_0 = \mathcal{F}_{Up}(M^{s-1}_0) \,, \tag{2}$$

where $\mathcal{F}_{Up}(\cdot)$ denotes $2\times$ up-sampling operation used in [18, 22], $s \in \{1, 2, 3\}$, and s corresponds to the sampling factor ($s = i$ for the scaling factor is 2^i). $\mathbf{M}_0$ are the extracted inital multi-scale features to be sent to the first progressive residual module (PRM).

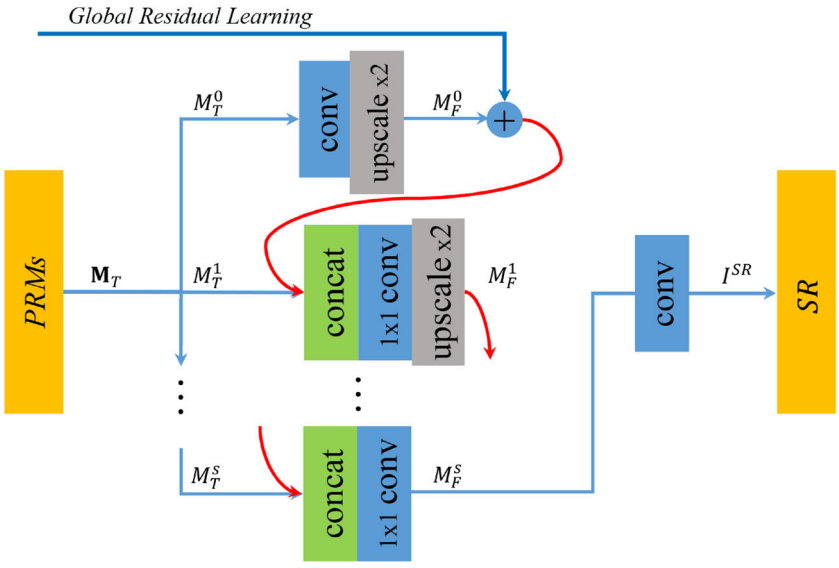

Fig. 4 Reconstruction net with multi-scale features fusion

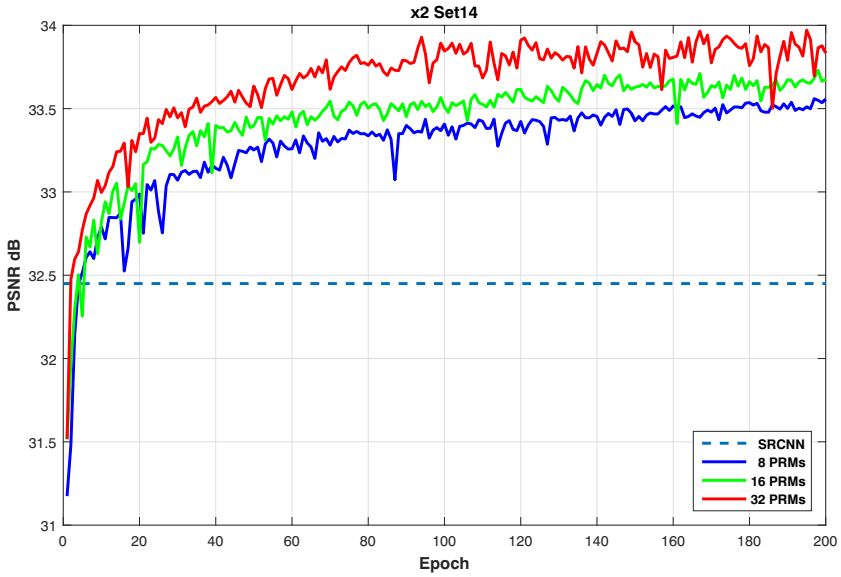

Fig. 5 Convergence analysis of PRNet with different number of PRM

Supposing T progressive residual modules are stacked to act as the feature mapping, the output $\mathbf{M}_t$ of the $t-$th PRM can be obtained by:

$$\begin{aligned}\mathbf{M}_t &= \mathcal{F}_{Prn,t}(\mathbf{M}_{t-1}) \\ &= \mathcal{F}_{Prn,t}(\mathcal{F}_{Prn,t-1}(\ldots(\mathcal{F}_{Prn,1}(\mathbf{M}_0))\ldots)), \quad t \in 1,\ldots,T, \end{aligned} \tag{3}$$

where $\mathcal{F}_{Prn,t}(\cdot)$ denotes the operations of the t-th PRM. $\mathcal{F}_{Prn,t}(\cdot)$ can be a composite function of operations, such as convolution and rectified linear units (ReLU).

Finally, our model uses the up-sampling and convolution layers in ReconNet to fuse multi-scale features and reconstruct residual images. Therefore, our PRNet can be formulated as:

$$\begin{aligned}I^{SR} &= \mathcal{F}_{PRNet}(I^{LR}) \\ &= \mathcal{F}_{Rec}(\mathcal{F}_{Prn,t}(\mathcal{F}_{Prn,t-1}(\ldots(\mathcal{F}_{Prn,1}(\mathbf{M}_0))\ldots)) \\ &\quad + M_0^0)\,, \end{aligned} \tag{4}$$

where $\mathcal{F}_{PRNet}(\cdot)$ and $\mathcal{F}_{Rec}(\cdot)$ denote the function of our PRNet and the reconstruction net with multi-scale features fusion respectively.

Given a training set $\{I_i^{HR}, I_i^{LR}\}_{i=1,\ldots,N}$, where N is the number of training patches and I_i^{HR} is the ground truth of the low-quality patch I_i^{LR}, the loss function of our PRNet with the parameter set $\boldsymbol{\Theta}$, is

$$\mathcal{L}^{SR}(\boldsymbol{\Theta}) = \frac{1}{N}\sum_{i=1}^{N}\|\mathcal{F}_{PRNet}(I_i^{LR}) - I_i^{HR}\|_1. \tag{5}$$

3.2 Progressive residual module

In order to synthesize local hierarchical features and local multi-scale features, we propose a progressive residual module (PRM), as show in Fig. 3. Here we will present more details of this structure, which contains local cross-scale feature fusion (LSCFF) and local residual learning (LRL).

Local cross-scale feature fusion is exploited to adaptively fuse the states from preceding PRM and fully utilize hierarchical information and multi-scale information in current PRM. Different from previous studies, we construct

Table 1 Ablation investigation of local residual learning (LRL), local cross-scale feature fusion (LCSFF), and multi-scale features fusion (MSFF)

Different combinations of LRL, LCSFF and MSFF								
LRL	×	✓	×	×	×	✓	✓	✓
LCSFF	×	×	✓	×	✓	×	✓	✓
MSFF	×	×	×	✓	✓	✓	×	✓
PSNR	31.34	33.43	33.39	33.34	33.45	33.56	33.70	33.73

We observe the best performance (PSNR) on Set14 [29] with scaling factor ×2 in 200 epochs

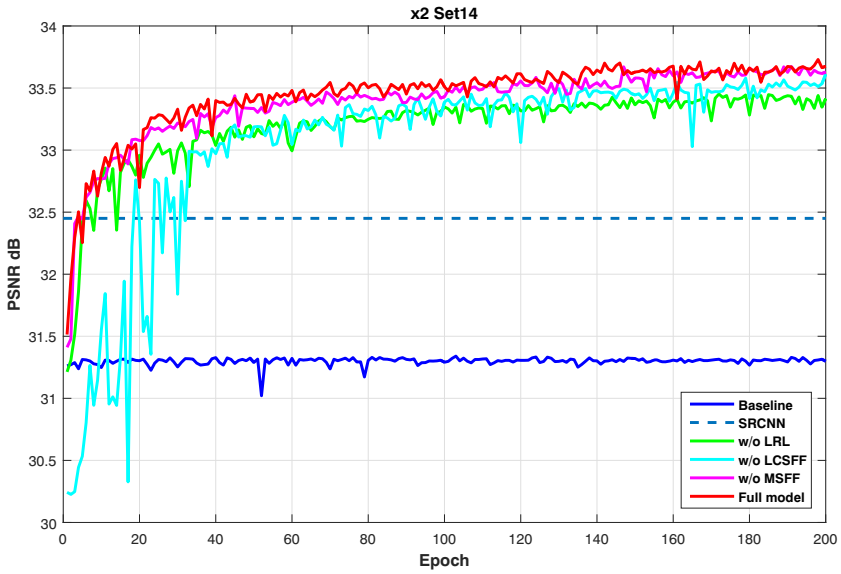

Fig. 6 Convergence analysis on MCSFF, LRL, and MSFF. The curves for each combination are based on the PSNR on Set14 [29] with scaling factor $\times 2$ in 200 epochs

a multi-stream sub-network in which complementary multi-scale feature information can be detected and fused on different streams. In this way, the hierarchical and multi-scale features of the deep network can be integrated to provide more clues for image reconstruction. The operation can be formulated as:

$$\begin{aligned} M^0_{t,LC} &= p_t * \boldsymbol{\sigma}(q_t * M^0_{t-1}), \\ M^1_{t,LC} &= W^1_t * [M^1_{t-1}, (M^0_{t,LC})\uparrow_2], \\ &\cdots\cdots \\ M^s_{t,LC} &= W^s_t * [M^s_{t-1}, (M^{s-1}_{t,LC})\uparrow_2] \\ &\quad +((M^{s-2}_{t,LC})\uparrow_{2^2} + \ldots + (M^0_{t,LC})\uparrow_{2^s}) , \end{aligned} \tag{6}$$

$$\mathbf{M}_{t,LC} = \{M^s_{t,LC}, s = 0, 1, 2, 3\}, \tag{7}$$

where $*$ is the spatial convolution operator, $\uparrow_{2^s}$ is the up-sampling operator with scaling factor 2^s, and $p_t, q_t, \mathbf{W}_t = \{W^s_t, s = 1, 2, 3\}$ are convolutional layers at stage t. $\sigma(*) = max(0, x)$ stands for the ReLU function, and $[x, y]$ denotes the concatenation operation with x and y.

Local residual learning is applied to further facilitate the information flow. Formally, we define LRL as:

$$\mathbf{M}_t = \mathbf{M}_{t,LC} + \mathbf{M}_{t-1}, \tag{8}$$

where $\mathbf{M}_{t-1}$ and $\mathbf{M}_t$ represent the input and output of the PRM, respectively. The operation $+$ is performed by a shortcut connection and element-wise addition. LRL not only makes the computational complexity greatly reduced, but also promotes the performance of the network.

3.3 Reconstruction net with multi-scale features fusion

After extracting local features at different scales with a set of PRMs, we further propose a multi-scale features fusion structure for reconstruction, as shown in Fig. 4, which can adaptively fuse global multi-scale feature information. It consists of global residual fusion (GRF) and multi-scale features fusion (MSFF).

Global residual learning is introduced to gain the feature-maps before conducting up-sampling by

$$M^0_F = h_f * M^0_T + M^0_0, \tag{9}$$

where M^0_0 denotes the initial shallow feature-maps, M^0_T denotes the zeroth element of $\mathbf{M}_T$, which is the output of the last PRM module of PRNet, $*$ is the spatial convolution operator, and h_f is the convolutional layer at the global residual stage. All the other layers before global feature fusion are fully exploited with our proposed progressive residual modules (PRMs). PRM generates multi-level local features of different sizes, which are further adaptively fused into the image up-sampling process to enhance the reconstruction performance.

Multi-scale features fusion is then utilized to adaptively fuse global feature information of different scales in the up-

Table 2 Quantitative evaluation of state-of-the-art SR algorithms: average PSNR / SSIM for scale factors 2×, 3×, 4× and 8×

Dataset	Scale	Set5 PSNR/SSIM	Set14 PSNR/SSIM	BSD100 PSNR/SSIM	Urban100 PSNR/SSIM	Manga109 PSNR/SSIM
Bicubic	×2	33.69 / 0.931	30.25 / 0.870	29.57 / 0.844	26.89 / 0.841	30.86 / 0.936
SRCNN [3]		36.72 / 0.955	32.51 / 0.908	31.38 / 0.889	29.53 / 0.896	35.76 / 0.968
SelfExSR [9]		36.60 / 0.955	32.24 / 0.904	31.20 / 0.887	29.55 / 0.898	35.82 / 0.969
FSRCNN [4]		37.05 / 0.956	32.66 / 0.909	31.53 / 0.892	29.88 / 0.902	36.67 / 0.971
VDSR [10]		37.53 / 0.959	33.05 / 0.913	31.90 / 0.896	30.77 / 0.914	37.22 / 0.975
DRRN [23]		37.74 / 0.959	33.23 / 0.914	32.05 / 0.897	31.23 / 0.919	37.92 / 0.976
SRDCNN [13]		37.26 / 0.957	32.69 / 0.899	31.55 / 0.891	- / -	- / -
LapSRN [15]		37.52 / 0.959	33.08 / 0.913	31.80 / 0.895	30.41 / 0.910	37.27 / 0.974
EDSR [18]		38.11 / 0.960	33.92 / 0.920	32.32 / 0.901	32.93 / 0.935	39.10 / 0.977
MSRN [17]		38.08 / 0.961	33.74 / 0.917	32.23 / 0.901	32.22 / 0.933	38.82 / 0.977
D-DPBN [7]		38.09 / 0.960	33.85 / 0.919	32.27 / 0.900	32.55 / 0.932	38.89 / 0.978
RDN [32]		38.24 / 0.961	34.01 / 0.921	32.34 / 0.902	32.89 / 0.935	39.18 / 0.978
PRNet (ours)		38.24 / 0.961	34.02 / 0.921	32.35 / 0.902	32.93 / 0.936	39.12 / 0.978
PRNet+ (ours)		38.30 / 0.962	34.09 / 0.922	32.40 / 0.903	33.14 / 0.937	39.34 / 0.979
Bicubic	×3	30.41 / 0.869	27.55 / 0.775	27.22 / 0.741	24.47 / 0.737	26.99 / 0.859
SRCNN [3]		32.78 / 0.909	29.32 / 0.823	28.42 / 0.788	26.25 / 0.801	30.59 / 0.914
SelfExSR [9]		32.66 / 0.910	29.18 / 0.821	28.30 / 0.786	26.45 / 0.810	27.57 / 0.821
FSRCNN [4]		33.18 / 0.914	29.37 / 0.824	28.53 / 0.791	26.43 / 0.808	31.10 / 0.921
VDSR [10]		33.67 / 0.921	29.78 / 0.832	28.83 / 0.799	27.14 / 0.829	32.01 / 0.934
DRRN [23]		34.03 / 0.924	29.96 / 0.835	28.95 / 0.800	27.53 / 0.764	32.74 / 0.939
SRDCNN [13]		33.59 / 0.923	29.54 / 0.824	28.80 / 0.797	- / -	- / -
LapSRN [15]		33.82 / 0.922	29.87 / 0.832	28.82 / 0.798	27.07 / 0.828	32.21 / 0.935
EDSR [18]		34.65 / 0.928	30.52 / 0.846	29.25 / 0.809	28.80 / 0.865	34.17 / 0.948
MSRN [17]		34.38 / 0.926	30.34 / 0.840	29.08 / 0.804	28.08 / 0.855	33.44 / 0.943
D-DPBN [7]		- / -	- / -	- / -	- / -	- / -
RDN [32]		34.71 / 0.930	30.57 / 0.847	29.26 / 0.809	28.80 / 0.865	34.13 / 0.948
PRNet (ours)		34.67 / 0.929	30.57 / 0.847	29.25 / 0.809	28.77 / 0.865	34.05 / 0.948
PRNet+ (ours)		34.79 / 0.930	30.65 / 0.848	29.31 /0.810	28.96 / 0.868	34.37 / 0.950
Bicubic	×4	28.43 / 0.811	26.01 / 0.704	25.97 / 0.670	23.15 / 0.660	24.93 / 0.790
SRCNN [3]		30.50 / 0.863	27.52 / 0.754	26.91 / 0.712	24.53 / 0.725	27.66 / 0.859
SelfExSR [9]		30.34 / 0.862	27.41 / 0.753	26.84 / 0.713	24.83 / 0.740	27.83 / 0.866
FSRCNN [4]		30.72 / 0.866	27.61 / 0.755	26.98 / 0.715	24.62 / 0.728	27.90 / 0.861
VDSR [10]		31.35 / 0.883	28.02 / 0.768	27.29 / 0.726	25.18 / 0.754	28.83 / 0.887
DRRN [23]		31.68 / 0.888	28.21 / 0.772	27.38 / 0.728	25.44 / 0.764	29.46 / 0.896
SRDCNN [13]		31.16 / 0.879	27.85 / 0.764	27.08 / 0.709	- / -	- / -
LapSRN [15]		31.54 / 0.885	28.19 / 0.772	27.32 / 0.727	25.21 / 0.756	29.09 / 0.890
EDSR [18]		32.46 / 0.897	28.80 / 0.788	27.21 / 0.742	26.64 / 0.803	31.02 / 0.915
MSRN [17]		32.07 / 0.890	28.60 / 0.775	27.52 / 0.727	26.04 / 0.790	30.17 / 0.903
D-DPBN [7]		32.47 / 0.898	28.82 / 0.786	27.72 / 0.740	26.38 / 0.795	- / -
RDN [32]		32.47 / 0.899	28.81 / 0.787	27.72 / 0.742	26.61 / 0.803	31.00 / 0.915
PRNet (ours)		32.49 / 0.899	28.86 / 0.788	27.74 / 0.742	26.68 / 0.805	31.15 / 0.917
PRNet+ (ours)		32.63 / 0.900	28.95 / 0.790	27.81 / 0.744	26.87 / 0.808	31.52 / 0.920

Table 2 (continued)

Bicubic	×8	24.40 / 0.658	23.10 / 0.566	23.67 / 0.548	20.74 / 0.516	21.47 / 0.650
SRCNN [3]		25.33 / 0.690	23.76 / 0.591	24.13 / 0.566	21.29 / 0.544	22.46 / 0.695
SelfExSR [9]		25.49 / 0.703	23.92 / 0.601	24.19 / 0.568	21.81 / 0.577	22.99 / 0.719
FSRCNN [4]		25.60 / 0.697	24.00 / 0.599	24.31 / 0.572	21.45 / 0.550	22.72 / 0.692
VDSR [10]		25.93 / 0.724	24.26 / 0.614	24.49 / 0.583	21.70 / 0.571	23.16 / 0.725
DRRN [23]		26.18 / 0.738	24.42 / 0.622	24.59 / 0.587	21.88 / 0.583	23.60 / 0.742
LapSRN [15]		26.15 / 0.738	24.35 / 0.620	24.54 / 0.586	21.81 / 0.581	23.39 / 0.735
EDSR [18]		26.97 / 0.775	24.94 / 0.640	24.80 / 0.596	22.47 / 0.620	24.58 / 0.778
MSRN [17]		26.59 / 0.725	24.88 / 0.596	24.70 / 0.541	22.37 / 0.598	24.28 / 0.752
MSLapSRN [14]		26.34 / 0.756	24.57 / 0.627	24.65 / 0.590	22.06 / 0.596	23.90 / 0.756
D-DPBN [7]		27.21 / 0.784	25.13 / 0.648	24.88 / 0.601	22.73 / 0.631	- / -
PRNet (ours)		27.18 / 0.784	25.15 / 0.648	24.89 / 0.602	22.77 / 0.633	24.97 / 0.793
PRNet+ (ours)		27.34 / 0.788	25.27 / 0.652	24.98 / 0.604	22.95 / 0.640	25.25 / 0.799

Red indicates the best and blue indicates the second best performance

sampling process. We name this operation as multi-scale features fusion (MSFF) formulated as:

$$\begin{aligned} M_F^1 &= W_f^1 * [M_T^1, (M_F^0)\uparrow_2] \,, \\ &\cdots\cdots \\ M_F^s &= W_f^s * [M_T^s, (M_F^{s-1}))\uparrow_2] \,, \end{aligned} \tag{10}$$

$$I^{SR} = r_f * M_F^s, \tag{11}$$

where $*$ is the spatial convolution operator, $\uparrow_2$ is the up-sampling operator with scaling factor 2, and $\mathbf{W}_f = \{W_f^s, s = 1, 2, 3\}$ and r_f are convolutional layers at the multi-scale features fusion stage. $[x, y]$ denotes concatenation operation with x and y.

4 Experiments

In this section, we first describe the implementation and training details of our network. We then validate the contributions of different components in the proposed network and compare the proposed PRNet with several state-of-the-art SR methods on benchmark datasets. We present the quantitative evaluation and qualitative comparison. Finally, we apply our approach to real photos.

4.1 Implementation and training details

Datasets and Metrics In our work, we choose DIV2K [1] as the training dataset, a new high-quality image dataset for image restoration challenge. DIV2K consists of 800 training images, 100 validation images, and 100 test images. We train all of our models with 800 training images and use 14 validation images in the training process. For testing, we use five standard benchmark datasets: Set5 [2], Set14 [29], B100 [19], Urban100 [9], and Manga109 [20]. These datasets contain a wide variety of images that can fully verify our model. Following previous works, all testing is based on luminance channel in YCbCr space, and scaling factors: ×2, ×3, ×4, and ×8 are used for training and testing.

Training Setting Following settings of [18], in each training batch, we randomly extract 16 LR RGB patches with the size of 48×48 as inputs. We randomly augment the patches by flipping and rotating before training. To maintain the image details, instead of transforming the RGB patches into a YCbCr space, we use the 3-channel image information from the RGB for training. The entire network is optimized by Adam [12] with L_1 loss by setting $\beta_1 = 0.9$ and $\beta_2 = 0.999$. The learning rate is initially set to 10^{-4} and halved at every 2×10^{-5} mini-batch updates for 3×10^{-5} total mini-batch updates. All experiments are conducted using Pytorch [5], MATLAB R2015b on NVIDIA TITAN Xp GPUs.

4.2 Model analysis

In this section, we first analyze the effects of the number of PRM, which is closely related to the depth of the network. We then use ablation experiments to evaluate several key design methodologies of PRNet.

PRMs and Network depth We investigate the basic network parameters: the number of PRM (denote as T for short).

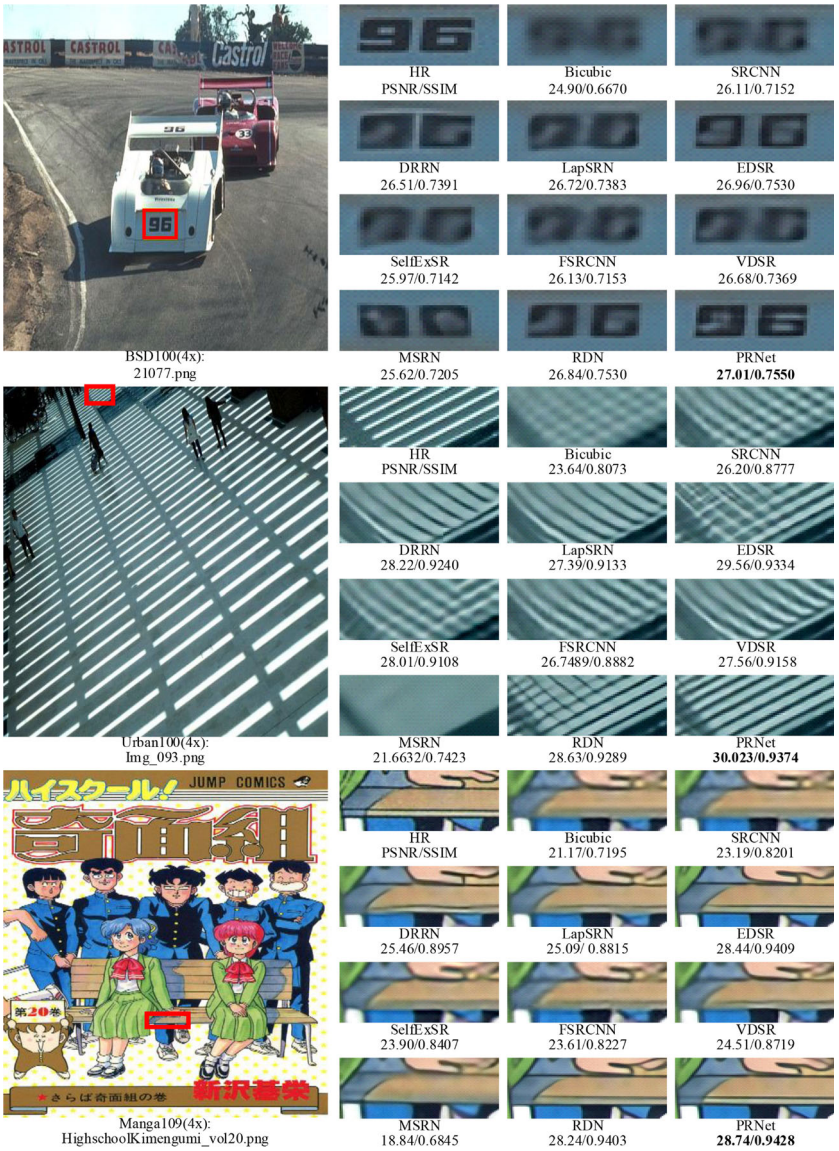

Fig. 7 Visual comparison for 4× SR on BSD100 [19], Urban100 [9], and Manga109 [20] datasets. The best results are highlighted. (Zoom in for best view)

We use the performance of SRCNN [3] as a reference. For illustration purpose, we train the proposed model with different number of PRM, that is the different depth of whole PRNet. We choose $T = 8, 16, 32$. As shown in Fig. 5, larger T would lead to higher performance. This is mainly because the network becomes deeper with larger T. On the other hand, although PRNet with smaller T would suffer some performance degradation during training, it still outperforms SRCNN [3] due to its powerful feature representation and fusion capabilities. More importantly, PRNet allows deeper and wider network, with bigger T and s, where more hierarchical features and more multi-scale features are extracted for higher performance.

Ablation Investigation Table 1 shows the ablation investigation on the effects of local cross-scale feature fusion (LCSFF) , local residual learning (LRL), and multi-scale features fusion (MSFF). The eight networks have the same number of PRMs ($T = 16$). The baseline is obtained without LCSFF, LRL, or MSFF and performs very poorly (PSNR = 31.34 dB). This is due to the difficulty of training and also demonstrates that stacking many basic blocks in a very deep network does not yield better performance.

We then add one of the LRL, LCSFF or MSFF to the baseline (the 2^{nd} to 4^{th} combinations in Table 1). The results prove that each component can greatly improve the performance of the baseline. We further remove one of LRL, LCSFF, or MSFF from PRNet to verify the validity of each component design. The quantitative results in Table 1 show that each component can significantly improve the performance of the network.

This is because LRL can promote the flow of information and gradients. LCSFF can fully combine multi-scale features and hierarchical features in feature extraction. MSFF

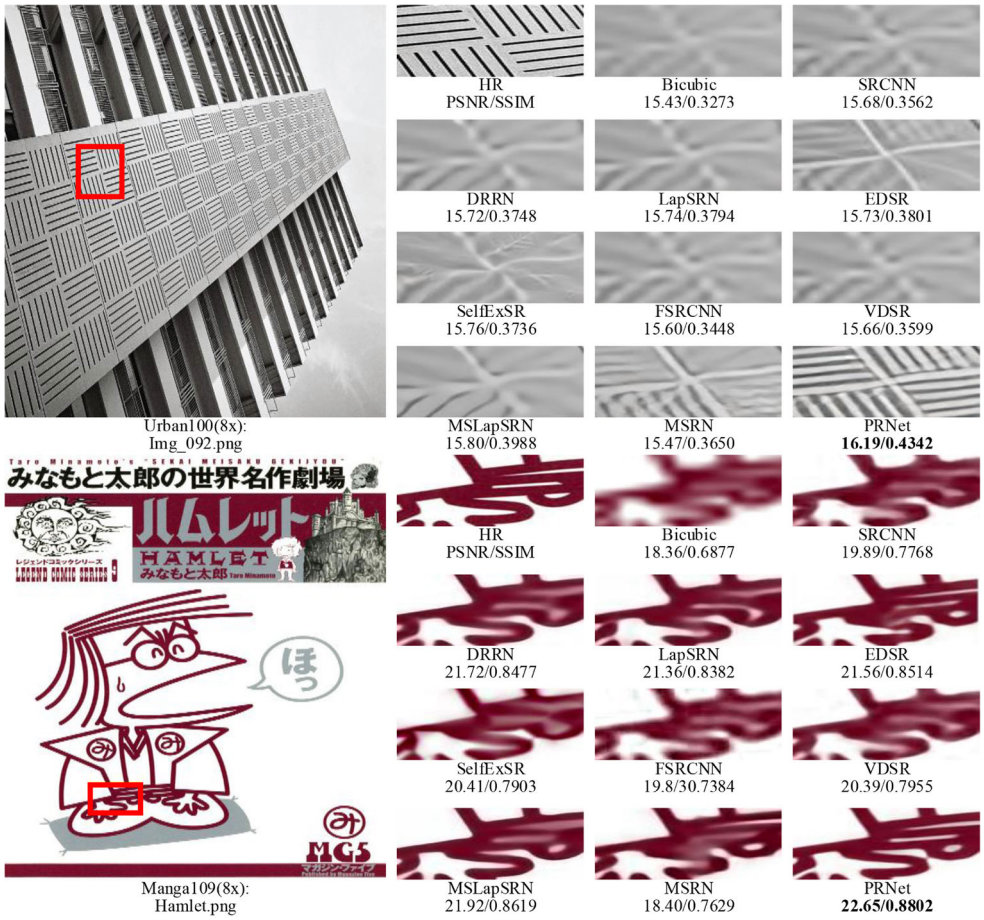

Fig. 8 Visual comparison for 8× SR on Urban100 [9] and Manga109 [20] datasets. The best results are highlighted. (Zoom in for best view)

can effectively fuse multi-scale features in reconstruction. A similar phenomenon can be seen when we use these three components simultaneously (denote as the full model). PRNet using three components performs the best.

We can also visualize the convergence process of the above combinations in Fig. 6. The convergence curves are consistent with our analyses and indicate that LCSFF, LRL or MSFF can stabilize the training process without obvious performance degradation. These quantitative and visual analysis prove the benefits and effectiveness of the proposed LCSFF, LRL or MSFF.

4.3 Comparisons with state-of-the-art methods

To confirm the ability of the proposed network, we perform several experiments and analysis. We compare our network with 11 state-of-the-art SR algorithms: SRCNN [3], SelfExSR [9], FSRCNN [4], VDSR [10], DRRN [23], SRDCNN [13], LapSRN [15], EDSR [18], MSRN [17], D-DBPN [7], and RDN [32]. Similar to [18, 32], we also adopt the self-ensemble strategy [26] to further improve our PRNet and denote the self-ensembled PRNet as PRNet+.

Table 2 shows quantitative comparisons for ×2, ×3, ×4, and ×8 SR. It is worth noted that D-DBPN [7] has to divide each image in Manga109 into four parts and then calculate PSNR separately which will significantly improve the super-resolution performance. For a fair comparison, we do not compare the results of D-DBPN [7] on the Manga109 dataset in the 8× up-sampling sampling factor, and the result of other datasets are cited from their paper.

When compared with all previous methods, our PRNet+ performs the best on all the datasets with all scaling factors. Even without self-ensemble, our PRNet also achieves the best average results on most datasets. Specifically, for the scaling factor ×3, our PRNet would not hold a similar advantage over RDN [32]. This is mainly because we adjust the number of feature map channels of our network (252 and 28). However, our results are still better than the rest of the models. Moreover, we have better applicability than D-DBPN [7]. For the scaling factor ×2, ×4, and ×8, our PRNet performs the best on all datasets. On the other hand, when the scaling factor becomes larger (e.g., ×8), the gains of our PRNet over EDSR [18] also becomes larger.

We also provide visual comparison results as qualitative comparisons. Figure 7 shows the visual comparisons on the 4× scale. For image "21077" and "HighschoolKimengumi_vol 20", we observe that most of the compared methods would produce blurred artifacts and distorted

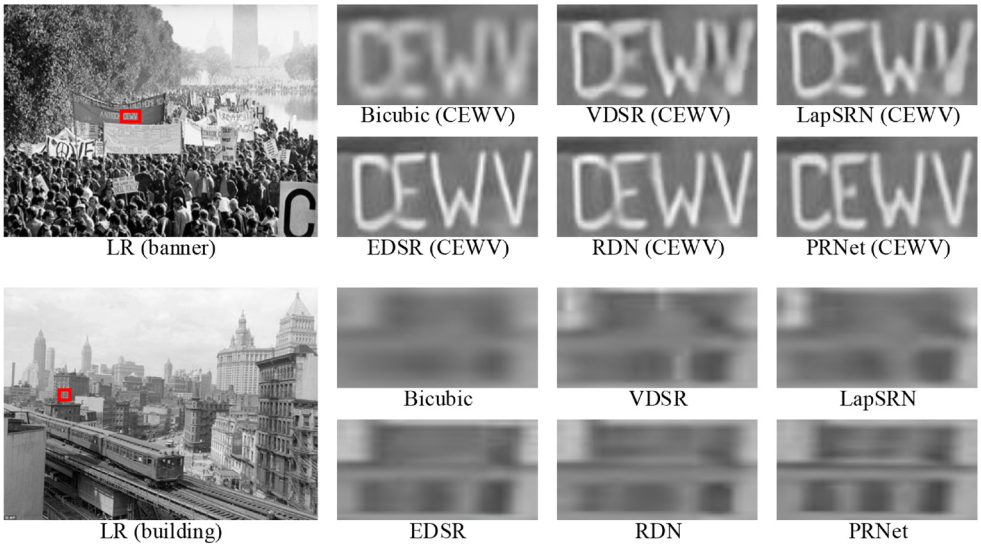

Fig. 9 Visual results on real-world images with scaling factor x4. The two rows show SR results for images "banner" and "building" respectively. "CEWV" is the ground-truth of "banner" and denotes the "Committee to End the War in Vietnam". (Zoom in for best view)

edges. By contrast, our PRNet can restore sharper and more natural edges, and produce more faithful results. For the visually farther texture in the image "img_093", all methods of comparison can't reconstruct it correctly. While our PRNet can obviously reconstruct it. This is mainly because PRNet uses multi-scale feature information through multi-scale features fusion.

To further illustrate the analysis above, we show visual comparisons for 8× SR in Fig. 8. For image "img_092", due to the large scaling factor, the results of Bicubic would lose details and produce an incorrect texture structure. This false pre-amplification result would also lead some state-of-the-art methods (*e.g.*, SRCNN, VDSR, and DRRN) to reconstruct totally erroneous structures. Even the original LR as input, other methods cannot reconstruct the right structure either. While, our PRNet can recover them correctly and clearly. Similar observations are shown in image "Hamlet". Our proposed PRNet can integrate multi-scale and hierarchical feature information to enhance the ability of feature representation and improve the performance of reconstruction.

4.4 Super-resolving real-world photos

We also conduct SR experiments on two historical real-world images, "banner" (with 400 × 270 pixels) and "building" (with 400 × 327 pixels). In this case, the original HR images are not available and the degradation model is unknown either. We compare our PRNet with VDSR [10], LapSRN [15], EDSR [18], and RDN [32]. As shown in Fig. 9, our PRNet recreates finer details and more loyal to real-world scenarios than other state-of-the-art methods. These results further indicate the benefits of learning multi-scale features from the original input image. Combining the hierarchical features and multi-scale features performs robustly for unknown degradation models.

5 Conclusion

In this paper, we propose a Progressive Residual Network (PRNet) for image SR, where progressive residual module (PRM) serves as the basic build module. In each PRM, dense connected up-sampling convolution layers allow full usage of local multi-scale features. The local residual leaning (LRL) further improves the flow of information and gradient. Moreover, we propose the multi-scale features fusion (MSFF) to fuse multi-scale features extracted from previous PRMs during reconstruction. By fully using local and global multi-scale features, our PRNet leads to a dense fusion of hierarchical and scale features. We use the same PRNet structure to handle the bicubic degradation model and real-world data. Extensive benchmark evaluations demonstrate that our PRNet yields superior results and successfully outperforms other state-of-the-art methods on large scaling factors such as 4× and 8× enlargement.

Acknowledgments This work is supported by National Nature Science Foundation of China (61472029, 51827813, 61473031), National Key R&D Program of China (2017YFB1201104, 2016YFB1200100), and Scientific Research Project of Beijing Educational Committee (SM20191001107; PXM 2019_014213_000007).

References

1. Agustsson E, Timofte R (2017) Ntire 2017 challenge on single image super-resolution: Dataset and study. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp 126–135

2. Bevilacqua M, Roumy A, Guillemot C, AlberiMorel ML (2012) Low-complexity singleimage super-resolution based on nonnegative neighbor embedding. In: Proceedings of the 23rd British Machine Vision Conference, pp 1–10
3. Dong C, Loy CC, He K, Tang X (2014) Learning a deep convolutional network for image super-resolution. In: Proceedings of European conference on computer vision, pp 184–199
4. Dong C, Loy CC, Tang X (2016) Accelerating the super-resolution convolutional neural network. In: Proceedings of European conference on computer vision, pp 391–407
5. Facebook (2017) Pytorch. https://pytorch.org/
6. Gao X, Zhang K, Tao D, Li X (2012) Image super-resolution with sparse neighbor embedding. IEEE Trans Image Process 21(7):3194–3205
7. Haris M, Shakhnarovich G, Ukita N (2018) Deep back-projection networks for super-resolution. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, pp 1664–1673
8. Huang G, Liu Z, Maaten LVD, Weinberger KQ (2017) Densely connected convolutional networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 4700–4708
9. Huang JB, Singh A, Ahuja N (2015) Single image super-resolution from transformed self-exemplars. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, pp 5197–5206
10. Kim J, Kwon Lee J, Mu Lee K (2016) Accurate image super-resolution using very deep convolutional networks. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, pp 1646–1654
11. Kim J, Kwon Lee J, Mu Lee K (2016) Deeply-recursive convolutional network for image super-resolution. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, pp 1637–1645
12. Kingma D, Ba J (2014) Adam: a method for stochastic optimization. In: Proceedings of International Conference on Learning Representations
13. Kuang P, Ma T, Chen Z, Li F (2019) Image super-resolution with densely connected convolutional networks. Appl Intell 49(1):125–136
14. Lai W, Huang J, Ahuja N, Yang M (2018) Fast and accurate image super-resolution with deep laplacian pyramid networks. IEEE Transactions on Pattern Analysis and Machine Intelligence, pp 1–1
15. Lai WS, Huang JB, Ahuja N, Yang MH (2017) Deep laplacian pyramid networks for fast and accurate super-resolution. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, pp 624–632
16. Ledig C, Theis L, Huszar F, Caballero J, Cunningham A, Acosta A, Aitken A, Tejani A, Totz J, Wang Z, Shi W (2017) Photo-realistic single image super-resolution using a generative adversarial network. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, pp 4681–4690
17. Li J, Fang F, Mei K, Zhang G (2018) Multi-scale residual network for image super-resolution. In: Proceedings of European Conference on Computer Vision, pp 517–532
18. Lim B, Son S, Kim H, Nah S, Mu Lee K (2017) Enhanced deep residual networks for single image super-resolution. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp 136–144
19. Martin D, Fowlkes C, Tal D, Malik J (2001) A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In: Proceedings of IEEE International conference on computer vision, pp 416–423
20. MatsuiY IK, Aramaki Y, Fujimoto A, Ogawa T, Yamasaki T, Aizawa K (2017) Sketch-based manga retrieval using manga109 dataset. Multimedia Tools and Applications, pp 21811–21838
21. Schulter S, Leistner C, Bischof H (2015) Fast and accurate image upscaling with super-resolution forests. In: The IEEE Conference on Computer Vision and Pattern Recognition, pp 3791–3799
22. Shi W, Caballero J, Huszar F, Totz J, Aitken AP, Bishop R, Rueckert D, Wang Z (2015) Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, pp 1874–1883
23. Tai Y, Yang J, Liu X (2017) Image super-resolution via deep recursive residual network. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, pp 3147–3155
24. Tai Y, Yang J, Liu X, Xu C (2017) Memnet: A persistent memory network for image restoration. In: Proceedings of IEEE International Conference on Computer Vision, pp 4539–4547
25. Timofte R, De Smet V, Van Gool L (2013) Anchored neighborhood regression for fast example-based super-resolution. In: Proceedings of IEEE International Conference on Computer Vision, pp 1920–1927
26. Timofte R, Rothe R, Van Gool L (2016) Seven ways to improve example-based single image super-resolution. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, pp 1865–1873
27. Tong T, Li G, Liu X, Gao Q (2017) Image super-resolution using dense skip connections. In: Proceedings of IEEE International Conference on Computer Vision, pp 4799–4807
28. Zagoruyko S, Komodakis N (2017) Diracnets: Training very deep neural networks without skip-connections. arXiv:170600388
29. Zeyde R, Elad M, Protter M (2012) On single image scale-up using sparse-representations. In: Curves and Surfaces, pp 711–730
30. Zhang K, Gao X, Tao D, Li X (2012) Single image super-resolution with non-local means and steering kernel regression. IEEE Trans Image Process, pp 4544–4556
31. Zhang L, Wu X (2006) An edge-guided image interpolation algorithm via directional filtering and data fusion. IEEE Trans Image Process, pp 2226–2238
32. Zhang Y, Tian Y, Kong Y, Zhong B, Fu Y (2018) Residual dense network for image super-resolution. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, pp 2472–2481
33. Zhao L, Li M, Meng D, Li X, Zhang Z, Zhuang Y, Tu Z, Wang J (2018) Deep convolutional neural networks with merge-and-run mappings. In: Proceedings of the 27th International Joint Conference on Artificial Intelligence, pp 3170–3176

Jin Wan received the B.E. degree from Changchun University of Science and Technology in 2017. He is currently pursuing the Ph.D. degree with the School of Computer and Information Technology, Beijing Jiaotong University, Beijing, China. His research interests are in machine learning, pattern recognition, image processing and algorithms.

Ai-Xin Chong received his B.Sc. degree from Shandong Agricultural University in 2017. Now, he is a Ph.D. candidate in Beijing Jiaotong University. His main research interests include computer stereo vision and pattern recognition.

Hui Yin received the Ph.D. degree in computer application technology from Beijing Jiaotong University, Beijing, China. She is currently a Full Professor of the School of Computer and Information Technology, Beijing Jiaotong University. Her current research interests include the machine vision, intelligent information processing and their application in the railway industry.

Zhi-Hao Liu received the B.E. degree from Beijing Institute of Graphic Communication, Beijing, China, in 2016. He is currently pursuing the Ph.D. degree with the School of Computer and Information Technology, Beijing Jiaotong University, Beijing, China. His current research interests include computer vision and pattern recognition.