Neurocomputing 358 (2019) 285–293

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Detail-preserving image super-resolution via recursively dilated residual network

Feng Li[a,b], Huihui Bai[a,b,*], Yao Zhao[a,b]

[a] Institute of Information Science, Beijing Jiaotong University, Beijing 100044, China
[b] Beijing Key Laboratory of Advanced Information Science and Network Technoloy, Beijing 100044, China

ARTICLE INFO

Article history:
Received 30 October 2018
Revised 27 April 2019
Accepted 15 May 2019
Available online 17 May 2019

Communicated by Dr Yan Bo

Keywords:
Image super-resolution
Spatial modulated dilated residual block
Contextual information
Image detail

ABSTRACT

Convolutional neural network (CNN) methods have been successfully applied in single image super-resolution (SR). However, existing very deep CNN based SR methods face with the challenge of memory footprint and computational complexity for real-world applications. Besides, many previous methods lack flexible ability to emphasize local spatial informative areas, which is limited to recover the high-frequency detail of LR input. In this paper, to address these problems, we implement a spatial modulated residual unit (SMRU) upon the dilated residual unit and propose a recursively dilated residual network (RDRN) to reconstruct high-resolution (HR) images from low-resolution (LR) observations. The proposed RDRN can effectively exploit the contextual information over larger regions and pay attention to the high-frequency parts for image detail recovery. Furthermore, such spatial modulation mechanism (SPM) in SMRU can incorporate well with existing SR models for better reconstruction performance. Extensive evaluations on public benchmark datasets demonstrate that our proposed method achieves superior performance in terms of quantitative and qualitative assessments.

1. Introduction

The objective of single image super-resolution (SR) is to recover a visual-pleasant high-resolution (HR) image from a given low-resolution (LR) image. Since that image SR can overcome the limitation of image resolution in a small scale, it has been applied in various applications, such as medical imaging [1], face recognition [2], satellite imaging [3], and other fields. The image SR is an essentially ill-posed problem since there are multiple solutions existed to reconstruct HR images from LR ones with non-invertible operations. Most recent methods typically resolve this inverse problem by incorporating various image priors to constrain the solution space. Tai et al. [4] propose an approach that reconstructs edges while recovering the image detail by adding learning-based detail synthesis to edge-directed image SR in a mutually consistent framework. In [5], Zhang et al. present a non-local kernel regression method for image and video restoration tasks, which exploits both the non-local self-similarity and local structural regularity in natural images.

To reveal the high-frequency detail of reconstructed HR images, recently image SR methods employ example-based approaches. Yang et al. [6] propose a fast image SR method based on the regression on in-place examples, which utilizes two fundamental SR approaches of learning from external-examples and self-examples. In [7], Huang et al. introduce a self-similarity driven SR algorithm, which can effectively increase the size of the limited internal dictionary without any external training images. Timofte et al. [8] address the problem of image upscaling based on a dictionary of low- and high-resolution exemplars, which combines the best quality of the anchored neighborhood regression and simple functions for image SR. By exploiting the connection between sparse coding based approaches and local linear regression, Schulter et al. [9] present a new approach for image SR via random forests. To produce super-resolved LR images with better objective quality, Song et al. [10] introduce a gradient field sharpening transform that converts the blurry gradient field of upsampled LR image to a much sharper gradient field of original HR image. Inspired by the success of convolutional neural networks (CNN) in computer vision field, Dong et al. [11] firstly propose to directly learn an end-to-end mapping between low- and high-resolution images via fully convolutional neural network for image SR (SR-CNN), which shows notable superior accuracy compared with previous example-based methods. The authors further replaces the

* Corresponding author at: Institute of Information Science, Beijing Jiaotong University, Beijing 100044, China.
E-mail address: hhbai@bjtu.edu.cn (H. Bai).

https://doi.org/10.1016/j.neucom.2019.05.042
0925-2312/

bicubic upsampling operation with one deconvolutional layer at the end of the network for fast image SR (FSRCNN) [12], which achieves a high acceleration without the loss of restoration quality. Kim et al. [13] improve the performance over SRCNN via very deep convolutional network with residual learning (VDSR). To ease the difficulty of training very deep network and control the model parameters, Kim et al. [14] further propose a deeply-recursive convolutional network (DRCN), which uses recursive-supervisions and skip connections to improve the performance of predicted HR images. Recently, the major trend of CNN based image SR algorithms [15–17,17–19] is stacking more layers to boost the SR performance. Mao et al. [15] introduce a deep convolutional encoder-decoder network which uses symmetrically lines to combine the convolutional and deconvolutional layers with skip-layer connections for image restoration. In [17], Tai et al. propose a deep recursive residual network (DRRN), which recursively learns the residual image to control the model parameters while increasing the depth (up to 52 layers) of the networks. Zhang et al. [20] especially propose a 400-layers channel-wise attention based residual network (RCAN) for accurate image SR.

Due to the contextual information spreading over very large regions, the information contained in small patches are usually not sufficient to restore the high-frequency detail. It is necessary to expand the receptive field of networks to exploit the contextual information over larger regions. One option is to increase the kernel size of convolutional layers to enlarge the receptive field. Nevertheless, simply increasing the filter size can involve more weight parameters and computational cost. In addition, according to existing popular networks [21,22], stacking more 3×3 convolutional layers can incorporate more non-linear rectification layers to make the model more discriminative. As mentioned above, another option is stacking more convolutional layers to obtain large receptive field. Unfortunately, utilizing too many convolutional layers can also inevitably increase the parameters and demand more memory space. Inspired the recursive learning in [14], there are some methods [17,23,24] combine the recursive learning strategy with deep networks to increase the depth without introducing additional weight parameters. Motivated by the characteristic that dilated convolution can expand the receptive field while keeping the small kernel size of the standard convolution, which means that we can use relative fewer convoltional layers to effectively exploit the contextual information over larger regions. Besides, there are different types of information across the space in LR inputs and deep features, which have different contributions for high-frequency detail recovery. Most previous CNN-based methods [13,17] treat the LR features equally in networks and lack flexible ability to emphasize the local spatial informative areas. Although generative adversarial networks (GAN) [19] can help to recover a photo-realistic HR image, sometimes the local detail in HR images may not always consistent with the LR images.

To solve the drawbacks mentioned above, we propose a deep recursively dilated residual network (RDRN) for fast and accurate image SR. Our proposed RDRN mainly consists of a recursion in recursion module (RIR), a detail refinement module (DRM) and an upscale module. The RIR module is composed of multiple cascading spatial modulated residual units (SMRU) implemented upon the dilated residual unit(DRU). The SMRU first employs two dilated convolutional layer to extract the input features with larger receptive field and then introduces a spatial modulated unit (SPU) to model the contextual information over local representations within each feature map. Besides, recursive learning is adopted in these SMRUs to efficiently reuses the weight parameters while increasing the depth of networks. Then, after the effective feature extraction via RIR, we employ a DRM composed of multiple convolutional layers to refine the local detail of input features for highly accurate image SR. After that, an upscale module is adopted to aggregate the obtained the residual representations to generate HR residual images. Finally, we implement an element-wise addition operation on the HR residual images and the bicubic amplified LR images to generate the SR results.

In summary, the main contributions of this work can be summarized as follows:

- We propose a novel deep recursively dilated residual network (RDRN) to effectively exploit the contextual information over larger regions and emphasize meaningful features for fast and accurate image SR via recursive and residual learning schemes.
- In the proposed RDRN, we present a RIR module composed of multiple SMRUs, which combines the dilated convolution and spatial modulation mechanism (SPM) to extract more high-frequency features. The recursive learning in RIR can efficiently train our dilated network in deeper visions without introducing additional weight parameters.
- We present a spatial modulated unit (SPU), which incorporates the SPM with the dilated residual unit to model the contextual information over local representations within each feature map. Such SPM can incorporates well with exising CNN based models to obtain better SR results. Our method demonstrates superior SR performance in common benchmarks compared with many state-of-the-art methods.

The rest parts of this paper are organized as follows. Section 2 briefly introduces related work in image SR. Section 3 illustrates the detail of our proposed RDRN. The experimental results on several benchmarks are presented in Section 4. The conclusion of this paper is summarized in Section 5.

2. Related work

2.1. Deep learning based image SR

Deep learning methods have been widely used in computer vision tasks including image classification [21,22], image segmentation [25], and object detection [26]. Recent image SR methods tend to build end-to-end CNN models to learn the mapping function from LR to HR images using large training datasets. Dong et al. [11] first introduce a three-layer CNN to conduct the feature extraction and learn the non-linear mapping function between LR and HR patches for image SR. Kim et al. [13] adopt the residual learning and deeper layers to improve the reconstruction accuracy. The authors [14] further additionally utilize the recursive learning and multi-path skip connections to boost the SR performance. Therefore, most CNN based methods employ different kinds of residual learning strategy [16–18,27] to achieve remarkable improvement in image SR. Motivated by the densely connected network (DenseNet) [28], some methods [23,29,30] utilize the densely-liked connections to fully exploit the hierarchical features for image SR.

Although the increase in depth can enlarge the receptive field and improve the discriminative capacity for learning more complex LR-to-HR mappings, this approach dramatically suffers from the enormous parameters and harder training process. There are some methods [14,17,24] adopt the recursive learning strategy to mitigate the difficulty of training very deep networks and control the model parameters. DRCN [14] introduces a recursive layer into the network and repeatedly applies the same convolutional layer to reduce the number of parameters. Tai et al. [17] introduce the global residual learning (GRL) and local residual learning (LRL) in DRRN to solve the gradients vanishing/exploding problem. The GRL learns the residual image from the input and output of the networks. The LRL is utilized to carry the rich features to later layers.

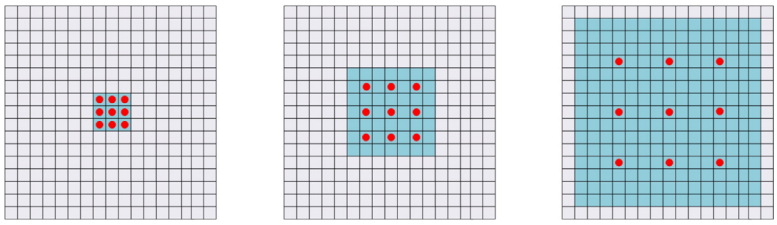


Fig. 1. The receptive fields of dilted convolution with 3×3 kernel size and different rates. (a). Standard convolution corresponds to dilated convolution with $r = 1$. (b) The receptive field of 7×7 produced by the dilated convolution with $r = 2$. (c). The receptive field of 15×15 produced by the dilated convolution with $r = 4$.

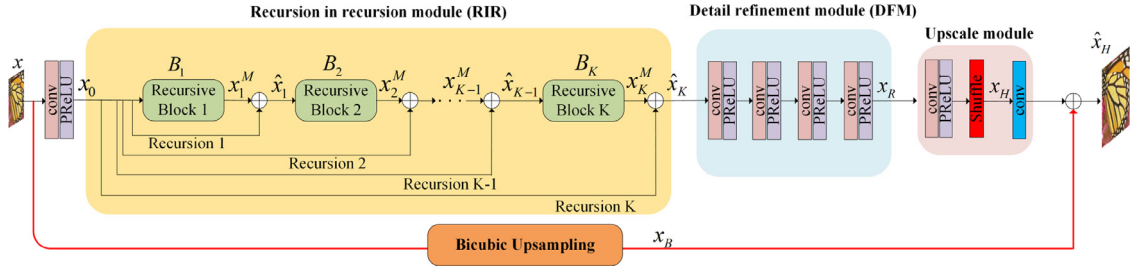


Fig. 2. The overall framework of our proposed RDRN for image SR, which is mainly composed of three components: a recursion in recursion (RIR) module, a detail refinement module (DRM) and an upscale module.

2.2. Dilated convolution

The main idea of dilated convolution is to insert "holes" between pixels in convolutional kernels to exponentially expand the receptive fields without losing resolution in deep CNNs, such as image classification [25], image segmentation [31–33]. For a convolutional filter with the size of $k \times k$ and dilation rate r, the size of resulted dilated convolutional filter is $k_d \times k_d$, where $k_d = k + (k-1) \times (r-1)$. As shown in Fig. 1, a dilated convolution with filter size of 3×3 and $r = 1$ followed by a dilated convolution with the same kernel and $r = 2$ can produce the receptive field of 7×7 with identical parameters, which demonstrates that we can adopt different dilation rates to enlarge the receptive field while keeping the merits of small filters. The standard convolution corresponds to the dilated convolution with $r = 1$. From Fig. 1, we can find that the number of parameters associated with each layer is identical. The receptive field grows exponentially while the number of parameters grows linearly.

To preserve the spatial resolution in deep CNNs for image classification, Yu et al. [25] propose a dilated residual network (DRN) upon the original residual network (ResNet) [22], which uses dilated convolution to increase the receptive field of higher layers. The authors remove the downsampling layers of the two final groups in ResNet and replace these layers by dilated convolutional layers. The converted DRN has the same number of parameters and layers as the ResNet yet with higher resolution output and more accurate classification. For image SR, there are also some methods [34–36] employ dilated convolution to effectively expand the receptive field of networks without additional computational complexity and memory consumption.

3. Proposed method

In this section, we first provide a global view of the proposed network RDRN. Then, we elaborate each component of the proposed RDRN, which is simply illustrated in Fig. 2. The proposed RDRN mainly consists of three components: a recursion in recursion module (RIR), a detail refinement module (DRM) and an upscale module. Finally, we introduce the loss function for training our RDRN.

3.1. Overview

Let $x \in \mathbb{R}^{H \times W \times C}$ denotes the LR input image, and $y \in \mathbb{R}^{sH \times sW \times C}$ denotes its original corresponding HR image. The degradation process of y can be formulated

$$x = (D \otimes y) \downarrow_s + n \tag{1}$$

where D denotes various blurring degradations and $\downarrow_s$ denotes the bicubic downsampling operation with scale factor s. n is the additional noise. Our goal is to restore the HR image from an observed LR image x. Therefore, as illustrated in Fig. 2, our proposed RDRN takes the LR image x as input and predict the HR image

$$\hat{x}_H = F_{SR}(x) \tag{2}$$

where $F_{SR}(\cdot)$ denotes the mapping function for LR to HR images of RDRN. $\hat{x}_H \in \mathbb{R}^{sH \times sW \times C}$ is the estimated HR image.

3.2. Recursion in recursion module

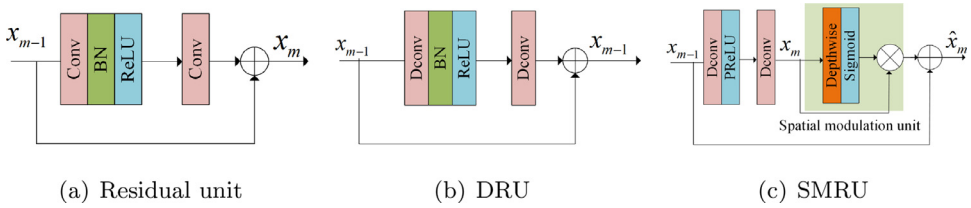
Given an LR input x, we first use a standard convolutional layer with the kernel size of 3×3 and stride 1 to extract the LR features

$$x_0 = f_e(x) \tag{3}$$

where $f_e(\cdot)$ denotes the feature extraction function and x_0 is the extracted feature fed into the RIR module.

3.2.1. Spatial modulated residual unit

The LR input and features contains low-frequency parts and high-frequency areas. The high-frequency areas usually contains abundant edge or texture detail, while the smooth areas have more low-frequency information. These types of features have different contributions for the high-frequency detail recovery. Therefore, in the RIR module, we present a spatial modulated residual unit (SMRU) which is implemented on a dilated residual unit (DRU) [25]. The start point of DRU is from the residual unit in ResNet



(a) Residual unit (b) DRU (c) SMRU

Fig. 3. The convertion of the residual unit into DRU and the SMRU in RDRN. (a). The residual unit consists of two standard convolutional layers in ResNet. (b) The converted DRU which consists of two dilated convolutional layers in RDN, where the "Dconv" denotes dilated convolution. (c) The proposed SMRU in our RDRN, which consists of two dilated convolutional layers and a spatial modulation unit.

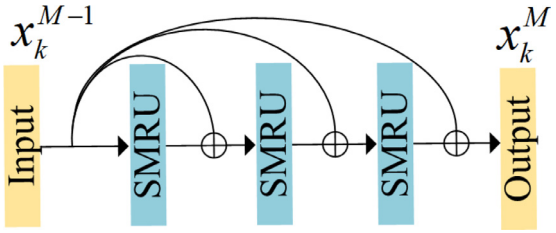


Fig. 4. The structure of the k^{th} recursive block which consists of multiple SRMUs in the proposed RIR module. Here, we simply depict three SMRUs in the recursive block.

[22]. The converted DRN has the same number of parameters and layers as the ResNet yet with higher resolution output. The original residual unit is depicted in Fig. 3(a). Fig. 3(b) simply illustrates the conversion from the residual unit to DRU. In our proposed SMRU, we first use two dilated convolutional layers to exploit the contextual information with larger receptive field and small kernel size 3×3

$$x_m = f_m^2(\tau(f_m^1(x_{m-1}))) \tag{4}$$

where x_{m-1} denotes the output of the $(m-1)^{th}$ SMPU and serves as the input of the m^{th} SMRU. $f_m(\cdot)$ and $f_{m-1}(\cdot)$ represent the convolution operation of the two dilated layers, respectively. We use the parametric rectified linear unit (PReLU) [37] as activation function $\tau(\cdot)$.

From the viewpoint of image SR, the importance of the channels varies by the spatial regions. Each feature map has different sense depending on the convolutional filters. In the case of edges and textures, those channels from complex filters perform more important role for image detail recovery. In order to make our proposed network pay more attention to informative regions, we introduce a spatial modulation mechanism (SPM) which uses a depthwise convolution to capture the spatial relations. The depthwise convolution applies a single filter to each input channel to modulate the spatial information within each feature map. Then, as shown in Fig. 3(c), we employ a sigmoid function to normalized the modulated features between a range from 0 to 1, and then we conduct the spatial-wise multiplication with the features fed into the spatial modulated unit (SPU). Finally, we adopt local residual learning to obtain the final output $\hat{x}_m$ of the SMRU

$$\hat{x}_m = x_{m-1} + \sigma(f_m^3(x_m)) \otimes x_m \tag{5}$$

where $f_m^3(\cdot)$ denotes the depthwise convolution operation and $\sigma(\cdot)$ denotes the sigmoid function. $\otimes$ represents the element-wise multiply operation.

3.2.2. Recursive block

In the RIR module, we share the weight parameters within each SMRU. Specifically, as sketched in Fig. 4, we form M SMRU as a recursive block and employ the shared-source skip connections to carry rich image detail to late layers and help gradient flow. Therefore, for the k^{th} recursive block, we have

$$\begin{aligned} x_k^M &= \mathcal{B}_k(x_{k-1}^M) \\ &= \mathcal{R}_k^M(\mathcal{R}_k^{M-1}(...(\mathcal{R}_k^2(\mathcal{R}_k^1(x_{k-1}^M)))...)) \end{aligned} \tag{6}$$

where x_{k-1}^M and x_k^M denotes the output of the $(k-1)^{th}$ and k^{th} recursive block. $\mathcal{R}_k^M(\cdot)$ denotes the mapping function of the M^{th} SMRU in the k^{th} recursive block. $\mathcal{B}_k(\cdot)$ denotes the mapping function of the k^{th} block. Supposing there are K recursive blocks in the RIR module, the first block B_1 is denoted as recursion 1. As shown in Fig. 2, the former recursion is employed as an unit stacked with current block to make up a new recursion, which means recursion in recursion (RIR). Therefore, the output $\hat{x}_K$ of our proposed RIR module can be represented as

$$\begin{aligned} \hat{x}_K &= f_{RIR}(x_0) = x_K^M + x_0 \\ &= \mathcal{B}_K(\hat{x}_{K-1}) + x_0 \\ &= \mathcal{R}_K^M(\mathcal{R}_K^{M-1}(...(\mathcal{R}_K^2(\mathcal{R}_K^1(\hat{x}_{K-1})))...)) + x_0 \end{aligned} \tag{7}$$

where $f_{RIR}(\cdot)$ denotes the mapping function of the RIR module and $\hat{x}_{K-1}$ is the output of the $(M-1)^{th}$ recursion.

3.3. Detail refinement module

Since dilated convolution pads zeros between two pixels in the convolutional kernel, according to [25], the receptive field can only cover the area with checkerboard pattern, which can cause gridding artifacts. To address this problem, we design a detail refinement module (DRM) to fine-tune the detail of the features produced by RIR. As illustrated in Fig. 2, considering that skip connections can propagate the gridding artifacts from RIR module to later layers, in our proposed DRM, 4 standard convolutional layers are simply stacked without any skip connection to solve this problem

$$x_R = f_D(\hat{x}_K) \tag{8}$$

where $f_D(\cdot)$ denotes the function of DRM, and x_R is the output of DRM.

3.4. Upscale module

After obtaining the refine LR residual features, we stack an upscale module in the HR space. In the upscale module, we utilize the sub-pixel magnification algorithm in [38] to increase the resolution of the feature maps. Then we adopt a convolutional layer with 3 output channels to reconstruct the HR residual image. Finally, we conduct the global residual learning between the generated HR residual image and the bicubic amplified LR input to obtain the final SR results $\hat{x}_H$. This process can be described as follows:

$$\begin{aligned} x_H &= \mathcal{PS}(f_u^1(x_R)) \\ \hat{x}_H &= f_u^2(x_H) + x_B \end{aligned} \tag{9}$$

where $f_u^1(\cdot)$ denotes the mapping function of the convolutional layer with $H \times W \times s^2C$ channels before the pixel shuffle operation.

The pixel shuffle operator $\mathcal{PS}$ is used to rearrange the elements of a $H \times W \times s^2C$ tensor to a tensor $x_H \in \mathbb{R}^{sH \times sW \times C}$. $f_u^2(\cdot)$ represents the convolution operation of the final reconstruction layer. x_B is bicubic upsampled HR image from the LR input x.

3.5. Training

Given a training set $[x_i, \tilde{x}_i]_{i=1}^N$, where N is the number of training patches and $\tilde{x}_i$ is the ground truth of the LR patch x_i. We train our RDRN with L1 loss. The loss function of the proposed network with the parameter set Θ is

$$L(\Theta) = \frac{1}{N}\sum_{i=1}^{N} \|F_{SR}(x_i) - \tilde{x}_i\|_1 = \frac{1}{N}\sum_{i=1}^{N} \|\hat{x}_H - \tilde{x}_i\|_1 \quad (10)$$

4. Experiments

In this section, we evaluate the performance of our model on several datasets. We first describe the datasets which are used for training and testing our models. Next, we introduce the implementation detail in our experiments. Then we investigate the effectiveness of different components in our proposed RDRN for image SR. Finally, we compare the proposed RDRN with several state-of-the-art methods on subjective assessment, objective assessment, and inference time.

4.1. Datasets

As for training, following [18], we use a high-quality (2K resolution) dataset DIV2K [39] as our training data. DIV2K includes 800 images training images, 100 validation images and 100 testing images. The training images is augmented by flipping horizontally or vertically, randomly rotating 90°, 180°, and 270°, and scaling in a range from 0.6 to 0.9. For testing, we evaluate our proposed RDRN algorithm on four public benchmark datasets: *Set5* [40], *Set14* [41], *BSD100* [42], and *Urban100* [7]. The SR results are evaluated with two subjective metrics: PSNR [43] and SSIM [44] on Y channel (*i.e.*, luminance) of transformed YCbCr space and compare the performance on $2\times$, $3\times$, and $4\times$ SR.

4.2. Implementation detail

In our proposed RDRN, except the upscale module, all of the standard convolutional layers and dilated convolutional layers consist of 64 filters with the kernel size of 3×3 and stride 1. In the SMRU, we set 3×3 as the size of depthwise convolutional layer. In the upscale module, the first convolutional layer consists $64 \times s \times s$ filters corresponding the scale factor s with the kernel size of 3×3 and stride 1. The final reconstruction convolutional layer has 3 filters, as we output color images. To produce the LR images, we use the bicubic interpolation to downscale the original HR images by MATLAB *imresize* function. In each training batch, we randomly extract 20 LR RGB patches with the size of 64×64 as the input fed into our proposed network. We train our model using the Adam optimizer [45] with $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\epsilon = 10^{-8}$. For weight initialization, we use the method introduced in He et al. [22]. The initial learning rate is set to 10^{-4} for all layers and decreases to half every 200 epochs. We use PyTorch [46] on a NVIDIA Titan Xp GPU for training and testing our models.

4.3. Ablation study

In this section, we first investigate the effectiveness of dilation schemes using in our models for image SR. Then, we investigate another two basic network parameters: the number of recusions K and the number of SMRUs M in each recursion block. Finally, we study the effect of the proposed SPM, RIR, and DRM in the SMRU.

Table 1
Results of different variant dilation rate schemes with a scale factor of 4. The **Text** indicates the best performance.

Dilation scheme	Receptive field	*Set5*	*Set14*
1-2-1	65×65	31.35 / 0.883	28.03 / 0.767
1-2-3	97×97	31.39 / 0.883	28.05 / **0.768**
1-3-4	129×129	**31.41** / **0.884**	**28.08** / **0.768**
1-4-8	235×235	31.22 / 0.880	27.96 / 0.765
RDRN-ND	49×49	31.31 / 0.882	28.01 / 0.766

Table 2
Results of RDRN with different values of K and M for $2\times$ SR. The **text** indicates the best performance.

Structure	Parameters	PSNR / SSIM
K2M6	456K	37.47 / 0.957
K3M4	535K	37.58 / 0.959
K3M6	535K	37.65 / 0.960
K4M4	677K	37.67 / 0.960
K3M8	535K	**37.73** / **0.961**

4.3.1. Effectiveness of dilation rate

We conduct experiments on the impact of receptive field for image SR. With the same depth of networks, the receptive fields are controlled by the dilation rate where a higher dilation rate r can obtain larger receptive field. We employ a baseline model of our proposed RDRN, which contains 3 recursions and each recursive block is composed of 4 SMRUs. The SMRUs in each recursive block share the same weight parameters and dilation rate. We have researched the dilation scheme as follows: 1-2-1, 1-2-3, 1-3-4, and 1-4-8. For example 1-2-1, which has the first block with $r = 1$, the second block with $r = 2$, and the third with $r = 1$. We adopt the model without dilation convolution (denote as RDRN-ND) as a reference. All of the models are evaluated on *Set5* and *Set14* for $4\times$ SR. The results are shown in Table 1, where the receptive fields in Table 1 calculated on the layers in RIR module. We can observe that larger receptive field can achieve higher SR performance and the scheme 1-3-4 achieves the best performance. Besides, we can also find that the scheme 1-4-8 obtain the largest receptive field but achieves lower PSNR than other three schemes. The reason is that when we use the dilated convolution to enlarge the receptive field, the feature maps should be padded which is consistent with the dilation rate r to maintain the same resolution of feature maps at each layer. When r becomes larger in top layers, the pixels from the input can be very sparse, which can destroy the correlations between pixels and then missing the local information. Therefore, we adopt the framework with the dilation scheme 1-3-4 as the final arrangement in our RDRN.

4.3.2. Study of K and M

The number of recusions K and the number of SMRUs M can directly affect the parameters and layers of the network and determines the model parameter and the SR performance. Now, we study on the effects of K and M for image SR. Table 2 shows the results of our models with different K and M for $2\times$ SR on *Set5*. We can observe that larger K or M would lead to higher performance. This is mainly because that the network becomes deeper with larger K or M. Besides, benefiting from the local recursive learning in RIR module, the models which employ the same number of recursions involve the same amount of weight parameters while increasing the values of M. As a result, we chose the model *K3M8* as the best structure for comparing with other methods.

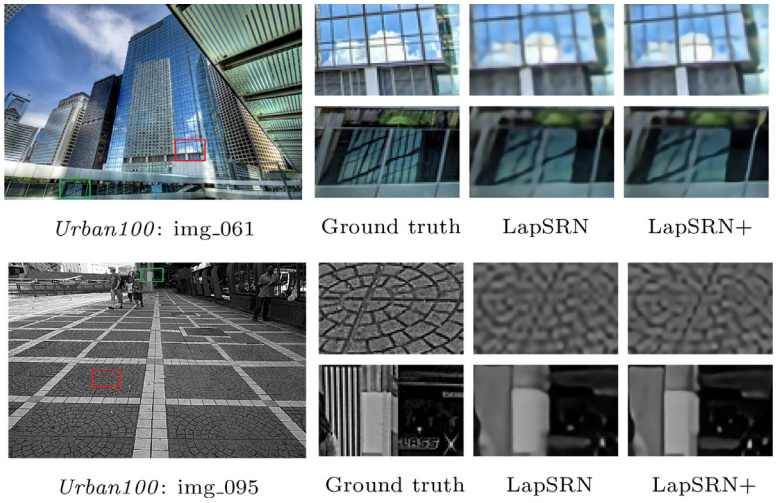


Fig. 5. Visual comparison of the LapSRN and LapSRN+ for 4 × SR on the *Urban100* datasets, where the LapSRN+ can produce clearer local texture and more straight lines.

Fig. 6. Visual comparison on the "ppt3" image from *Set14* [41] for ×4 scale. The lines are straightened and sharpened in our result, whereas other methods give blurry boundary.

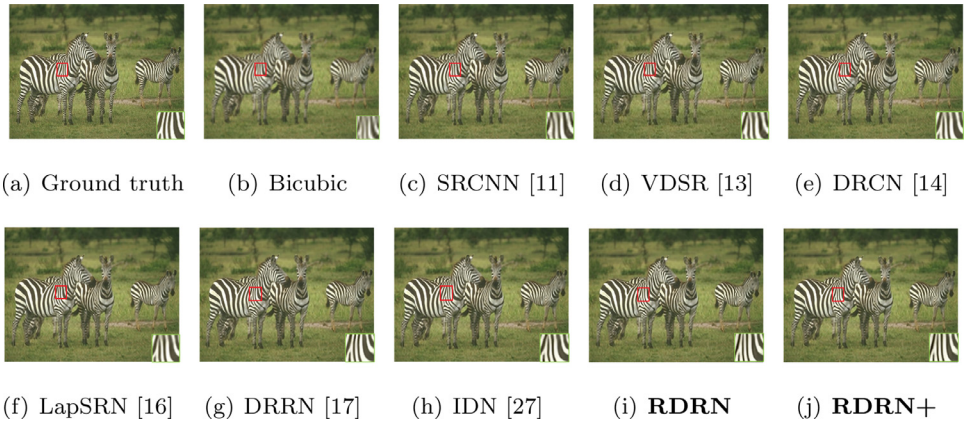


Fig. 7. Visual comparison on the "253027" image from *BSD100* [42] for 3 × SR. The stripes on the zebra are more clear and natural in our results, whereas other methods produce blurry content.

Table 3
Investigation of SPM, RIR, and DRM for 2 × SR on *Set5*. The **Text** Indicates the Best Performance.

Structure	Different combinations of SPM, RIR, and DRM							
SPM	×	✓	×	×	✓	✓	×	✓
RIR	×	×	✓	×	✓	×	✓	✓
DRM	×	×	×	✓	×	✓	✓	✓
Weights shared	No	No	Yes	No	Yes	No	Yes	Yes
PSNR	37.73	37.75	37.54	37.79	37.59	**37.86**	37.70	**37.73**
Parameters	2037K	2042K	373K	2075K	387K	2190K	520K	535K

Table 4
Quantitative evaluation of state-of-the-art SR algorithms: average PSNR/SSIM for scale factors 2, 3 and 4. Bold text indicates the best performance and italic text indicates the second best performance.

Methods	Scale	*Set5*	*Set14*	*BSD100*	*Urban100*
Bicubic	2	33.64 / 0.929	30.31 / 0.869	29.55 / 0.843	26.88 / 0.841
SRCNN [11]	2	36.35 / 0.952	32.29 / 0.905	31.15 / 0.885	29.10 / 0.890
VDSR [13]	2	37.53 / 0.959	33.15 / 0.913	31.90 / 0.896	30.77 / 0.914
DRCN [14]	2	37.63 / 0.959	32.98 / 0.913	31.85 / 0.894	30.76 / 0.913
LapSRN [16]	2	37.52 / 0.959	33.08 / 0.913	31.80 / 0.895	30.41 / 0.910
DRRN [17]	2	37.74 / 0.959	33.23 / 0.914	32.05 / 0.897	31.23 / 0.919
IDN [27]	2	*37.83* / *0.960*	*33.30* / *0.915*	*32.08* / *0.899*	31.27 / *0.920*
RDRN (our)	2	37.73 / *0.961*	33.25 / *0.915*	*32.08* / 0.898	*31.25* / *0.920*
RDRN+(our)	2	**37.86** / **0.961**	**33.31** / **0.916**	**32.12** / **0.901**	**31.30** / **0.921**
Bicubic	3	30.39 / 0.868	27.64 / 0.776	27.21 / 0.740	24.46 / 0.736
SRCNN [11]	3	32.76 / 0.908	29.41 / 0.823	28.41 / 0.787	26.24 / 0.800
VDSR [13]	3	33.66 / 0.921	29.77 / 0.834	28.83 / 0.798	27.14 / 0.829
DRCN [14]	3	33.82 / 0.922	29.76 / 0.833	28.80 / 0.797	27.15 / 0.828
LapSRN [16]	3	33.78 / 0.921	29.87 / 0.833	28.81 / 0.797	27.06 / 0.827
DRRN [17]	3	34.03 / 0.924	29.96 / 0.835	28.95 / 0.800	*27.53* / *0.838*
IDN [27]	3	*34.11* / *0.925*	*29.99* / 0.835	28.95 / 0.801	27.42 / 0.836
RDRN (our)	3	34.10 / *0.925*	*29.99* / *0.836*	*28.96* / *0.801*	*27.53* / *0.838*
RDRN+(our)	3	**34.17** / **0.928**	**30.05** / **0.837**	**29.09** / **0.803**	**27.60** / **0.839**
Bicubic	4	28.42 / 0.810	26.00 / 0.703	25.96 / 0.669	23.15 / 0.659
SRCNN [11]	4	30.48 / 0.862	27.50 / 0.752	26.90 / 0.712	24.16 / 0.707
VDSR [13]	4	31.35 / 0.884	28.02 / 0.768	27.29 / 0.725	25.18 / 0.753
DRCN [14]	4	31.53 / 0.885	28.02 / 0.767	27.23 / 0.723	25.14 / 0.751
LapSRN [16]	4	31.54 / 0.885	28.19 / 0.772	27.32 / 0.728	25.21 / 0.756
DRRN [17]	4	31.68 / 0.889	28.21 / 0.772	27.38 / 0.728	*25.44* / *0.763*
IDN [27]	4	31.82 / *0.890*	28.25 / *0.773*	27.41 / 0.730	25.41 / *0.763*
RDRN (our)	4	*31.77* / *0.890*	*28.26* / **0.774**	*27.43* / *0.731*	25.43 / *0.763*
RDRN+(our)	4	**31.89** / **0.891**	**28.29** / **0.774**	**27.51** / **0.732**	**25.48** / **0.764**

4.3.3. *Effectiveness of RIR, DRM, and SPM*

Now we turn our attention to the RIR, DRM, and the SPM in RIR. We research on the effects of these components for image SR. We first train models without any type of the three components. Then we investigate different combination of these components. All of the models adopt the same dilation scheme 1-3-4 and the same structure *K*3*M*8. There are 8 different combinations and the results for scale factor 2 on *Set5* are shown in Table 3. It is noted that the models without RIR means the same structure as others but no recursing learning. As shown in Table 3, we can find that the SPM or DRM can help to achieve higher PSNR performance. When we add the two components to the RIR structure can achieve the highest PSNR score but very large amount of parameters. We can also observe that the model with the all three components and weight sharing can achieve acceptable SR performance but much fewer parameters. To balance the trade-off between the parameters and SR performance, we combine the recursive learning with the three components as our final model, termed as RDRN. Besides, we chose the model which also contains the three components but no recursive learning strategy as the enhance vision of RDRN, termed as RDRN+ (the model with 2190K parameters in Table 3).

To demonstrate the proposed SPM can incorporate well with existing CNN based SR models for better SR performance, we add the SPM in the LapSRN [16] at the end of the feature extraction module within each scale level. Therefore, we add two SPUs in LapSRN for 4 × SR, termed as LapSRN+. We show the visual comparison of the two networks for 4 × SR on *BSD100* dataset. As shown in Fig. 5, we can see that the LapSRN+ can recover the HR images with clearer local texture detail and sharper edges compared with the original LapSRN, which demonstrates that the proposed SPM can help to enhance the discriminative capacity of networks for high-frequency detail recovery.

4.4. Comparing with the state-of-the-arts

In this section, we compare the two models of our proposed method, *i.e.* RDRN and RDRN+, with several state-of-the-art methods which include SRCNN [11], VDSR [13], DRCN [14], LapSRN [16], DRRN [17], and IDN [27]. Table 4 summarizes the quantitative results on the four benchmarks for 2 ×, 3 ×, and 4 × SR in terms of PSNR and SSIM. Table 4 summarizes the quantitative results on the four benchmarks, which illustrates that the RDRN can achieve the promising PSNR and SSIM performance compared with other methods. Besides, the RDRN+ significantly outperforms the state-of-the-art method IDN [27] in a considerable margin.

To fully investigate how the models perform in terms of visual quality, some promising results from several state-of-the-art methods with larger scales on *Set14* [41], *BSD100* [42], and *Urban100* [7] are visualized in Figs. 6–8. It is observed that our proposed method can reconstruct the lines, texture and stripes more accurately and clearly, which demonstrates that our method is able to restore the HR images with preserving more high-frequency detail while other methods reconstruct the HR images with more blurry contents.

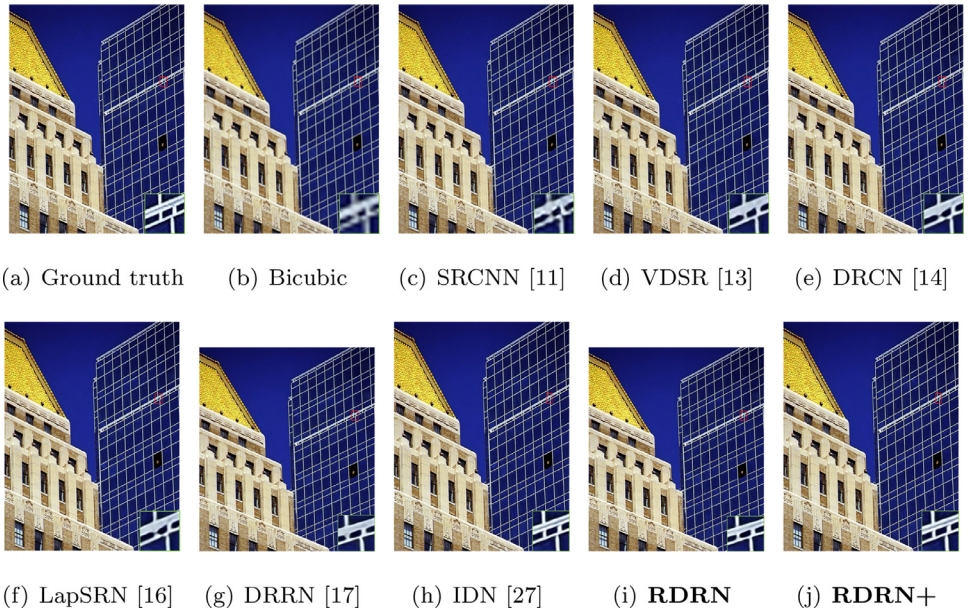


Fig. 8. Visual comparison on the "img_063" image from *Urban100* [7] for 4 × SR. The boundary is sharper in our results, whereas other methods give blurry lines.

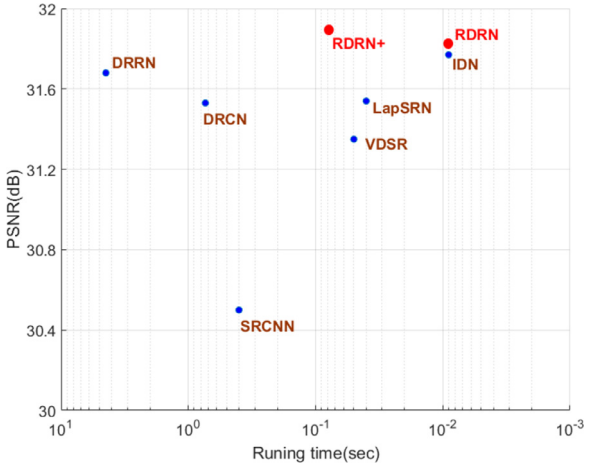


Fig. 9. PSNR performance versus runtime (evaluated in seconds). The results are evaluated on the *Set5* dataset for 4 × SR. The proposed RDRN achieves the comparable performance with the fastest reconstruction speed. The RDRN+ achieves the highest PSNR with acceptable execution time.

As for inference time, we use the public codes of the compared algorithms to evaluate the runtime on the machine with 3.4GHz Intel i7 CPU (128G RAM) and NVIDIA Titan Xp GPU (12G memory). Fig. 9 shows the shows the trade-offs between the execution time and PSNR performance on the *Set5* dataset for 4 × SR. As shown in Fig. 9, our proposed models RDRN achieves the best runtime performance with comparable SR results. Besides, the enhancement vision RDRN+ of our proposed method can stride the balance between the reconstruction accuracy and runtime, which outperforms the state-of-the-art method IDN [27] by a considerable margin.

5. Conclusion

In this paper, we propose a novel deep recursively dilated residual network (RDRN) to effectively exploit the contextual information over larger regions and emphasize local meaningful features for fast and accurate image SR via recursive and residual learning schemes. We present a spatial modulated unit (SPU), which incorporates the spatial modulation mechanism (SPM) with the dilated residual unit to model the contextual information over local representations within each feature map. Such SPM can incorporates well with exising CNN based models to obtain better SR results. The extensive experiments demonstrate that the proposed model is feasible and desirable for real-time applications.

Conflict of interest

None.

Acknowledgment

This work was supported in part by Fundamental Research Funds for the Central Universities (2019JBZ102).

References

[1] H. He, S. Mandal, A. Buehler, Improving optoacoustic image quality via geometric pixel super-resolution approach, IEEE Trans. Med. Image Process. 35 (3) (2016) 812–818.
[2] W. Zou, P.C. Yuen, Very low resolution face recognition problem, IEEE Trans. Image Process. 21 (1) (2012) 327–340.
[3] F. Li, X. Jia, D. Fraser, A. Lambert, Super resolution for remote sensing images based on a universal hidden Markov tree model, IEEE Trans Geosci. Remote. 48 (3) (2010) 1270–1278.
[4] Y. Tai, S. Liu, M.S. Brown, S. Lin, Super resolution using edge prior and single image detail synthesis, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2010, pp. 2400–2407.
[5] H. Zhang, J. Yang, Y. Zhang, T.S. Huang, Non-local kernel regression for image and video restoration, in: Proceedings of the European Conference on Computer Vision, 2010, pp. 566–579.
[6] J. Yang, Z. Lin, S. Cohen, Fast image super-resolution based on in-place example regression, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2013, pp. 1059–1066.
[7] J. Huang, A. Singh, N. Ahuja, Single image super-resolution from transformed self-exemplars, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 5197–5206.
[8] R. Timofte, V.D. Smet, L.V. Gool, A+: adjusted anchored neighborhood regression for fast super-resolution, in: Proceedings of the Asian Conference on Computer Vision, 2014, pp. 111–126.
[9] S. Schulter, C. Leistner, H. Bischof, Fast and accurate image upscaling with super-resolution forests, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 3791–3799.
[10] Q. Song, R. Xiong, D. Liu, Z. Xiong, F. Wu, W. Gao, Fast image super-resolution via local adaptive gradient field sharpening transform, IEEE Trans Image Process. 27 (4) (2018) 1966–1980.

[11] C. Dong, C. Loy, K. He, X. Tang, Image super-resolution using deep convolutional networks, IEEE Trans. Pattern Anal. Mach. Intell. 38 (2) (2015) 295–307.
[12] C. Dong, C. Loy, X. Tang, Accelerating the super-resolution convolutional neural network, in: Proceedings of the European Conference on Computer Vision, 2016, pp. 391–407.
[13] J. Kim, J.K. Lee, K.M. Lee, Accurate image super-resolution using very deep convolutional networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 1646–1654.
[14] J. Kim, J.K. Lee, K.M. Lee, Deeply-recursive convolutional network for image super-resolution, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 1637–1645.
[15] X. Mao, C. Shen, Y. Yang, Image restoration using very deep convolutional encoder-decoder networks with symmetric skip connections, in: Proceedings of Annual Conference on Neural Information Processing Systems, 2016.
[16] W.S. Lai, J.B. Huang, N. Ahuja, M.H. Yang, Deep Laplacian pyramid networks for fast and accurate super-resolution, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 5835–5843.
[17] Y. Tai, J. Yang, X. Liu, Image super-resolution via deep recursive residual network, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 2790–2798.
[18] B. Lim, S. Son, H. Kim, S. Nah, K.M. Lee, Image super-resolution using very deep residual channel attention networks, in: Proceedings of the European Conference on Computer Vision, 2018, pp. 8–14.
[19] C. Ledig, L. Theis, F. Huszar, J. Caballero, A.P. Aitken, A. Tejani, J. Totz, Z. Wang, W. Shi, Deep Laplacian pyramid networks for fast and accurate super-resolution, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 105–114.
[20] Y. Zhang, K. Li, K. Li, L. Wang, B. Wang, Y. Fu, Enhanced deep residual networks for single image super-resolution, in: Proceedings of the European Conference on Computer Vision, 2017, pp. 1132–1140.
[21] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, in: Proceedings of the International Conference on Learning Representations, 2015.
[22] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.
[23] Y. Tai, J. Yang, X. Liu, C. Xu, Memnet: a persistent memory network for image restoration, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 4549–4557.
[24] W. Han, S. Chang, D. Liu, M. Yu, M. Witbrock, T.S. Huang, Image super-resolution via dual-state recurrent networks, in: Proceedings of the IEEE International Conference on Computer Vision, 2018, pp. 1654–1663.
[25] F. Yu, V. Koltun, T. Funkhouser, Dilated residual networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 636–644.
[26] R. Girshick, J. Donahue, T. Darrell, J. Malik, Rich feature hierarchies for accurate object detection and semantic segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 580–587.
[27] H. Zheng, X. Wang, X. Gao, Fast and accurate single image superresolution via information distillation network, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 723–731.
[28] G. Huang, Z. Liu, L. Maaten, K.Q. Weinberger, Densely connected convolutional networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 2261–2269.
[29] T. Tong, G. Li, X. Liu, Q. Gao, Image super-resolution using dense skip connections, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 4809–4817.
[30] Y. Zhang, Y. Tian, Y. Kong, B. Zhong, Y. Fu, Residual dense network for image super-resolution, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 2472–2481.
[31] F. Yu, V. Koltun, Multi-scale context aggregation by dilated convolutions, in: Proceedings of the International Conference on Learning Representations, 2016.
[32] L. Chen, G. Papandreou, I. Kokkinos, K. Murphy, A.L. Yuille, Deeplab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs, IEEE Trans. Pattern Anal. Mach. Intell. 40 (4) (2018) 834–848.
[33] L. Chen, G. Papandreou, I. Kokkinos, K. Murphy, A.L. Yuille, Semantic image segmentation with deep convolutional nets and fully connected CRFs, in: Proceedings of the International Conference on Learning Representations, 2015.
[34] Z. Zhang, X. Wang, C. Jung, DCSR: Dilated convolutions for single image super-resolution, IEEE Trans. Image Process. 28 (4) (2019) 1625–1635.
[35] G. Lin, Q. Wu, L. Qiu, X. Huang, Image super-resolution using a dilated convolutional neural network, Neurocomputing 275 (2018) 1219–1230.
[36] W. Shi, F. Jiang, D. Zhao, Single image super-resolution with dilated convolution based multi-scale information learning inception module, in: Proceedings of the IEEE International Conference on Image Processing, 2017, pp. 977–981.
[37] K. He, X. Zhang, S. Ren, J. Sun, Delving deep into rectifiers: surpassing human-level performance on imagenet classification, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 1026–1034.
[38] W. Shi, J. Caballero, F. Huszár, J. Totz, A. P. A., et al., Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network, in: Proceedings of the IEEE International Conference on Computer Vision, 2016, pp. 1874–1883.
[39] E. Agustsson, R. Timofte, Ntire 2017 challenge on single image super-resolution: dataset and study, in: Proceedings of the IEEE International Conference on Computer Vision Workshop, 2017, pp. 1110–1121.
[40] M. Bevilacqua, A. Roumy, C. Guillemot, M.L. Alberi-Morel, Low-complexity single-image super-resolution based on nonnegative neighbor embedding, in: Proceedings of the British Machine Vision Conference, 2012, pp. 135.1–135.10.
[41] R. Zeyde, M. Elad, M. Protter, On single image scale-up using sparse-representations, in: Proceedings of the 7th International Conference on Curves Surfaces, 2012, pp. 711–730.
[42] D. Martin, C. Fowlkes, D. Tal, J. Malik, A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics, in: Proceedings of the IEEE International Conference on Computer Vision, 2001, pp. 416–423.
[43] Q. Huynh-Thu, M. Ghanbari, Scope of validity of PSNR in image/video quality assessment, Electr. Lett. 44 (13) (2008) 800–801.
[44] Z. Wang, A.C. Bovik, H.R. Sheikh, E.P. Simoncelli, Image quality assessment: from error visibility to structural similarity, IEEE Trans. Image Process. 13 (4) (2004) 600–612.
[45] D.P. Kingma, J. Ba, Adam: a method for stochastic optimization, CoRR (2014). arXiv: 1412.6980
[46] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z.D., et al., Automatic differentiation in pytorch, in: Proceedings of the Advances in Neural Information Processing Systems Workshop Autodiff, 2015, pp. 1–4.

Feng Li received his B.S. degree in Anhui Normal University, China, in 2012. Now, he is pursuing his Ph.D. degree in Institute of Information Science, Beijing Jiaotong University, Beijing, China. His research interests are image and video compression, video super resolution, and image restoration, such as image superresolution, and compression reduction.

Huihui Bai received her B.S. degree from Beijing Jiaotong University, China, in 2001, and her Ph.D. degree from Beijing Jiaotong University, China, in 2008. She is currently a professor in Beijing Jiaotong University. She has been engaged in R&D work in video coding technologies and standards, such as HEVC, 3D video compression, multiple description video coding (MDC), and distributed video coding (DVC).

Yao Zhao received the BS degree from Fuzhou University, China, in 1989, and the ME degree from Southeast University, Nanjing, China, in 1992, both from the Radio Engineering Department, and the PhD degree from the Institute of Information Science, Beijing Jiaotong University (BJTU), China, in 1996. He became an associate professor at BJTU in 1998 and became a professor in 2001. From 2001 to 2002, he was a senior research fellow with the Information and Communication Theory Group, Faculty of Information Technology and Systems, Delft University of Technology, Delft, The Netherlands. He is currently the director of the Institute of Information Science, BJTU. His current research interests include image/video coding, digital watermarking and forensics, and video analysis and understanding. He serves on the editorial boards of several international journals, including as associate editors of IEEE Transactions on Cybernetics, IEEE Signal Processing Letters, and an area editor of Signal Processing: Image Communication (Elsevier), etc. He was named a distinguished young scholar by the National Science Foundation of China in 2010, and was elected as a Chang Jiang Scholar of Ministry of Education of China in 2013. He is a senior member of the IEEE.