FilterNet: Adaptive Information Filtering Network for Accurate and Fast Image Super-Resolution

Feng Li, Huihui Bai, and Yao Zhao, *Senior Member, IEEE*

***Abstract*— Deep convolutional neural network (CNN) approaches have achieved impressive performance for image super-resolution (SR). The main issue of image SR is to effectively recover the high-frequency detail of low-resolution (LR) input. However, existing CNN methods often inevitably exhibit a large amount of memory consumption and computational cost. In addition, in most SR networks, the low-frequency and high-frequency components of the LR features are treated equally in the training process, which can ignore the local detailed information and hinder the representational capacity of networks. To solve these issues, in this paper, we propose a deep adaptive information filtering network (FilterNet) for accurate and fast image SR. In contrast to the existing methods that adopt fully CNN methods to directly predict the HR images, the proposed FilterNet concentrates on more useful features and adaptively filters the redundant low-frequency information. In general, we present the dilated residual group (DRG), which consists of multiple dilated residual units. The DRGs can directly expand the receptive field of the network to efficiently exploit the contextual information of the LR input. In the dilated residual unit, a gated selective mechanism is proposed to adaptively learn more high-frequency information and filter the low-frequency information. Besides, we introduce a novel adaptive information fusion structure, which builds long scaling skip connections among the DRGs to rescale the hierarchical features and fuse more detailed information. The scaling weights can be deemed as the part parameters of our network and trained adaptively. The Extensive evaluations on benchmark datasets demonstrate that our FilterNet achieves superior performance both on accuracy and speed compared with recent state-of-the-art methods.**

***Index Terms*— Image super-resolution, dilated convolution, gated selective mechanism, adaptive information fusion.**

I. Introduction

SINGLE image super-resolution (SISR) aims to recover the corresponding high-resolution (HR) image from its low-resolution (LR) observation. Image SR has been widely used in various image and video processing tasks, such as surveillance imaging, medical imaging, and video streaming. The image SR problem is typically an ill-posed problem since there are numerous solutions for LR input. To alleviate this inverse problem, plenty of image SR methods constrain the solution with strong prior knowledge, which can be roughly categorized into interpolation-based methods, reconstruction-based methods, and learning-based methods.

Interpolation-based methods utilize fixed filters [1] or adaptive filters [2]–[4] to produce HR images from given LR images by estimating the unknown pixels in HR grids. Although these methods are generally simple and fast for real-time SR applications, they are prone to overlook the high-frequency detail, which can lead to blurry edges and noticeable artifacts. Reconstruction-based methods assume that the corrupted images are produced by multiple degradation factors and restore the HR images by reversing the degradation process. Based on the maximum a posteriori probability (MAP), most methods impose specific prior knowledge as regularization terms to regularize the recovery process. Typically, image priors include edge priors [5]–[7], gradient profile priors [8], non-local similarity [9], and total variation [10]. Though these types of SR approaches can preserve sharper edges to some extent, they usually result in unpleasant artifacts and overly smooth reconstruction results.

Learning-based methods estimate HR pixels by exploiting the internal similarities of image patches or learning the relationships between external LR and HR exemplar pairs. Based on the sparse signal representation, Yang *et al.* [11] jointly train two dictionaries for LR and HR image patches. Dong *et al.* [12] present an adaptive sparse domain selection (ASDS) and an adaptive regularization scheme to learn various sets of bases from a pre-collected dataset of example image patches for image SR. Zhu *et al.* [13] propose a fast image SR method based on self-example patch-based dictionary learning and sparse representation, which exploits the sparse signal representation theory in the framework of compressive sensing (CS) and dictionary learning of image patches. In [14], Timofte *et al.* propose an anchored neighbor regression (ANR) approach, which combines sparse dictionary learning with neighbor embedding methods for fast image SR reconstruction. Timofte *et al.* further introduce an improved variant of ANR (A+) [15], which super-resolves the problem of image upscaling based on the features and anchored regressors from ANR. Peleg and Elad [16] address the image SR problem using a statistical prediction model based on the sparse representations of low- and high-resolution image patches. By exploiting one-to-many correspondences

Manuscript received October 6, 2018; revised January 31, 2019; accepted March 9, 2019. Date of publication March 20, 2019; date of current version June 4, 2020. This work was supported in part by the Fundamental Research Funds for the Central Universities under Grant 2018JBZ001 and in part by the Key Innovation Team of Shanxi 1331 Project under Grant KITSX1331. This paper was recommended by Associate Editor G. Valenzise. *(Corresponding author: Huihui Bai.)*

The authors are with the Beijing Key Laboratory of Advanced Information Science and Network Technology, Beijing Jiaotong University, Beijing 100044, China, and also with the Institute Information Science, Beijing Jiaotong University, Beijing 100044, China (e-mail: 11feng@bjtu.edu.cn; hhbai@bjtu.edu.cn; yzhao@bjtu.edu.cn).

Color versions of one or more of the figures in this article are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TCSVT.2019.2906428

between LR and HR patches, Xiong *et al.* [17] propose a statistical method called soft information and decision to improve the reconstruction accuracy. In [18], Huang *et al.* present a self-similarity driven SR method that uses the transformed self-exemplars (SelfExSR), which exploits the patch search space expansion for improving the self-exemplar search. Schulter *et al.* [19] propose an approach for SISR via random forests, which directly maps from low to high-resolution patches using random forests.

Recently, due to the superior capacity of deep learning models, especially the convolutional neural network (CNN), various CNN methods have been developed to perform SISR and have achieved impressive performance. In [20], Dong *et al.* first introduce a fully convolutional neural network for image SR (SRCNN), which directly learns an end-to-end mapping between LR and HR images. Inspired by the VGG-net [21] used in image classification, Kim *et al.* [22] present a very deep convolutional network (VDSR) for accurate SR reconstruction. Kim *et al.* [23] further introduce a deeply recursive convolutional network (DRCN) which uses recursive-supervision and skip connections to ease the training process. Motivated by the success of residual learning networks (ResNet) [24]–[28] utilize the residual learning formulation to train very deep networks for better SR performance.

Though the above methods have achieved promising performance, most of these deep networks still exhibit some drawbacks. First, some networks [20], [22], [23] have small receptive fields, which are limited to exploiting the contextual information of LR input images. Although deeper models [25], [28], [29] can increase the receptive fields by stacking more convolutional layers, they tend to face the challenges of memory consumption and computational complexity. The efficiency of SR reconstruction must be sacrificed to maintain the reconstruction accuracy, which is seriously detrimental to real-time applications. Second, the main issue of image SR is to recover the high-frequency detail of input LR counterparts. In most CNN based SR models, the redundant low-frequency and high-frequency information are treated equally in the training process, which limits the ability to extract more detailed information and maximize the representational capacity of networks. Finally, most image SR methods commonly minimize the mean squared error (MSE) between the reconstructed HR image and the ground truth image, which can fail to recover sharp edges and lead to misleading artifacts.

To address the above problems, we propose an adaptive information filtering network (FilterNet) to adaptively learn more useful features for fast and accurate image SR. In the proposed FilterNet, a feature extraction module is first adopted to extract the shallow LR features from the observed LR images. Then, motivated by the fact that dilated convolution supports exponentially expanding the receptive field without increasing parameters or loss resolution, we present dilated residual groups (DRGs) to efficiently exploit the contextual information of the LR input image. Each DRG consists of multiple dilated residual units composed of dilated convolutional layers to expand the receptive field of the network. To concentrate on more useful information for high-frequency detail reconstruction, we incorporate the gated selective mechanism (GSM) into our dilated residual unit, termed the selective residual unit (SRU), to rescale the internal features and suppress the less valuable features. Moreover, instead of directly using traditional skip connections to speed up the training process, we propose an adaptive information fusion structure (AIFS), which adaptively assigns different scaling weights to different inputs of the bypass connections. Such SRU and AIFS allow our proposed FilterNet to learn more useful features to help better recover the image details and maximize the pixel-wise fitting capacity for highly accurate image SR reconstruction. Furthermore, we employ an upscale module for upsampling the previous feature maps to a finer level. Finally, a convolutional layer is used to predict the HR residual images. We conduct an element-wise addition operation on the predicted HR residual images and the correspondingly upsampled LR images to obtain the final HR images.

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

1) We propose a novel FilterNet framework to solve the image SR problem. The proposed network can efficiently exploit the contextual information of LR input images and adaptively rescale different types of features to produce highly accurate results.
2) The proposed FilterNet employs multiple cascaded dilated residual groups (DRG) that consist of many selective residual units (SRU) with a stacked style. The SRU first utilizes dilated convolutional layers to expand the receptive field of the network, which can effectively exploit the contextual information. To concentrate on more useful information for high-frequency detail reconstruction, we propose the gated selective mechanism (GSM) and incorporate it in the SRU to adaptively learn more high-frequency information and filter the low-frequency information.
3) We present an adaptive information fusion structure (AIFS), which builds adaptively weighted skip connections among these DRGs to pass more useful features and improve the representational capacity of the network. With the structure of the proposed FilterNet, our model outperforms state-of-the-art methods in terms of both speed and accuracy.

The rest of this paper is organized as follows. The related works are briefly reviewed in Section II. Section III introduces the proposed FilterNet with a detailed analysis of every component. The implementation details and experimental results on several public benchmarks are presented in Section IV. The conclusion of this paper is summarized in Section V.

II. Related Work

A. Deep Neural Networks for Image Super-Resolution

Since Dong *et al.* [20] first propose a three-layer CNN for image SR (SRCNN), the majority of recent works have adopted deep neural networks to super-resolve the ill-posed problem and have achieved dramatic improvements. Mao *et al.* [25] propose a very deep convolutional encoder-decoder network for image restoration (RED), which uses

multiple convolutional and deconvolutional layers on the end-to-end mappings from the corrupted images to the original HR images. Zhang *et al.* [26] propose a deep CNN for image denoising (DnCNN) with residual learning and extend it to the image SR task. Yang *et al.* [27] integrate edge priors into a residual network to recurrently perform image SR. In [28], Tai *et al.* propose a deep recursive residual network (DRRN) for image SR, which adopts both local and global learning strategies to mitigate the difficulty of training very deep networks. Tai *et al.* [29] further propose a persistent memory network (MemNet) with much a deeper architecture for image restoration.

However, all of these methods utilize bicubic interpolation to upsample the original LR images to the HR space before applying them into networks, which can increase the computational complexity for image SR. To investigate more accurate and efficient SR methods, Dong *et al.* [30] employ a deconvolutional layer at the end of the network to replace the bicubic interpolation and propose a fast convolutional neural network for image SR (FSRCNN). Lai *et al.* [31] propose the Laplacian pyramid SR network (LapSRN) to progressively reconstruct the sub-band residual HR images and use a deconvolutional layer for upsampling the LR images. Lai *et al.* [32] further present a multi-scale model (MS-LapSRN) to learn the inter-scale correlation and improve the reconstruction accuracy compared with single-scale models. In [33], Han *et al.* argue that many deep SR models can be reformulated as a single-state recurrent neural network (RNN) with finite foldings and propose a dual-state recurrent network (DSRN) for image SR. Shi *et al.* [35] propose an efficient sub-pixel convolutional neural network (ESPCN), which introduces a sub-pixel convolutional layer to learn the upscaling operation for image and video SR. Ledig *et al.* [36] propose an SR generative adversarial network (SRGAN) for a 4× upscaling factor, which also employs the sub-pixel convolutional layer proposed in ESPCN to increase the resolution of the input images. In [37], Wang *et al.* present a resolution-aware network (RAN) to address the SR problem, which designs an upsampling network consisting of multiple submodules to learn from the training samples of different resolutions.

B. Dilated Convolution

Dilated convolution was originally developed in *algorithm àtrous* for wavelet decomposition in signal processing [38], [39], which removes the downsampling operator from the usual implementation of discrete wavelet transform (DWT). In this implementation, the responses of the filters are upsampled, thereby inserting "holes (zeros)" between nonzero filter taps in convolutional filters, which is equivalent to a convolution with a larger filter derived from the original filter by inserting it with zeros. In a 1D signal, the output $y[i]$ of the dilated convolution of a 1D signal $x[i]$ with a filter $w[i]$ of length K is defined as:

$$y[i] = \sum_{k=1}^{K} x[i + r \cdot k] \cdot w[k] \tag{1}$$

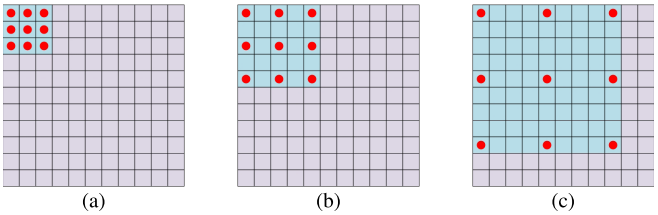


Fig. 1. The receptive field of a dilated convolution with a kernel size of 3×3 and different dilated rates r of 1, 2, and 4. The standard convolution is a special case for rate $r = 1$. (a) $r = 1$. (b) $r = 2$. (c) $r = 4$.

The parameter r is termed the dilated rate, which corresponds to the input stride with which we sample the input signal. As a special case, the dilated convolution with $r = 1$ yields the standard convolution.

In image processing, 2D dilated convolution can be interpreted as inserting "holes (zeros)" between each pixel in the convolution kernel. Therefore, the dilated convolution of a 2D signal $x[m, n]$ with a filter $w[i, j]$ of length M and width N can be defined as

$$y[m, n] = \sum_{i=1}^{M} \sum_{j=1}^{N} x[m + r \cdot i, n + r \cdot j] \cdot w[i, j] \tag{2}$$

where $y[m, n]$ is the output of the dilated convolution operation from $x[m, n]$. Given a convolution kernel with a size of $k \times k$, the resulting kernel size can be increased to $k + (k - 1) \times (r - 1)$ with the dilated rate r. As shown in Fig. 1, with the same kernel size 3×3, one dilated convolution layer can obtain the receptive field of 5×5 with $r = 2$, but the standard convolution can only obtain a 3×3 receptive field. This demonstrates that we can adopt a dilated convolution to effectively enlarge the receptive fields of networks with fewer layers or parameters compared with the standard convolution. Dilated convolution has been used in various tasks, such as image segmentation [40]–[42], object detection [43], image classification [45], and scenes understanding [46].

III. Proposed Method

In this section, we elaborate on each main component of our proposed FilterNet for image SR. As shown in Fig. 2, our FilterNet consists of four parts: a feature extraction module, multiple cascaded dilated residual groups (DRG), an upscale module, and a reconstruction layer. Then, we analyze the architecture of our network, generally involving the depth of the network and the parameters of each layer. Finally, we introduce the loss function for training our models.

A. Feature Extraction

In image SR, the quality degradation of an HR image x_{HR} to an LR image x_{LR} can be caused by blurring and downsampling. The corrupted low-quality image x_{LR} can be represented as

$$x_{LR} = BD(x_{HR}) + n \tag{3}$$

where B and D denote the blurring and downsampling operations, respectively, and n is the additive noise in the

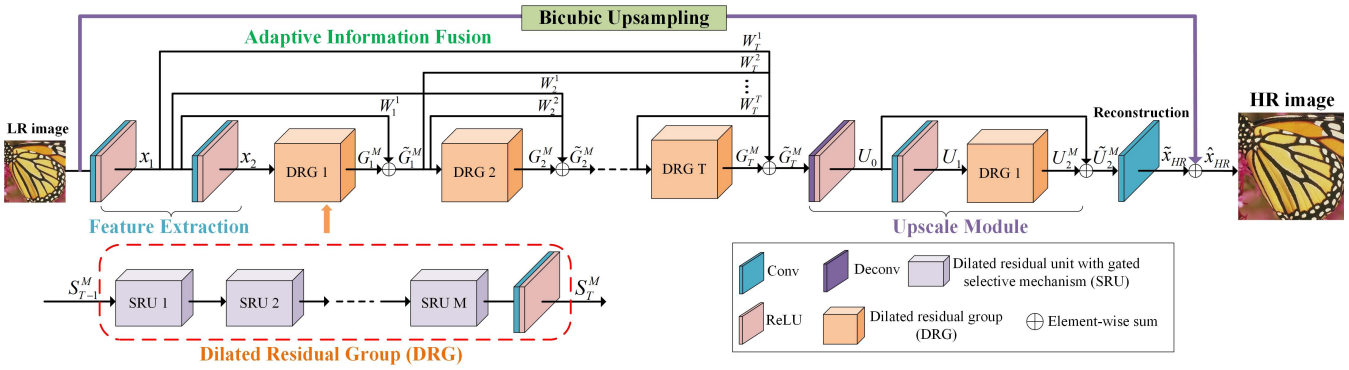


Fig. 2. The network architecture of our proposed adaptive information filtering network (FilterNet) for image SR.

degradation process. Let us denote the x_{LR} and $\hat{x}_{HR}$ as the observed LR input and the estimated HR output of FilterNet.

With respect to the feature extraction module, we use two convolutional layers with a kernel size of 3×3 to extract the shallow LR features from the original LR input. The first convolutional layer extracts the features x_1 and can be represented as

$$x_1 = f_1(x_{LR}) \tag{4}$$

where $f_1(\cdot)$ denotes the first feature extraction function. x_1 is used for further feature extraction and serves as the input for information fusion. Therefore, we have

$$x_2 = f_2(x_1) \tag{5}$$

where $f_2(\cdot)$ represents the second feature extraction function and x_2 denotes the further extracted LR features used as the input of the following state.

B. Dilated Residual Group

Now we introduce the detail of our proposed dilated residual group (DRG) in Fig. 2, which is composed of multiple selective residual units (SRU) and a 1×1 convolutional layer with a stacked style. Each SRU combines the dilated convolution with the gated selective mechanism (GSM) to efficiently exploit the contextual information and ensure that the proposed network is more sensitive to informative features. In addition, the adaptive information fusion structure (AIFS), which adopts long scaling skip connections among these DRGs, is used to pass more useful features to later stages for detailed information fusion.

1) Selective Residual Unit: By assuming that optimizing residual mapping is easier than the original unreferenced mapping, He *et al.* [24] propose the deep residual network (ResNet), which explicitly utilizes few stacked layers to fit the residual mapping rather than directly fitting the desired underlying mapping. As shown in Fig. 3(a), for the i^{th} residual unit, denoting the input as x_{i-1} and the desired underlying mapping as $H(x_i)$, the residual mapping is defined as $F(x_{i-1}) = H(x_i) - x_{i-1}$. Thus, the desired formulation of $F(x_{i-1}) + x_{i-1}$ can be realized with the identity branch illustrated in Fig. 3(a). Starting with the ResNet architecture, Yu *et al.* [45] propose the Dilated Residual Network (DRN), which uses dilated convolutions to increase the receptive

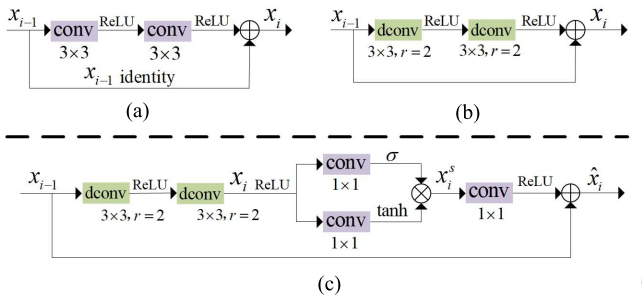


Fig. 3. (a) The architecture of the residual unit in ResNet. (b) The architecture of the dilated residual unit in DRN, where "dconv" denotes the dilated convolution. (c) The architecture of the dilated residual unit with a gated selective mechanism in our proposed FilterNet, where "⊗" denotes the element-wise product operation.

field of higher layers and compensate for the reduction of the receptive field by removing the subsampling layers. The dilated residual unit corresponding to the residual unit in ResNet is shown in Fig. 3(b). Here, we use the dilation rate $r = 2$ to give a simple example. With the kernel size of 3×3, we can see that the residual unit only gains the receptive field of 5×5 but the dilated residual unit in Fig. 3(b) can achieve the receptive field of 9×9.

In CNN based image SR, the predicted pixel value of the desired HR image depends on the receptive field of networks. The size of the receptive field determines that the amount of contextual information of the LR input image can be captured to infer high-frequency components. To obtain sufficient contextual information and provide more clues for better high-frequency detail prediction, as sketched in Fig. 3(c), in our proposed SRU, we first use two dilated convolutional layers to obtain a larger receptive field. With the input x_{i-1} of the i^{th} SRU, the output x_i of the first two dilated convolutional layers can be represented as

$$x_i = f_{i,2}(\tau(f_{i,1}(x_{i-1}))) \tag{6}$$

where $f_{i,1}(\cdot)$ and $f_{i,2}(\cdot)$ denote the dilated convolution operations, and $\tau(\cdot)$ denotes the ReLU [47] activation function.

2) Gated Selective Mechanism: In image SR, the LR image contains redundant low-frequency components and necessary high-frequency components including edges, textures and

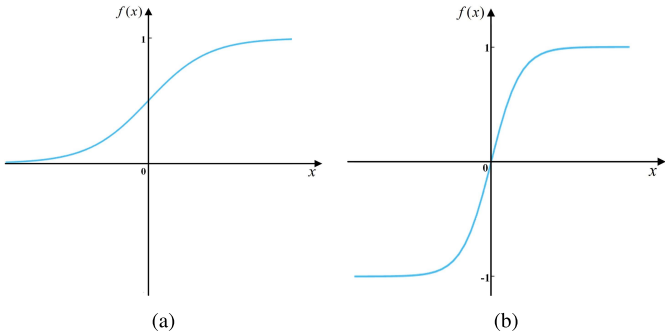


Fig. 4. The graphs of sigmoid and hyperbolic tangent (tanh) functions. (a) Sigmoid function. (b) Hyperbolic tangent (tanh) function.

other details. Previous CNN-based SR methods extract the features from the LR input and equally learn all the features during the training process, which lack flexibility in dealing with the low- and high-frequency information. To solve the above problems, we propose a gated selective mechanism (GSM) to exploit more important high-frequency information and suppress less valuable low-frequency information. Specifically, after the output of the second dilated convolutional layer x_i activated by ReLU, and the information flow is input into two branches. As illustrated in Fig. 3(c), in each branch, a convolutional layer with a kernel size of 1×1 is employed to integrate the information on each pixel in different channels. Next, we use the sigmoid function to adaptively rescale the internal features and control the output information flow into the later state. The graph of the sigmoid function is visualized in Fig. 4(a). We can observe that the sigmoid function has a domain of all real numbers, with a return value monotonically increasing from 0 to 1. Hence, in the top branch of GSM, the sigmoid function layer after the 1×1 convolutional layer can be viewed as the input gate, which decides what previous input features need to be updated. Compared to the sigmoid function, as shown in Fig. 4(b), the hyperbolic tangent (tanh) function squashes the real numbers to the range between $[-1, 1]$. In the bottom branch of GSM, with all of the input features from the previous state, the tanh function layer is utilized to create a new candidate value vector. Then, we combine these two to create the new updated features by conducting a multiply operation on them. The gated selective function can be formulated as

$$x_i^s = \sigma(g_{i,1}(\tau(x_i))) \otimes tanh(g_{i,2}(\tau(x_i))) \tag{7}$$

where $\sigma(\cdot)$ and $tanh(\cdot)$ represent the sigmoid and tanh functions, respectively. $g_{i,1}(\cdot)$ and $g_{i,2}(\cdot)$ are the convolution operations in the two branches. x_i^s is the new candidate value produced by GSM, and $\otimes$ is the element-wise product operation. Finally, followed by a convolutional layer with a kernel size of 1×1, the new candidate value x_i^s is integrated and further added to the original input x_{i-1} to obtain a new information flow that passes to the following state. This procedure can be expressed as

$$\hat{x}_i = \tau(g_{i,3}(x_i^s)) + x_{i-1} \tag{8}$$

where $\hat{x}_i$ denotes the final updated features and $g_{i,3}(\cdot)$ is the convolution operation of the third 1×1 convolutional layer in Fig. 3(c). Assuming there are M SRUs in each DRG, the output G_L^M of the L^{th} DRG can be formulated as

$$G_L^M = h_L(S_L^M(S_L^{M-1}(\ldots S_L^2(S_L^1(G_{L-1}^M))\ldots))) \tag{9}$$

where $[S_L^1, S_L^2, \ldots, S_L^{M-1}, S_L^M]$ denote the mapping function of M SRUs in the L^{th} DRG. $h_L(\cdot)$ is the convolution operation after M SRUs (see in Fig. 2). G_{L-1}^M is the output of the $(L-1)^{th}$ DRG.

3) Adaptive Information Fusion Structure: Instead of directly linking the output of previous states to the current state with skip connections, we build long scaling skip connections among these DRGs, which assign different scaling weights for different states to pass more detailed information for highly accurate prediction. The scaling weights can be deemed as the part parameters of our FilterNet and trained adaptively. In our proposed FilterNet, as shown in Fig. 2, the output x_1 serves as the first input for information fusion. Therefore, the output of the first fusion step can be represented as

$$\tilde{G}_1^M = W_1^1 x_1 + G_1^M \tag{10}$$

where W_1^1 denotes the first scaling weight of x_1 to the first DRG and $\tilde{G}_1^M$ is the output of the first fusion step. Supposing there are T DRGs before the upscale module in the FilterNet, each DRG has M SRUs, the output of the T^{th} information fusion $\tilde{G}_T^M$ can be expressed as

$$\tilde{G}_T^M = W_T^1 x_1 + \sum_{t=1}^{T-1} W_T^{t+1} \tilde{G}_t^M + G_T^M \tag{11}$$

where the G_T^M is the output of the T^{th} DRG and the $[W_T^1, W_T^2, \ldots, W_T^T]$ are the T scaling weights for the T^{th} fusion step. Benefiting from the SRU and AIFS, our network can efficiently exploit the contextual information of the LR input images and adaptively learn more useful features.

C. Upscale Module

For upscaling the LR input images to the desired spatial resolution, in this work, we utilize one deconvolutional layer to learn the upscaling filters. The deconvolution operation can be regarded as an inverse process of the convolution operation. In a sense, assuming that the stride of the input filter is s, it can be found that upsampling with factor s by deconvolution is the convolution operation with a fractional input stride $1/s$. Specifically, with the input $\tilde{G}_T^M$, we first adopt a 3×3 deconvolutional layer to learn the upscale filters. For different scale factors, we use the same kernel size with different strides corresponding to the scale factors. The upsample operation is performed as

$$U_0 = u_0(\tilde{G}_T^M) \tag{12}$$

where $u_0(\cdot)$ is the upsampling function, and U_0 denotes the upscaled HR features. Then, we use a convolutional layer with a kernel size of 3×3 to extract the HR features

$$U_1 = u_1(\tau(U_0)) \tag{13}$$

where $u_1(\cdot)$ denotes the convolution operation for HR features extraction, and U_1 is the extracted features from U_0. Afterwards, we only employ one DRG to learn the HR features and

TABLE I
DETAILED SETTING OF EACH MODULE IN OUR PROPOSED NETWORK WITH SCALE FACTORS OF s ($s = 2, 3, 4$) FOR IMAGE SR

Module	Output size	Layer	Filter size
Feature Extraction	$w \times h \times 64$	Conv, ReLU, Conv, ReLU	3×3
Multiple DRGs	$w \times h \times 64$	{SRUs: Dilated Conv, ReLU, Dilated Conv, ReLU Conv, Sigmoid & Conv, Tanh Element-product, Conv, ReLU, Element-wise} Conv, ReLU	3×3 1×1 1×1 1×1
Upscale Module	$sw \times sh \times 64$	deconv, ReLU, conv, ReLU 1 DRG	3×3 —
Reconstruction Layer	$sw \times sh \times 1$	Conv, Global residual	3×3

the shortcut connection without scaling weights to conduct the residual mapping

$$U_2^M = u_2(\tau(U_1))$$
$$\tilde{U}_2^M = U_2^M + U_0 \tag{14}$$

where U_2^M and $u_2(\cdot)$ are the output and mapping function, respectively, of the DRG in the upscale module. $\tilde{U}_2^M$ is the final learned HR residual features after the residual learning operation. A convolutional layer with a kernel size of 3×3 is performed as reconstruction layer to reconstruct the residual HR image $\tilde{x}_{HR}$

$$\tilde{x}_{HR} = R(\tilde{U}_2^M) \tag{15}$$

where $R(\cdot)$ is the reconstruction function. Finally, to subtract the smooth area of the original LR input, the global residual learning is performed on the residual between the estimated HR residual image and the bicubic upsampled image to improve the reconstruction performance. Therefore, the output $\hat{x}_{HR}$ of our proposed FilterNet can be expressed as

$$\hat{x}_{HR} = \tilde{x}_{HR} + B(x_{LR}) \tag{16}$$

where B denotes the bicubic upsampling operation.

D. Loss Function

According to [48], solely utilizing L2 loss can fail to recover sharp edges and lead to overly smooth results. L1 loss provides better convergence than L2 loss for image SR. In our experiments, we find that training the network with L1 loss can achieve better performance both on PSNR and visual quality than L2 loss. Therefore, we train our networks using L1 loss instead of L2. The evaluation of this comparison is provided in Section IV. Given a training set $\left\{x_{LR}^{(i)}, x_{HR}^{(i)}\right\}_{i=1}^{N}$, where N is the number of training patches and $x_{HR}^{(i)}$ is the ground truth HR patch of the LR patch $x_{LR}^{(i)}$, the loss function of the basic FilterNet with the parameter set Θ is

$$L(\Theta) = \frac{1}{N}\sum_{i=1}^{N} \|x_{HR}^{(i)} - \hat{x}_{HR}^{(i)}\|_1 \tag{17}$$

IV. EXPERIMENTS

A. Datasets

In this work, we use 400 images from the training and validation set of BSDS500 [49] and 800 images from the training set of the DIV2K dataset [50], totaling 1200 images without data augmentation for training our models. For testing, we compare our models with recent state-of-the-art SR methods on four popular benchmark datasets: SET5 [51], SET14 [52], BSDS100 [49], and URBAN100 [18]. All experiments are evaluated on the peak signal-to-noise ratio (PSNR), the structural similarity (SSIM) index, and inference time.

B. Implementation Details

In our proposed FilterNet, apart from the deconvolutional layer and the reconstruction layer, each convolutional layer and dilated convolutional layer consists of 64 filters with the stride of 1. The deconvolutional layer in our upscale module consists of 64 filters with the strides corresponding to the scale factors 2, 3, and 4. The reconstruction layer is a convolutional layer that has one channel with stride 1 to reconstruct the HR images. All of the weight layers except the reconstruction layer are followed by the ReLU [47] as an activation function. The detailed setting is summarized in Table I, where w, h denote the image width and height, respectively.

In the training phase, the original images are first converted to the YCbCr color space and only the Y-channel is processed. LR training patches are obtained by down-scaling the HR patches using bicubic interpolation with scale factors of 2, 3, and 4. In each training batch, we randomly sample 32 patches with the size of 120×120 without overlapping. The weights were initialized by the method proposed in [53], and the biases were initialized to zero. We implement our FilterNet with the Caffe package [54] and optimize the models using an Adam [55] optimizer. We set the momentum parameter to 0.9 and the weight decay to $1e-4$. The initial learning rate is set to $1e-4$ and decreased by a factor of 10 for every 60 epochs in the training phase. The training of a single FilterNet model can roughly take 2 days with a Titan Xp GPU.

C. Ablation Study

In this subsection, we first investigate the basic network parameters of the proposed FilterNet: the number of DRGs (denoted as T), and the number of SRUs per DRG (denoted as M). Then, we study the dilation rate schemes of these dilated convolutional layers. Next, we analyze the effects of the GSM and the AIFS. Finally, we discuss the performance of L1 and L2 loss functions for training our models.

1) Investigation of T and M: To clearly show how the parameters T and M affect the performance of FilterNet,

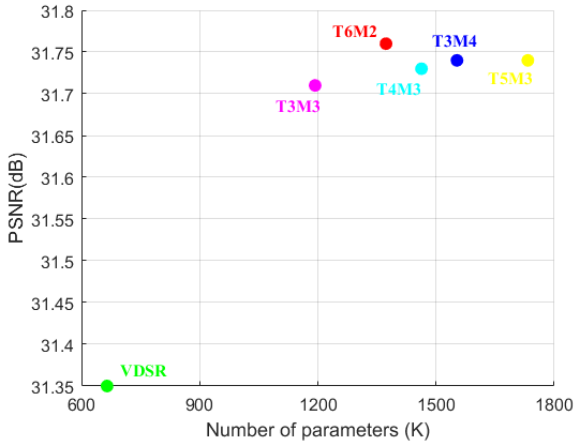


Fig. 5. The comparisons of various FilterNets at different T and M combinations in terms of PSNR and parameters for the scale factor of 3 on SET5 dataset. The green denotes the model of VDSR, while other colors denote the models of FilterNet with different structures.

we fix the number of DRGs in upscale modules to 1. Then, we change the number of DRGs in the front part and the number of SRUs in the DRGs. With different numbers of T and M, we obtain five networks, which are denoted as T3M3, T4M3, T3M4, T5M3, and T6M2. We apply the trained models on the SET5 dataset and then illustrate the PSNR and parameters of these structures for 4× SR. We use the performance of VDSR [22] as a reference. As shown in Fig. 5, compared with VDSR, our proposed method has approximately 0.37 to 0.42 dB improvement. In addition, one can see that, a larger T or M can achieve better performance, which is mainly attributable to a deeper network caused by a larger T or M. Additionally, the T6M2 architecture performs better than T5M3 but has fewer parameters, which suggests that the gains of our models are not only from the deep depth but also from the richer representations and hierarchical information fusion. Considering the trade-off between the PSNR performance and parameters of these models, we adopt the T6M2 as our baseline model, denoted as FilterNet_T6M2, for the following experiments.

2) Study of Dilation Rate Schemes: We now turn our attention to dilated convolution in the proposed FilterNet for image SR. To demonstrate the effectiveness of dilation convolution, we conduct experiments to reveal appropriate dilation rate schemes for image SR. We use the baseline model FilterNet_T6M2 with dilation rate $r = 1$ (no dilation), which contains 6 DRGs and 2 SRUs in each DRG, as the reference. Specifically, we conduct our experiments with several variants on the following assignments:

(i) 1-2: For 6 DRGs, we divide them into 2 blocks, where each block contains 3 DRGs. We set $r = 1$ for the first block and $r = 2$ for the second block.

(ii) 1-2-3, 1-2-5: The 6 DRGs are divided into 3 blocks, and each block consists of 2 DRGs. We fixed the first block and second block with $r = 1$ and $r = 2$, respectively. Then, we gradually change the dilation rate of the third block as $r = 3, 5$.

TABLE II

RESULTS OF DIFFERENT VARIANT DILATION RATE SCHEMES WITH A SCALE FACTOR OF 3. THE **TEXT** INDICATES THE BEST PERFORMANCE

Scheme	SET14	BSDS100
1-2	29.94 / 0.834	28.90 / 0.798
1-2-3	29.96 / 0.834	28.92 / 0.798
1-2-5	29.99 / 0.835	28.94 / 0.800
1-3-4	**30.03 / 0.837**	**28.95 / 0.801**
Baseline	29.91 / 0.832	28.87 / 0.796

TABLE III

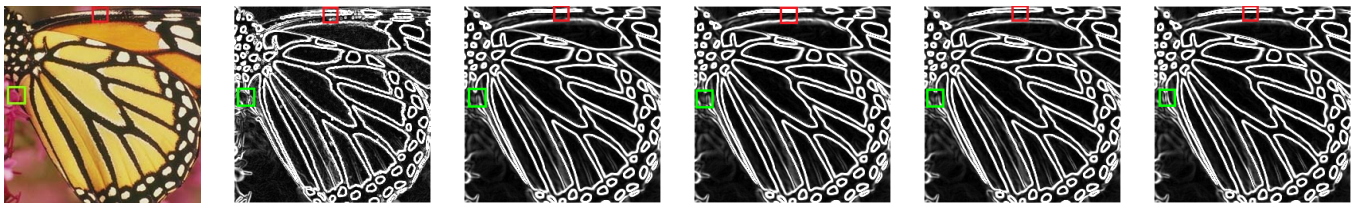
ABLATION STUDY OF GSM AND AIFS WITH THE SCALE FACTOR OF 4. THE **TEXT** INDICATES THE BEST PERFORMANCE

Method	Different combinations of GSM and AIFS			
SRU	×	✓	×	✓
AIFS	×	×	✓	✓
PSNR	28.09	28.17	28.16	**28.27**
SSIM	0.698	0.771	0.771	**0.773**
Time(sec.)	0.134	0.117	0.112	**0.108**

(iii) 1-3-4: Based on (ii), we set $r = 3$ for the second block and $r = 4$ for the third block to further expand the receptive field of the network.

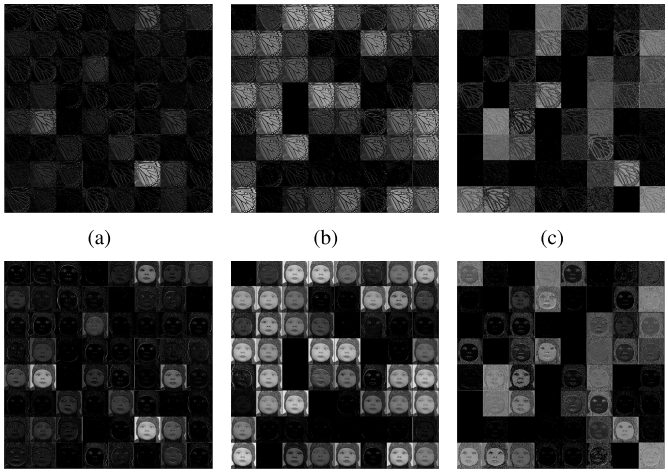
All of the trained models are compared on the SET14 and BSDS100 datasets with the scale factor of 3 and then illustrated in Table II in terms of PSNR and SSIM performance. With the same number of parameters, we can see that increasing the receptive field size generally yields higher performance compared to the baseline model. In addition, with 3 different dilation rates from 1-2-3 to 1-3-4, we observe that using a larger dilation rate to expand the receptive fields can consistently improve the performance. Since the dilation rate scheme 1-3-4 achieves the best performance on the two datasets compared to the other schemes, we select the 1-3-4 scheme to be incorporated into our baseline model FilterNet_T6M2 for our final network.

3) Study of GSM and AIFS: We now conduct detailed analyses on the proposed components, *i.e.*, the gated selective mechanism (GSM) and the adaptive information fusion structure (AIFS), for a better understanding of our proposed FilterNet. We fixed the basic design: 6 DRGs, 2 dilated residual units per DRG, and dilation rate scheme 1-3-4. We then add one of the GSM or AIFS to the basic architecture. We further add both the components to the basic design, resulting in our final network. Table III shows the ablation study of GSM and AIFS with the scale factor of 4 in terms of PSNR, SSIM and inference time on SET14. We observe that singly adding one of the components into the network can achieve higher performance on PSNR. In addition, the reconstruction speed can be further accelerated by the two components. With both components GSM and AIFS, our model can obtain higher SSIM and the PSNR of 0.18 dB compared with the first example network. To demonstrate that the GSM and AIFS can make the networks learn more high-frequency information for image detail recovery, we illustrate the visual comparisons of edge maps produced by the four architectures for 4× SR.



(a) (b) (c) (d) (e) (f)

Fig. 6. Visual comparisons of edge maps produced by the four architectures for 4× SR: (a) ground truth, (b) original edge map of the ground truth image, (c) w/o SRU or AIFS, (d) w/o AIFS, (e) w/o SRU, and (f) with both SRU and AIFS.



(a) (b) (c) (d) (e) (f)

Fig. 7. The visualization of the feature maps in the SRU, where the first two columns are the output of the two 3×3 dilated convolutional layers, and the third column are the feature maps of the last 1×1 convolutional layer in the GSM. Top: the "butterfly" image from SET5, bottom: the "baby" image from SET5.

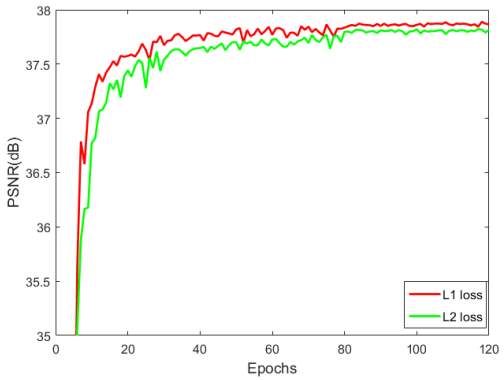
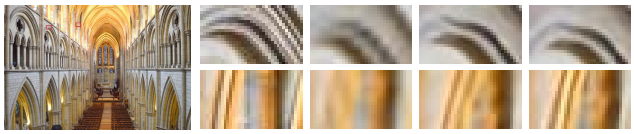


Fig. 8. Convergence analysis of our proposed FilterNet with L1 and L2 loss functions.



Ground-truth HR HR LR L2 loss L1 loss

Fig. 9. Visual comparisons of the L1 and L2 loss functions utilized to train our FilterNet for 3× SR.

As shown in Fig. 6, the network with SRU (Fig. 6(d)) or AIFS (Fig. 6(e)) provide better local detailed information and clearer boundaries than the network without the two components in Fig. 6(c). We can obviously observe that Fig. 6(f) achieves the best texture detail and sharpest edges with utilizing the GSM and AIFS together in our network.

To demonstrate that our proposed method can adaptively focus on high-frequency information in the training process, we need to inspect the output of the internal layers in SRU, as shown in Fig. 3(c), which includes two dilated convolutional layers and the GSM. For a better understanding the adaptive selection process, we visualize the feature maps of the first 3×3 dilated convolutional layers to show the extracted features containing both abundant low- and high-frequency components. Then, we illustrate the feature maps of the last 1×1 convolutional layers, which can be seen in the output of our GSM. As illustrated in Fig. 7, the first two dilated convolutional layers can effectively learn the contextual information of the LR input with a larger receptive field. In addition, in the third column of Fig. 7 (Fig. 7(c) and Fig. 7(f)), we can obviously see that the output of the GSM contains more edge detail and textures but fewer smooth low-frequency components, which demonstrates that the proposed GSM can effectively concentrate on the high-frequency information and suppress the low-frequency information.

4) Discussion of the Loss Functions: In this subsection, we discuss the loss functions for training our models. We first train our FilterNet with L1 loss. Then, we change the L1 loss function into L2 loss function to train our models. We visualize the convergence process of the two training strategies on SET5 for 2× SR in Fig. 8. We observe that the L1 loss can achieve a faster convergence speed and higher PSNR performance compared with L2 loss. To demonstrate the image visual quality of the two methods, in Fig. 9, we show the HR images reconstructed by the models with L1 and L2 loss functions for 3× SR. We find that training a network with L1 loss can guarantee better image quality in image SR.

D. Comparisons With the State-of-the-Arts

1) Objective Evaluation: We evaluate our proposed FilterNet with three commonly used image quality metrics: peak signal-to-noise ratio (PSNR), structural similarity index (SSIM), and information fidelity criterion (IFC) [57]. We compare the proposed method for 2×, 3× and 4× SR with 10 state-of-the-art methods including A+ [15], SelfExSR [18],

TABLE IV

QUANTITATIVE EVALUATION OF STATE-OF-THE-ART SR ALGORITHMS: AVERAGE PSNR AND THE CORRESPONDING STANDARD DEVIATION FOR SCALE FACTORS 2, 3 AND 4. **RED** TEXT INDICATES THE BEST AND BLUE TEXT INDICATES THE SECOND BEST PERFORMANCE

Datasets	Scale	Metric	Bicubic	A+	SelfEx	SRCNN	VDSR	DRCN	LapSRN	DRRN	DSRN	RAN	MS-LapSRN	**FilterNet**
SET5	2	PSNR	33.65	36.60	36.60	36.65	37.53	37.63	37.52	37.74	37.66	37.58	37.78	**37.86**
		Stdev	3.986	3.446	3.446	3.195	3.138	3.440	3.305	3.163	N.A.	N.A.	3.227	3.254
	3	PSNR	30.39	32.62	32.66	32.76	33.66	33.82	33.78	34.03	33.88	33.71	34.06	**34.08**
		Stdev	4.089	3.454	3.594	3.142	2.619	2.710	2.615	2.648	N.A.	N.A.	2.624	2.693
	4	PSNR	28.42	30.32	30.34	30.49	31.35	31.53	31.54	31.68	31.40	31.43	**31.74**	**31.74**
		Stdev	4.125	3.674	4.200	3.822	2.732	2.700	2.784	2.693	N.A.	N.A.	2.572	2.587
SET14	2	PSNR	30.34	32.32	32.24	32.29	32.97	32.98	33.08	33.23	33.15	33.10	33.28	**33.34**
		Stdev	3.558	3.936	4.120	4.025	4.137	4.174	4.118	4.236	N.A.	N.A.	4.206	4.249
	3	PSNR	27.64	29.15	29.18	29.41	29.77	29.76	29.87	29.96	**30.26**	29.84	29.97	30.03
		Stdev	3.421	3.883	3.917	3.839	4.089	4.158	3.964	4.177	N.A.	N.A.	4.027	4.095
	4	PSNR	26.10	27.34	27.41	27.61	28.03	28.04	28.19	28.21	28.07	28.09	28.26	**28.27**
		Stdev	3.297	3.705	3.735	3.616	3.836	3.892	3.920	3.914	N.A.	N.A.	3.948	3.936
BSDS100	2	PSNR	29.56	31.24	31.20	31.36	31.90	31.85	31.80	32.05	**32.10**	31.92	32.05	32.09
		Stdev	3.525	4.001	3.991	3.891	4.186	4.190	4.198	4.225	N.A.	N.A.	4.257	4.272
	3	PSNR	27.21	28.31	28.30	28.41	28.83	28.80	28.81	**28.95**	28.81	28.84	28.93	**28.95**
		Stdev	3.345	3.717	3.750	3.616	3.934	3.952	3.941	3.998	N.A.	N.A.	3.982	4.001
	4	PSNR	25.96	26.83	26.84	26.91	27.29	27.24	27.32	27.38	27.25	27.31	**27.43**	27.39
		Stdev	3.242	3.504	3.551	3.418	3.707	3.688	3.742	3.775	N.A.	N.A.	3.789	3.752
URBAN100	2	PSNR	26.88	29.25	29.55	29.52	30.77	30.76	30.41	31.23	30.97	N.A.	31.15	**31.24**
		Stdev	2.999	3.482	3.454	3.343	4.616	4.672	4.649	4.654	N.A.	N.A.	4.712	4.693
	3	PSNR	24.46	26.05	26.45	26.24	27.14	27.15	27.06	27.53	27.16	N.A.	27.47	**27.55**
		Stdev	3.381	3.543	3.768	3.834	4.194	4.240	4.223	4.281	N.A.	N.A.	4.311	4.379
	4	PSNR	23.15	24.34	24.83	24.53	25.18	25.14	25.21	25.44	25.08	N.A.	25.51	**25.53**
		Stdev	3.204	3.582	3.650	3.433	3.862	3.849	3.931	3.932	N.A.	N.A.	3.995	3.984

TABLE V

QUANTITATIVE EVALUATION OF STATE-OF-THE-ART SR ALGORITHMS: AVERAGE SSIM/IFC FOR SCALE FACTORS 2, 3 AND 4. **RED** TEXT INDICATES THE BEST AND BLUE TEXT INDICATES THE SECOND BEST PERFORMANCE

Datasets	Scale	Metric	Bicubic	A+	SelfEx	SRCNN	VDSR	DRCN	LapSRN	DRRN	DSRN	RAN	MS-LapSRN	**FilterNet**
SET5	2	SSIM	0.930	0.955	0.955	0.954	0.958	0.959	0.959	0.959	0.959	0.959	0.960	**0.961**
		IFC	6.166	8.715	8.404	8.165	8.190	8.326	9.010	8.671	8.585	N.A.	**9.305**	9.261
	3	SSIM	0.868	0.909	0.910	0.908	0.921	0.922	0.921	0.924	0.922	0.922	0.924	**0.925**
		IFC	3.596	4.979	4.911	4.682	5.088	5.202	5.194	5.397	5.221	N.A.	5.390	**5.564**
	4	SSIM	0.810	0.860	0.862	0.862	0.882	0.884	0.885	0.889	0.883	0.885	0.889	**0.890**
		IFC	2.337	3.260	3.249	2.997	3.496	3.502	3.559	3.703	3.500	N.A.	3.749	**3.758**
SET14	2	SSIM	0.870	0.906	0.904	0.903	0.913	0.913	0.913	0.914	0.913	0.913	**0.915**	**0.915**
		IFC	6.126	8.200	8.018	7.829	7.878	8.025	8.505	8.320	8.169	N.A.	8.748	**8.824**
	3	SSIM	0.776	0.820	0.821	0.823	0.834	0.833	0.833	0.835	**0.837**	0.833	0.836	**0.837**
		IFC	3.491	4.545	4.505	4.373	4.606	4.686	4.665	4.878	4.892	N.A.	4.806	**5.036**
	4	SSIM	0.704	0.751	0.753	0.754	0.770	0.770	0.772	0.772	0.770	0.769	**0.774**	0.773
		IFC	2.246	2.961	2.952	2.767	3.071	3.066	3.147	3.252	3.147	N.A.	3.261	**3.321**
BSDS100	2	SSIM	0.844	0.887	0.887	0.888	0.896	0.894	0.895	0.897	0.897	0.896	0.898	**0.899**
		IFC	5.695	7.464	7.239	7.242	7.169	7.220	7.715	7.513	7.541	N.A.	**7.927**	7.860
	3	SSIM	0.740	0.785	0.786	0.787	0.798	0.797	0.797	0.800	0.797	0.798	0.802	**0.803**
		IFC	3.168	4.028	3.923	3.879	4.043	4.070	4.057	4.235	4.051	N.A.	4.154	**4.358**
	4	SSIM	0.669	0.711	0.713	0.712	0.726	0.724	0.728	0.728	0.724	0.726	**0.731**	0.729
		IFC	1.993	2.565	2.512	2.412	2.627	2.587	2.677	2.746	2.599	N.A.	2.755	**2.800**
URBAN100	2	SSIM	0.841	0.895	0.898	0.895	0.914	0.913	0.910	0.919	0.916	N.A.	0.919	**0.920**
		IFC	6.319	8.440	8.414	8.092	8.270	8.527	8.907	8.917	8.598	N.A.	9.406	**9.583**
	3	SSIM	0.736	0.799	0.810	0.800	0.829	0.828	0.827	**0.838**	0.828	N.A.	0.837	**0.838**
		IFC	3.661	4.883	4.988	4.630	5.045	5.187	5.168	5.456	5.172	N.A.	5.409	**5.603**
	4	SSIM	0.659	0.721	0.740	0.724	0.753	0.752	0.756	0.764	0.747	N.A.	**0.768**	**0.768**
		IFC	2.386	3.218	3.381	2.992	3.405	3.412	3.530	3.676	3.297	N.A.	3.727	**3.759**

SRCNN [20], VDSR [22], DRCN [23], LapSRN [31], DRRN [28], DSRN [33], RAN [37], and MS-LapSRN [32]. Since the models of LapSRN and MS-LapSRN trained with $4\times$ samples can handle the upsample scales of $3\times$, we test the two methods for $3\times$ SR by using their 2-level models. The results of DSRN [33] and RAN [37] are cited from their corresponding papers. Table IV shows the evaluation results in terms of PSNR and the corresponding standard deviation (Stdev). We observe that our FilterNet achieves comparable average PSNR performance against existing methods on all datasets. In addition, the standard deviations of all methods mainly distribute between $[2.5, 4.9]$ on different datasets. With the same dataset, the Stdev of our proposed FilterNet is very close to the compared state-of-the-art methods. This demonstrates that our method achieves as stable performance as other methods on different datasets. In Table V,

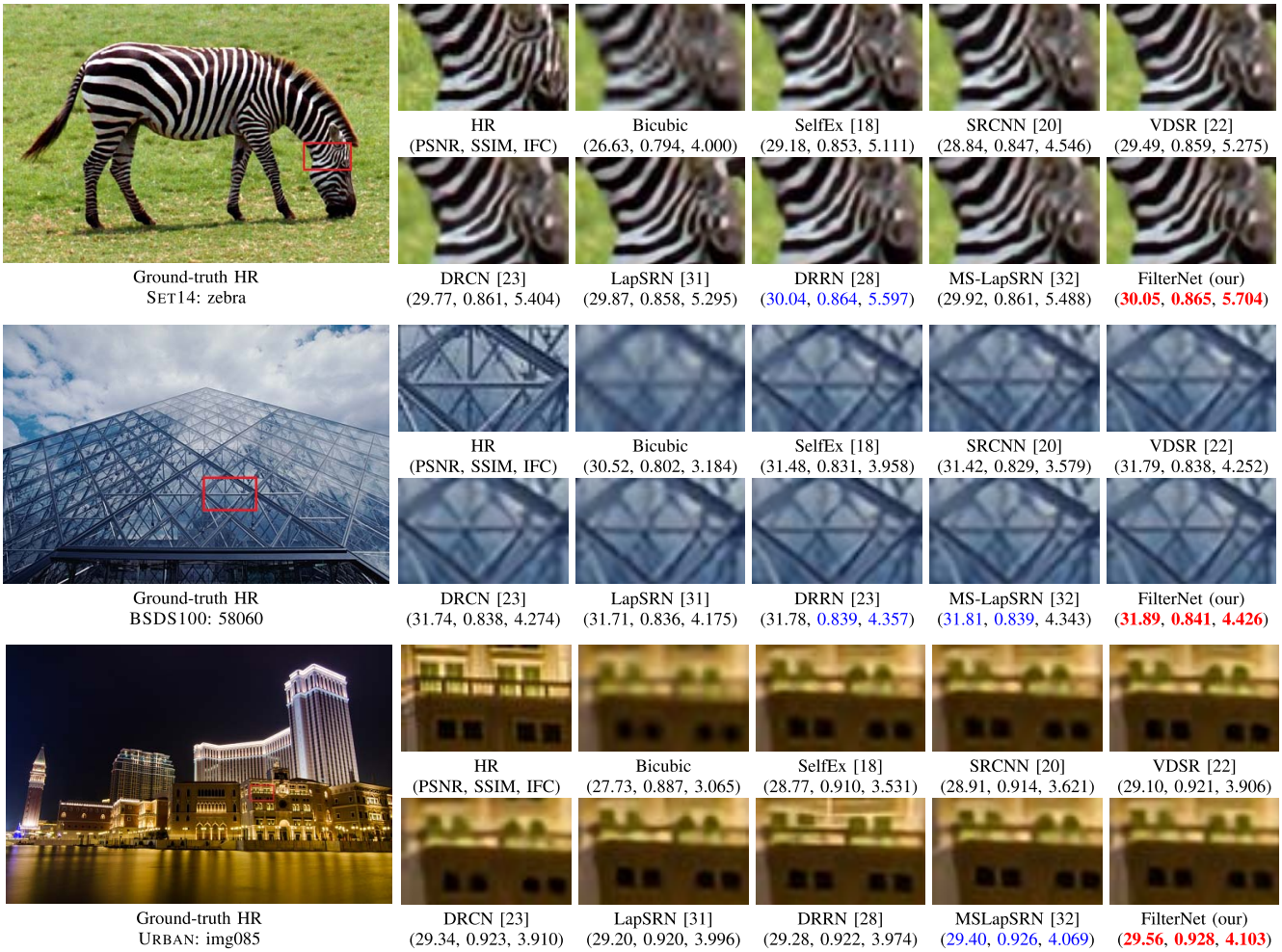


Fig. 10. Visual comparison of 3× SR on the SET14, BSDS100 and URBAN109 datasets.

we illustrate the evaluation results in terms of SSIM and IFC on all datasets. We observe that our proposed method performs favorably against the state-of-the-arts. Moreover, our algorithm achieves superior IFC values on most datasets, which has been claimed to be highly correlated with human perception of image SR. In particular, compared with the very deep models DRRN (up to 52 3×3 convolutional layers) and MS-LapSRN (up to 84 3×3 convolutional layers), the proposed FilterNet has a comparable quantitative performance with a shallower structure (only 32 3×3 convolutional layers).

2) Subjective Evaluation: To demonstrate the visual quality of our proposed method, we show the visual comparisons on the SET14, BSDS100 and URBAN100 for 3× SR in Fig. 10 and 4× SR in Fig. 10. For the image "zebra" from SET14 in Fig. 10, our method generates the HR image with clearer strips than those of the results produced by DRRN [28] and MS-LapSRN [32]. For the image "58060" from BSDS100, we can obviously see that our proposed method can produce the straightest and clearest lines compared with the other methods. For 4× SR, as shown in Fig. 11, our method can accurately produce textures, circles and parallel straight lines, whereas other methods generate results that still contain different extents of the fake information and noticeable artifacts.

E. Inference Time

We present our inference time and compare it with the state-of-the-art methods. All of the compared algorithms use the original public codes from the authors. We evaluate the runtime on the same machine with a 3.4 GHz Intel i7 CPU (128G RAM) and 1 NVIDIA Titan Xp GPU (12G Memory). Fig. 12 shows the trade-offs between the execution time and the PSNR performance on the SET5 dataset for 2× SR. Since the testing code of SRCNN [20] is implemented on a CPU, we rebuild the model as well as the VDSR [22] model in MatConvNet [58] with the same network parameters for evaluating the runtime on a GPU. As shown in Fig. 12, we can observe that SRCNN [20] achieves the fastest speed but very low reconstruction performance. Our proposed FilterNet can stride a balance between the reconstruction accuracy and runtime, which outperforms LapSRN [31] by 0.34 dB and the state-of-the-art method MS-LapSRN [32] by 0.08 dB with very similar execution times.

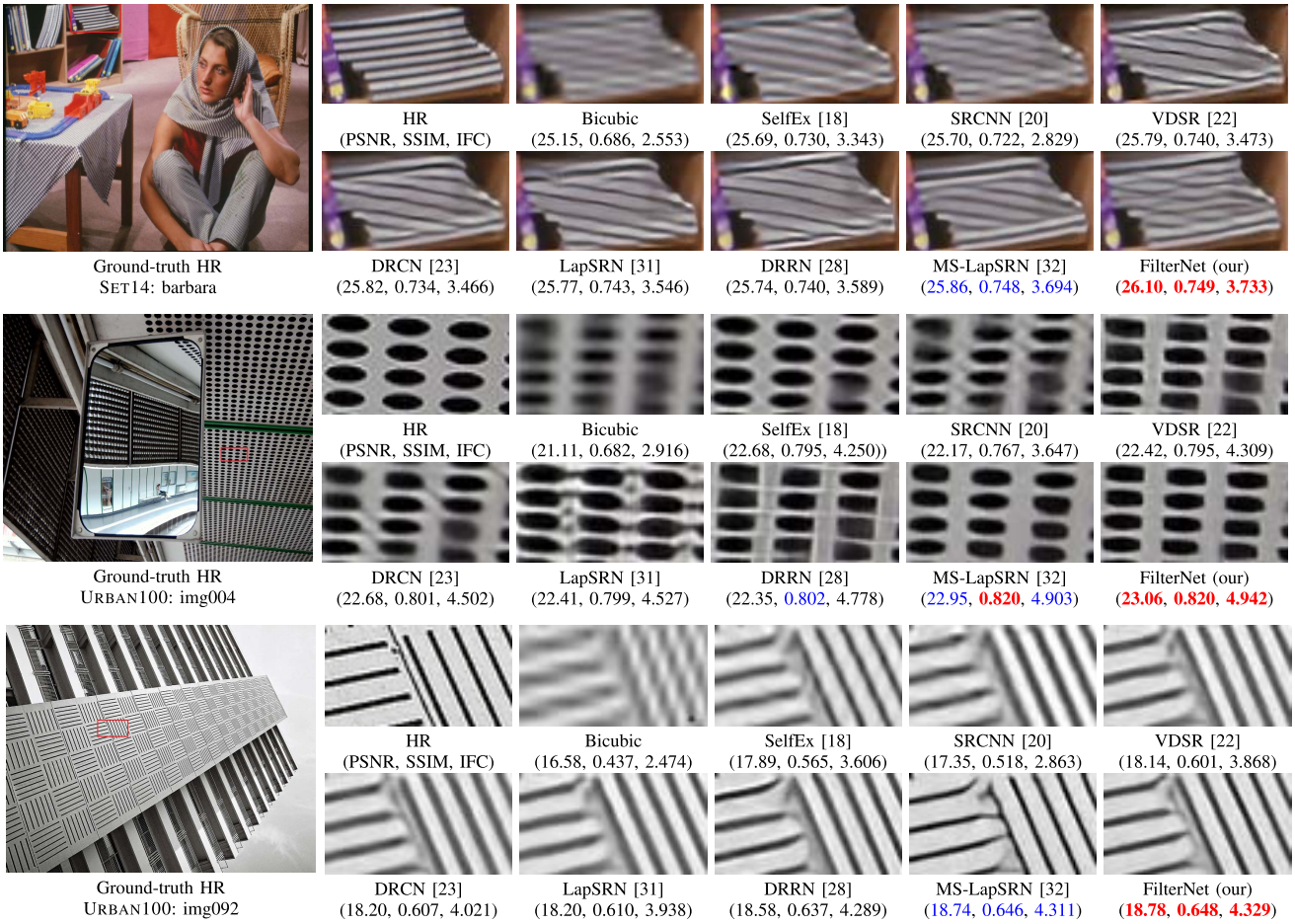


Fig. 11. Visual comparison for 4× SR on the SET14, and URBAN100 datasets.

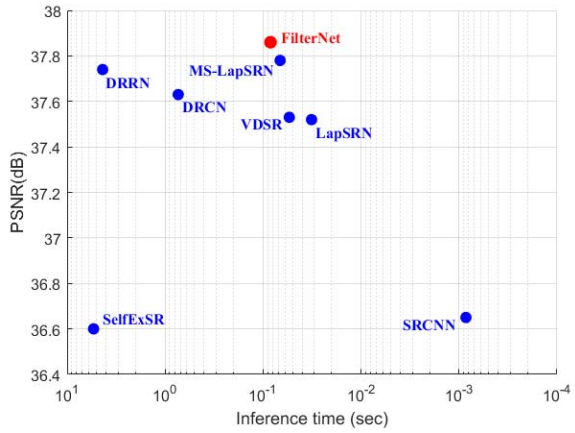


Fig. 12. PSNR performance versus runtime (evaluated in seconds). The results are evaluated on the SET5 dataset for 2× SR. The proposed FilterNet balances between reconstruction accuracy and inference time.

V. CONCLUSION

In this paper, we propose an adaptive information filtering network (FilterNet) for accurate and fast image super-resolution. In contrast to the existing methods that adopt full CNN to directly predict the HR images, the proposed FilterNet concentrates on more useful features and adaptively filters the redundant low-frequency information. The proposed FilterNet employs cascaded dilated residual groups (DRG) which consist of multiple selective residual units (SRU) with stacked style. With the SRUs, the network can efficiently exploit the contextual information of LR input images and adaptively concentrate on more useful information for high-frequency detail reconstruction. We present an adaptive information fusion structure (AIFS), which builds adaptively weighted skip connections among these DRGs to pass more useful features and improve the pixel-wise fitting capacity of the network. Comprehensive evaluations on benchmark datasets demonstrate that our method achieves superior performance compared with state-of-the-art methods in terms of quantitative and qualitative evaluations with promising inference time.

REFERENCES

[1] T. Blu, P. Thevenaz, and M. Unser, "Linear interpolation revitalized," *IEEE Trans. Image Process.*, vol. 13, no. 5, pp. 710–719, May 2004.
[2] X. Li and M. T. Orchard, "New edge-directed interpolation," *IEEE Trans. Image Process.*, vol. 10, no. 10, pp. 1521–1527, Oct. 2001.
[3] X. Li, H. He, R. Wang, and D. Tao, "Single image superresolution via directional group sparsity and directional features," *IEEE Trans. Image Process.*, vol. 24, no. 9, pp. 2874–2888, Sep. 2015.
[4] Y. Hu, N. Wang, D. Tao, X. Gao, and X. Li, "SERF: A simple, effective, robust, and fast image super-resolver from cascaded linear regression," *IEEE Trans. Image Process.*, vol. 25, no. 9, pp. 4091–4102, Sep. 2016.
[5] R. Fattal, "Image upsampling via imposed edge statistics," *ACM Trans. Graph.*, vol. 26, no. 3, pp. 95:1–95:8, Jul. 2007.

[6] Y.-W. Tai, S. Liu, M. S. Brown, and S. Lin, "Super resolution using edge prior and single image detail synthesis," in *Proc. Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2010, pp. 2400–2407.

[7] L. Wang, S. Xiang, G. Meng, H.-Y. Wu, and C. Pan, "Edge-directed single-image super-resolution via adaptive gradient magnitude self-interpolation," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 23, no. 8, pp. 1289–1299, Aug. 2013.

[8] H. Xu, G. Zhai, and X. Yang, "Single image super-resolution with detail enhancement based on local fractal analysis of gradient," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 23, no. 10, pp. 1740–1754, Oct. 2013.

[9] K. Zhang, X. Gao, D. Tao, and X. Li, "Single image super-resolution with non-local means and steering kernel regression," *IEEE Trans. Image Process.*, vol. 21, no. 11, pp. 4544–4556, Nov. 2012.

[10] A. Marquina and S. J. Osher, "Image super-resolution by TV-regularization and bregman iteration," *J. Sci. Comput.*, vol. 37, no. 3, pp. 367–382, Dec. 2008.

[11] J. Yang, J. Wright, T. S. Huang, and Y. Ma, "Image super-resolution via sparse representation," *IEEE Trans. Image Process.*, vol. 19, no. 11, pp. 2861–2873, Nov. 2010.

[12] W. Dong, L. Zhang, G. Shi, and X. Wu, "Image deblurring and super-resolution by adaptive sparse domain selection and adaptive regularization," *IEEE Trans. Image Process.*, vol. 20, no. 7, pp. 1838–1857, Jul. 2011.

[13] Z. Zhu, F. Guo, H. Yu, and C. Chen, "Fast single image super-resolution via self-example learning and sparse representation," *IEEE Trans. Multimedia*, vol. 16, no. 8, pp. 2178–2190, Dec. 2014.

[14] R. Timofte, V. De Smet, and L. Van Gool, "Anchored neighborhood regression for fast example-based super-resolution," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 1920–1927.

[15] R. Timofte, V. De Smet, and L. Van Gool, "A+: Adjusted anchored neighborhood regression for fast super-resolution," in *Proc. Asian Conf. Comput. Vis.*, Nov. 2014, pp. 111–126.

[16] T. Peleg and M. Elad, "A statistical prediction model based on sparse representations for single image super-resolution," *IEEE Trans. Image Process.*, vol. 23, no. 6, pp. 2569–2582, Jun. 2014.

[17] Z. Xiong, D. Xu, X. Sun, and F. Wu, "Example-based super-resolution with soft information and decision," *IEEE Trans. Multimedia*, vol. 15, no. 6, pp. 1458–1465, Oct. 2013.

[18] J.-B. Huang, A. Singh, and N. Ahuja, "Single image super-resolution from transformed self-exemplars," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 5197–5206.

[19] S. Schulter, C. Leistner, and H. Bischof, "Fast and accurate image upscaling with super-resolution forests," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 3791–3799.

[20] C. Dong, C. C. Loy, K. He, and X. Tang, "Image super-resolution using deep convolutional networks," *IEEE Trans. Pattern. Anal. Mach. Intell.*, vol. 38, no. 2, pp. 295–307, Feb. 2016.

[21] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Learn. Represent.*, Apr. 2015, pp. 1–14.

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

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

[24] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 770–778.

[25] X. Mao, C. Shen, and Y. Yang, "Image restoration using very deep convolutional encoder-decoder networks with symmetric skip connections," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 2802–2810.

[26] K. Zhang, W. Zuo, Y. Chen, D. Meng, and L. Zhang, "Beyond a Gaussian denoiser: Residual learning of deep CNN for image denoising," *IEEE Trans. Image Process.*, vol. 26, no. 7, pp. 3142–3155, Jul. 2017.

[27] W. Yang *et al.*, "Deep edge guided recurrent residual learning for image super-resolution," *IEEE Trans. Image Process.*, vol. 26, no. 12, pp. 5895–5907, Dec. 2017.

[28] Y. Tai, J. Yang, and X. Liu, "Image super-resolution via deep recursive residual network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 2790–2798.

[29] Y. Tai, J. Yang, X. Liu, and C. Xu, "MemNet: A persistent memory network for image restoration," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 4549–4557.

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

[31] W. Lai, J.-B. Huang, N. Ahuja, and M.-H. Yang, "Deep laplacian pyramid networks for fast and accurate super-resolution" in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.* Jul. 2017, pp. 5835–5843.

[32] W.-S. Lai, J.-B. Huang, N. Ahuja, and M.-H. Yang, "Fast and accurate image super-resolution with deep laplacian pyramid networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, to be pubilshed.

[33] W. Han, S. Chang, D. Liu, M. Yu, M. Witbrock, and T. S. Huang. (2018). "Image super-resolution via dual-state recurrent networks." [Online]. Available: https://arxiv.org/abs/1805.02704

[34] Z. Hui, X. Wang, and X. Gao. (2018). "Fast and accurate single image super-resolution via information distillation network." [Online]. Available: https://arxiv.org/abs/1803.09454

[35] W. Shi *et al.*, "Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network," in *Proc. IEEE Conf. Vis. Pattern. Recognit.*, Jun. 2016, pp. 1874–1883.

[36] C. Ledig *et al.*, "Photo-realistic single image super-resolution using a generative adversarial network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 105–114.

[37] Y. Wang, L. Wang, H. Wang, and P. Li, "RAN: Resolution-aware network for image super-resolution," *IEEE Trans. Circuits Syst. Video Technol.*, to be pubilshed.

[38] M. Holschneider, R. Kronland-Martinet, J. Morlet, and P. Tchamitchian, "A real-time algorithm for signal analysis with the help of the wavelet transform," *Wavelets*, pp. 286–297, Jan. 1989.

[39] M. J. Shensa, "The discrete wavelet transform: Wedding the atrous and mallat algorithms," *IEEE Trans. Signal Process.*, vol. 40, no. 10, pp. 2464–2482, Oct. 1992.

[40] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," in *Proc. ICLR*, Apr. 2016, pp. 1–13.

[41] L. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, Apr. 2018.

[42] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam. (2017). "Rethinking atrous convolution for semantic image segmentation." [Online]. Available: https://arxiv.org/abs/1706.05587

[43] J. Dai, Y. Li, K. He, and J. Sun. (2016). "R-FCN: Object detection via region-based fully convolutional networks." [Online]. Available: https://arxiv.org/abs/1605.06409

[44] W. Liu *et al.* (2015). "SSD: Single shot multibox detector." [Online]. Available: https://arxiv.org/abs/1512.02325

[45] F. Yu, V. Koltun, and T. Funkhouser, "Dilated residual networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 636–644.

[46] Y. Li, X. Zhang, and D. Chen. (2018). "CSRNet: Dilated convolutional neural networks for understanding the highly congested scenes." [Online]. Available: https://arxiv.org/abs/1802.10062

[47] V. Nair and G. Hinton, "Rectified linear units improve restricted boltzmann machines," in *Proc. Int. Conf. Mach. Learn.*, 2010, pp. 807–814.

[48] B. Lim, S. Son, H. Kim, S. Nah, and K. M. Lee, "Enhanced deep residual networks for single image super-resolution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, Jul. 2017, pp. 1132–1140.

[49] P. Arbelaez, M. Maire, C. Fowlkes, and J. Malik, "Contour detection and hierarchical image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 5, pp. 898–916, May 2011.

[50] R. Timofte *et al.*, "NTIRE 2017 challenge on single image super-resolution: Methods and results," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshop*, Jul. 2017, pp. 1110–1121.

[51] M. Bevilacqua, A. Roumy, C. Guillemot, and M. L. Alberi-Morel, "Low-complexity single-image super-resolution based on nonnegative neighbor embedding," in *Proc. Brit. Mach. Vis. Conf.*, 2012, pp. 135.1–135.10.

[52] R. Zeyde, M. Elad, and M. Protter, "On single image scale-up using sparse-representations," in *Proc. Int. Conf. Curves Surf.*, Jun. 2012, pp. 711–730.

[53] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 1026–1034.

[54] Y. Jia *et al.* (2014). "Caffe: Convolutional architecture for fast feature embedding." [Online]. Available: https://arxiv.org/abs/1408.5093

[55] D. Kingma and J. Ba. (2014). "Adam: A method for stochastic optimization." [Online]. Available: https://arxiv.org/abs/1412.6980
[56] H. Zhao, O. Gallo, I. Frosio, and J. Kautz. (2015). "Loss functions for neural networks for image processing." [Online]. Available: https://arxiv.org/abs/1511.08861
[57] H. R. Sheikh, A. C. Bovik, and G. de Veciana, "An information fidelity criterion for image quality assessment using natural scene statistics," *IEEE Trans. Image Process.*, vol. 14, no. 12, pp. 2117–2128, Dec. 2005.
[58] A. Vedaldi and K. Lenc, "MatConvNet: Convolutional neural networks for MATLAB," in *Proc. ACM Int. Conf. Multimedia.*, Oct. 2015, pp. 689–692.

Feng Li received the B.S. degree from Anhui Normal University, China, in 2012. He is currently pursuing the Ph.D. degree with the Institute of Information Science, Beijing Jiaotong University, Beijing, China. His research interests include image and video compression, image and video super resolution, computer vision, and deep learning.

Huihui Bai received the B.S. Ph.D. degrees from Beijing Jiaotong University, China, in 2001 and 2008, respectively. She is currently a Professor with Beijing Jiaotong University. She has been involved in research and development 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 (M'06–SM'12) received the B.S. degree from the Radio Engineering Department, Fuzhou University, China, in 1989, the M.E. degree from the Radio Engineering Department, Southeast University, Nanjing, China, in 1992, and the Ph.D. degree from the Institute of Information Science, Beijing Jiaotong University (BJTU), China, in 1996. He became an Associate Professor with BJTU in 1998 and a Professor in 2001, where he is currently the Director of the Institute of Information Science. 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. His current research interests include image/video coding, digital watermarking and forensics, and video analysis and understanding. 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 the Ministry of Education of China in 2013. He serves on the editorial boards of several international journals, including as an Associate Editor of the IEEE TRANSACTIONS ON CYBERNETICS and the IEEE SIGNAL PROCESSING LETTERS and an Area Editor of *Signal Processing: Image Communication* (Elsevier).