Large Factor Image Super-Resolution with Cascaded Convolutional Neural Networks

Dongyang Zhang, Jie Shao, Zhenwen Liang, Lianli Gao, and Heng Tao Shen

***Abstract*—Recently, convolutional neural networks (CNNs) have attracted considerable attention in single image super-resolution (SISR) and have enabled great performance improvements. However, most of the existing methods super-resolve input images to the desired size with an interpolation operation during the beginning stage, which brings about heavy aliasing artifacts and high computational costs. Especially for large upsampling factors (e.g., 8×), it remains a challenge to restore high-quality results for deeply degraded images. To tackle this problem, we propose a cascaded super-resolution convolutional neural network (CSRCNN), which takes a single low-resolution (LR) image as an input and reconstructs high-resolution (HR) images in a progressive way. At each cascaded level, to help converge and improve the accuracy, a novel U-net based block with backprojection is first introduced, which exploits the mutual relation between HR and LR feature spaces. A refined block following the U-net block is also used to reconstruct the realistic texture details. In addition, we naturally utilize the strategy of curriculum learning, organizing the learning process from easy (small factors) to hard (large factors). Comprehensive experiments on benchmark datasets demonstrate that the proposed network achieves superior results compared with those of other state-of-the-art methods, particularly with the 8× upsampling factor.**

***Index Terms*—Cascaded architecture, convolutional neural networks, image super-resolution**

I. Introduction

IMAGE super-resolution (SR) is fundamental to many problems in image processing and computer vision. Single image super-resolution (SISR) aims to recover a high-resolution (HR) image from a low-resolution (LR) image [1]. SR has a wide spectrum of applications, and it is especially useful to enhance the quality of large screen displays, such as in medical imaging [2], satellite imaging [3], security and surveillance [4]. In addition, the recovery from low resolution to high correspondence can reduce network bandwidth and storage requirements in the Internet environment [5], which is of significance to large video transmission.

Prediction methods based on interpolation (linear or bicubic) [1] represent the first attempts to address the SISR problem. Despite being fast and simple, these methods do not learn any prior information, and easily lead to overly smoothed textures. Afterwards, Farsiu et al. [6] proposed a method that focuses on edge preservation to partially handle this issue, and Xu et al. [7] used convolutional principal component analysis (CPCA) and random matching for self-learning super-resolution. Nevertheless, SR is an inherently underdetermined inverse problem where there are varieties of plausible high-resolution outputs for any single LR input. Therefore, determining how to constrain the transformation space by prior information is critical to recover the missing high-frequency details in the original HR image. To this end, an example-based strategy [8] adopted by recent advanced methods provides the prior information. Many powerful algorithms aim to establish complex nonlinear mapping between LR and HR patches, and they usually rely on large quantities of training data, such as [9]. More sophisticated algorithms exploit internal similarities of a given image [10], [11] or effectively learn mapping functions between external low-resolution and high-resolution counterparts [12]–[14]. Among the external example-based SR algorithms, a sparse representation method [15]–[17] is the most representative. However, this method involves several steps in its solution pipeline, and integration of all of these steps into one step with all of its parts being optimizable can be achieved by a neural network due to the strong data-fitting capability. In fact, the architecture of a neural network is inspired by sparse coding, and a typical practice is to compute a sequence of feature maps from the LR image through a convolution kernel, culminating with upsampling layers to increase the resolution of feature maps and ultimately restore the HR image, such as SRCNN [13] and ESPCN [14].

With the development of the convolutional neural network (CNN), many computer vision tasks have achieved fascinating

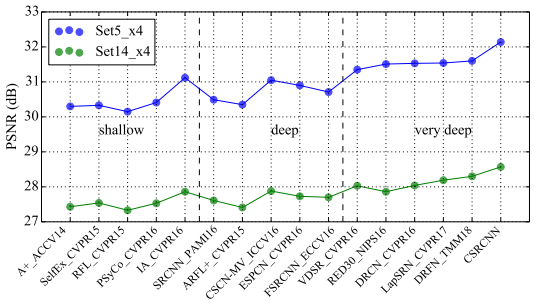


Fig. 1. From left to right: some shallow models and CNN-based models proposed in recent years with increasing depth. The performance of these models is measured by the peak signal-to-noise ratio (PSNR) for the upsampling factor 4× on the benchmark datasets Set5 and Set14.

D. Zhang, J. Shao, Z. Liang, L. Gao and H. T. Shen are with the Center for Future Media, School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu, 611731, China. E-mail: {dyzhang,zhenwenliang}@std.uestc.edu.cn, {shaojie,lianli.gao,shenhengtao}@uestc.edu.cn. J. Shao and H. T. Shen are also with Sichuan Artificial Intelligence Research Institute, Yibin, 644000, China. Corresponding author: Jie Shao.

This work is supported by National Natural Science Foundation of China (No. 61672133, No. 61832001 and No. 61632007) and Sichuan Science and Technology Program (No. 2019YFG0535).

TABLE I
SOME CNN-BASED SR ALGORITHMS ARE COMPARED. THE DEPTH STANDS FOR HOW MANY LAYERS ARE IN THE NETWORK, WHICH CONSISTS OF BOTH CONVOLUTION AND TRANSPOSED CONVOLUTION LAYERS. PROGRESSIVE RECONSTRUCTION REPRESENTS THE PREDICTION OF HR IMAGES IN MULTIPLE STAGES.

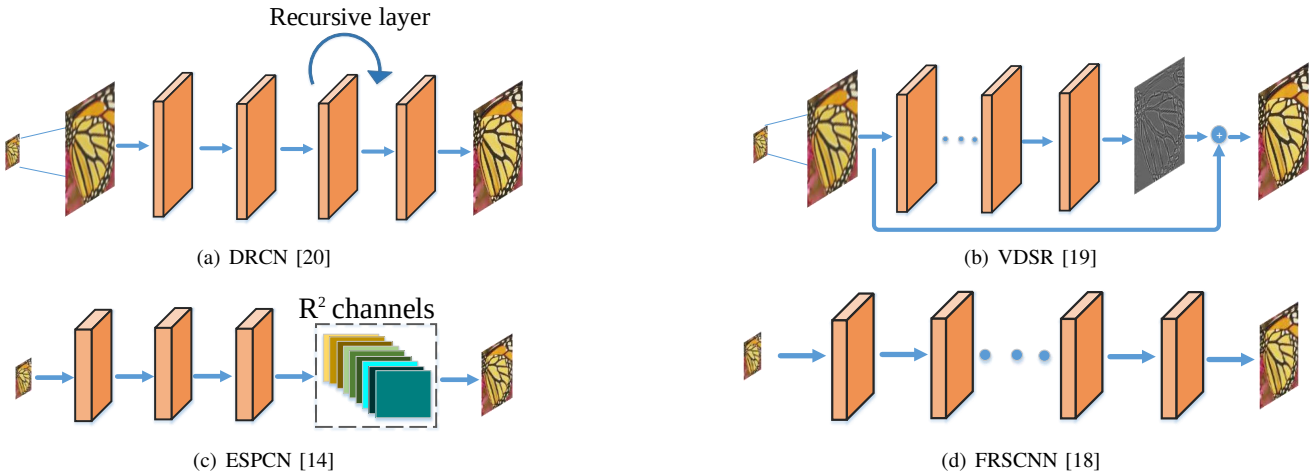
Method	Network input	Depth	Residual learning	Filters	Reconstruction	Loss function
SRCNN [13]	LR+bicubic	3	No	64	Direct	$L2$
FSRCNN [18]	LR	8	No	56	Direct	$L2$
ESPCN [14]	LR	3	No	64	Direct	$L2$
VDSR [19]	LR+bicubic	20	Yes	64	Direct	$L2$
DRCN [20]	LR+bicubic	5 (recursive)	No	256	Direct	$L2$
EDSR [21]	LR	52	Yes	256	Direct	$L1$
LapSRN [22]	LR	27	Yes	64	Progressive	$L1$
DRFN [23]	LR	65	Yes	64	Direct	$L1$
CSRCNN (ours)	LR	80	Yes	64	Progressive	$L1$

results. In contrast with the external example-based methods, the methods based on CNN implementation do not explicitly learn the dictionaries or manifolds [24] for modeling the patch space, but these can be achieved implicitly via hidden convolutional layers in an end-to-end manner. Super-resolution convolutional neural network (SRCNN) proposed by Dong et al. [13] represents the first application of the CNN architecture for the SR problem. Because of the merits of the CNN, using this type of network has become the mainstream practice to solve the SR problem, e.g., SRCNN [13], DRFN [23], CSN [25], ESPCN [14] and FSRCNN [18]. Some very deep networks have depths of more than 20 layers, such as VDSR [19], RED [26] and DRCN [20], all of which perform better than the shallow networks. Figure 1 shows some methods developed in recent years, in which we can find that the model depth is becoming increasingly deep, which also applies to the effect. For more detailed comparison, some key differences of the existing CNN-based methods and our proposed framework in this paper are given in Table I.

Although these methods above obtain remarkable results, there are still some major shortcomings. First, some existing methods [13], [19], [20] have a preprocessing step, in which an original low-resolution image must be upscaled to the desired spatial resolution with a bicubic or linear interpolation before being fed to the network for prediction (see Figure 2(a) and Figure 2(b)). It is well known that operations at high resolution increase the unnecessary computational cost. For an upsampling factor n, the computational cost of convolution with the interpolated LR image will be n^2 times that for the original LR image as an input. Second, the mean squared error (MSE) or $L2$ loss is the most widely used loss function for super-resolution, which is in agreement with the peak signal-to-noise ratio (PSNR) as one of the image evaluation criteria for SR. However, recent studies [21], [27], [28] demonstrate that, when compared with $L2$ loss, the $L1$ loss function is more powerful for performance and convergence. Consequently, an increasing number of works choose $L1$ loss to supervise the training to obtain less blurry results. Third, most existing SR methods only deal with small upsampling factors (e.g., $2\times$, $3\times$ or $4\times$), and as the upsampling factor becomes larger, these methods fail to restore high-quality visual details. Moreover, many methods super-resolve only at the end of the network through a subpixel convolution layer [14] or transposed convolution layer to upscale the resolution, which not only leads to checkerboard artifacts [29] due to simple concatenation of convolution layers but also increases the difficulties of training with large upsampling factors.

To handle the large factor image SR, we propose a cascaded super-resolution convolutional neural network (CSRCNN), which is progressive both in the architecture and training phase. Referring to Figure 3, different from the conventional network that directly super-resolves the LR input to the target resolution in an end-to-end manner, our network takes an LR image as an input and generates HR images through an intermediate level that performs a $2\times$ upsampling of the input from the previous level. At each level, we propose a novel backprojection U-net (BP-U) block and refined block to constitute a *hybrid link block* for accurate feature extraction and reconstruction. The traditional U-net upsamples and downsamples feature maps iteratively to maintain the resolution, and we improve the U-net with backprojection [30], which iteratively computes the reconstruction errors and reintegrates them for HR recovery. The refined block is a variant of the residual network [31], which can refine the easily lost information in HR space and accelerate convergence. In addition, we adapt our CSRCNN with an asymmetric structure, where more blocks are included in the lower levels to increase the receptive field, which is important for image restoration. The main contributions of our work can be summarized as follows:

- We propose a novel cascaded architecture in an asymmetric way, which is also progressive with respect to both architecture and training. At each level, we adopt residual learning by employing hybrid link blocks for feature extraction and reconstruction.
- Based on the traditional U-net, we propose a novel BP-U block with backprojection, which provides an error feedback mechanism to model the relation between LR space and HR space. Following is the refined block, and instead of normal convolution layer stacking, we combine the two different blocks to constitute a hybrid link block



(a) DRCN [20] (b) VDSR [19]

(c) ESPCN [14] (d) FRSCNN [18]

Fig. 2. Structure diagram of DRCN [20], VDSR [19], ESPCN [14], and FRSCNN [18]. DRCN [20] and VDSR [19] include a predefined upsampling operation to upscale LR input images to the size of the target by conventional interpolation, such as bicubic. ESPCN [14] and FRSCNN [18] perform convolution on LR features and reconstruct HR results only at the end of the processing pipeline through a subpixel convolution layer or transposed convolution layer.

for complex nonlinear mapping.

- Evaluation on benchmark datasets demonstrates that, compared with the other methods, our CSRCNN model with fewer parameters achieves compelling results in terms of accuracy and computational cost, particularly for large upsampling factors.

The remainder of this paper is organized as follows. Section II reviews the related work, and Section III introduces the proposed method. Section IV presents the experimental study. Finally, Section V concludes the paper.

II. Related Work

A. CNN-based Super-Resolution

With the success of AlexNet [32] in ImageNet [33], deep learning has become one of the most effective means to solve computer vision tasks. Super-resolution convolutional neural network (SRCNN), known as the pioneer CNN model for SISR, learns the nonlinear mapping between LR patches and HR patches via a convolutional network layer, which achieves significant performance compared with classical handcrafted methods. Afterwards, GoogleNet [34] revealed that the network depth is of crucial importance. Kim et al. proposed the deeply recursive convolutional network (DRCN) [20] (see Figure 2(a)), which employed a recursive structure to increase the receptiveness of the network field and avoid the expansion of network parameters at the same time. Kim et al. [19] proposed a deep model named VDSR (see Figure 2(b)), which exploited long-term contextual information with 20 stacked convolutional layers. To allow the gradient information to flow more fluently and accelerate convergence, VDSR adopted the idea of residual learning [31] and adjustable gradient clipping. Mao et al. [26] proposed symmetric skip connections within a 30-layer convolutional auto-encoder named RED for SISR, which alleviates the training difficulty. Based on DRCN, Tai et al. [35] proposed DRRN, combining the memory blocks and deeply recursive learning. Surprisingly, DRRN stacked the number of convolutional layers up to 52, but because of the recursive design, the parameters did not increase with the number of network layers. These methods achieve better performance than the previous simple CNN-based methods by a large margin, which also demonstrates that the golden words "the deeper the better" remain true in image super-resolution. Recently, attention mechanism has been embedded in CNN architecture for vision tasks, such as captioning [36], question answering [37] and cross-modal retrieval [38]–[40]. Attention has also been applied for image SR [41]. Similar to the work of ESPCN [14] and FSRCNN [18] corresponding to Figure 2(c) and Figure 2(d), the proposed CSRCNN also operates with low resolution to reduce the computational cost.

B. Large Factor Super-Resolution

A vast amount of literature has proposed algorithms for SR while only focusing on low upsampling factors (e.g., $2\times$, $3\times$, and $4\times$). However, for a large factor (e.g., $8\times$), few works can effectively address the problem, but in practice, large factor SR is often in great demand for small object recognition in traffic [42], sports [43], and surveillance [44]–[46] scenes. Lai et al. [22] proposed a Laplacian pyramidal network for SR, named LapSRN, which progressively reconstructed the sub-band residuals of HR images. Each pyramid level of LapSRN is supervised and can upsample $2\times$ of the pervious resolution by transposed convolutions. Moreover, LapSRN can accurately and rapidly generate multi-scale predictions in one feedforward pass. Lai et al. also proposed an improved model, named MsLapSRN [47], adopting the strategy of recursive architecture and multi-scale training. Recently, Yang et al. [23] proposed DRFN for the large factor SR problem. The authors first upsample raw features in front of the network by the transposed convolution, then use recurrent residual blocks to gradually recover high-frequency information in HR feature space, and finally use a convolutional layer to fuse feature maps from three different levels to reconstruct HR images. To obtain more photo-realistic results, ProGanSR [28] not only designed a progressive generator but also proposed a progressive discriminator following the adversarial strategy, which can

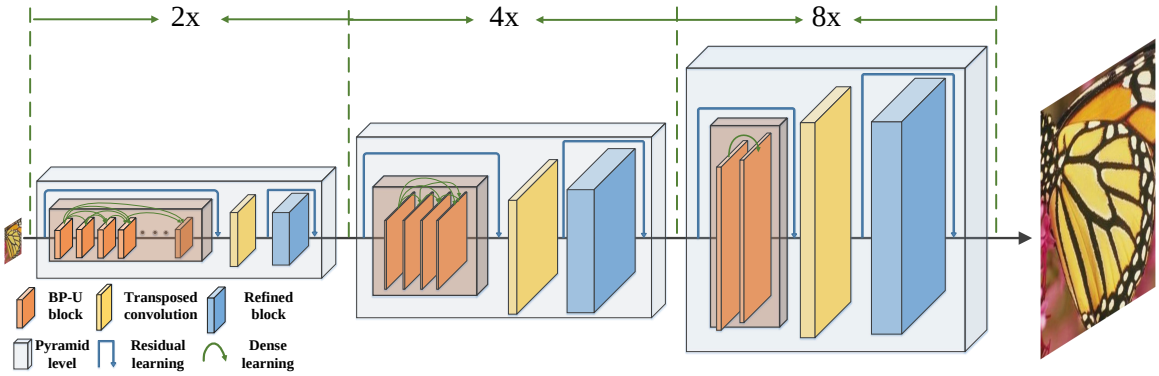


Fig. 3. An overview of the proposed asymmetric architecture. The BP-U block and refined block are adopted in our architecture. From low to high, our network is a typical cascaded pyramid structure where each level super-resolves a 2× factor using a transposed convolution layer.

simultaneously reconstruct high-quality visual results for all upsampling factors, particular 8×. In addition, with the aid of geometry prior knowledge, Kong et al. [48] proposed a novel Face Hallucination sub-Net (FHN) to super-resolve tiny face images to be 8× larger with high-fidelity facial details.

Also benefiting from the backprojection strategy, the deep backprojection networks (D-DBPN) proposed by Haris et al. [49] is most similar to our CSRCNN. DBPN exploits iterative upsampling and downsampling layers, providing an error feedback mechanism for projection errors at each stage. The authors also concatenate HR feature maps from all of the upsampling steps to build a dense connection for 8× SR, which achieves state-of-the-art results. Different from DBPN, we expand the power of U-nets that contain different resolution scales in one block while maintaining resolution between input and output. Improving the U-net with the error feedback mechanism, the proposed BP-U block is able to compute the errors between HR and LR spaces for better feature extraction. In contrast to DBPN, which adopts a dense connection to directly generate large factor SR, our method is progressive with respect to both the architecture and training, which significantly reduces the number of parameters and eases training.

C. GAN based Super-Resolution

Because MSE loss may fail to synthesize textures in accordance with human perception, perceptual loss is proposed by Johnson et al. [50], with the idea of ensuring the perceptual similarity between the estimated image and ground-truth in higher-level feature space such as VGG [51]. In addition to perceptual loss, generative adversarial networks (GANs) [52] have recently been shown to produce sharper results in the image generation task. Sønderby et al. [53] proposed AFFGAN, which can establish a connection between GANs and amortized variational inference for face image super-resolution. Zhang et al. [54] also took advantage of GANs and residual learning to make the face image super-resolution more perceptually friendly. Ledig et al. [55] developed an approach named SRGAN, inspired by [50]. The authors trained an adversarial network based on residual structure in conjunction with perceptual loss, which can recover photo-realistic textures from heavily downsampled images. Although these perception-based methods can synthesize realistic textures, their results have lower scores in terms of PSNR than the blurry samples optimized with MSE instead. In particular, Wang et al. [56] derived an Enhanced SRGAN (ESRGAN) by improving the network architecture, adversarial loss and perceptual loss of SRGAN, achieving better visual quality. Moreover, supervised by MSE loss, the proposed RRDB in [56] also exhibits superior performance over other PSNR-oriented methods. Overall, a comprehensive survey of the recent deep learning based SISR methods can be found in [57].

III. Proposed Method

In this section, we present the technical details of our proposed CSRCNN. We start with an overview of the architecture, followed by a description of the BP-U block and refined block. Finally, we introduce how to optimize the network with curriculum learning in a supervised manner as well as the implementation details for training.

A. Architecture

From Figure 3, we can see that our network is a typical cascaded structure that consists of three levels for the 8× upsampling factor. Following the principle of Laplacian pyramids, LapSRN [22] is a typical method that reconstructs HR images step by step. Similar to LapSRN, each level in our method performs a 2× upsampling of the input from the previous level, and multiple levels are assembled to super-resolve progressively. However, there are at least four differences between our CSRCNN and LapSRN. First, in LapSRN, the loss functions are computed at each level, which provides a form of intermediate supervision. Our method only calculates the loss at the end of the network. Second, LapSRN predicts a residual image that is used to reconstruct the HR image via addition with the upsampling image from the last level. Although we adopt residual learning, the HR image is generated by performing transposed convolution on feature maps instead of adding residual images. Third, referring to Figure 3, our network is an asymmetric pyramidal architecture where more blocks are in the lower levels and the higher levels

gradually decrease, which enables high upsampling factors while remaining efficient. However, each level in LapSRN is arranged with the same structure. Fourth, each level in LapSRN is just the combination of stacked convolution layers, while we propose the concept of the hybrid level including the BP-U block and refined block, which embraces both the superior reconstruction accuracy and high capacity of deep networks. The following part describes the two blocks in detail.

BP-U block: As Figure 4 shows, the proposed BP-U block is a typical U-net module consisting of convolutional and deconvolutional layers, which performs a chain of downsampling operations, followed by a chain of upsampling operations that process information from different spatial scales. Driven by the initial success of U-nets for semantic segmentation [58] and image translation [59], we first improve the U-net by backprojection which is well known as an efficient iterative procedure to minimize the reconstruction error [49]. Then, we expand the power of BP-U blocks by stacking them into deep architectures at each level. Initially, backprojection requires multiple LR inputs to iteratively reconstruct the errors. For the single image super-resolution, DBPN [49] proposed by Haris et al. adopts the iterative reconstruction procedure to guide the SR task, where the up-projection unit generates HR features and the down-projection unit then projects them back to the LR space and learns the errors. By creating an iterative up- and down-projection unit, this algorithm successfully exploits the nonlinear relation of LR and HR images. However, DBPN presents the shortcomings of an enormous number of parameters and difficulty in training. To overcome these deficiencies, we utilize U-nets in consideration of the following three aspects. First, the U-net is lightweight due to the encoder-decoder architecture, which is beneficial to reduce the GPU memory usage and time cost. Second, U-net has the ability to capture contextual information at multiple scales and propagate it to the higher resolution layers, which is critical for image reconstruction. Third, U-net can ease the training difficulty. Consequently, combining backprojection and the merits of U-nets, we integrate the up- and down-projection among feature space into the BP-U block for SISR.

Referring to Figure 4, the front part of the BP-U block is a convolutional layer with a 1×1 kernel, which can be viewed as feature pooling and dimensionality reduction to regulate the number of feature maps of the preceding blocks. All subsequent layers are 3×3 kernels, and the number of feature maps is fixed. The BP-U block is defined as follows:

$$downsampling \ \ 2\times: \qquad F_1 = (W_1 * F_0)\downarrow_2, \tag{1}$$

$$downsampling \ \ 2\times: \qquad F_2 = (W_2 * F_1)\downarrow_2, \tag{2}$$

$$concatenation: \qquad F_3 = [(W_3 * F_2), \quad F_2], \tag{3}$$

$$subtraction: \qquad F_4 = (W_4 * F_3)\uparrow_2 - F_1, \tag{4}$$

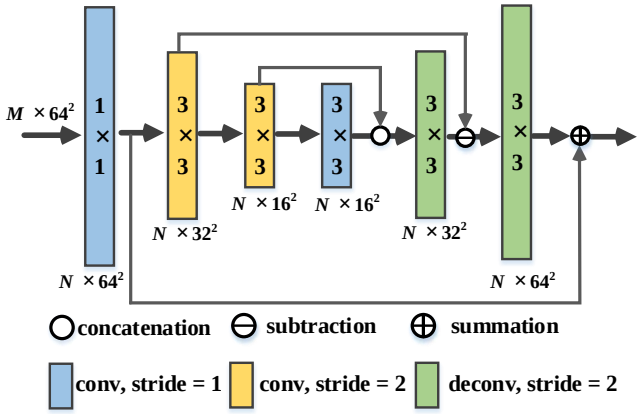


Fig. 4. An illustration of the proposed BP-U block. M is the number of input features. Across the first convolutional layer with a 1×1 filter, each layer has N output feature maps, which is helpful for dense learning. Different colors in the figure denote different kinds of convolutional layers with respect to filter and stride.

$$summation: \qquad F_5 = (W_5 * F_4)\uparrow_2 + F_0, \tag{5}$$

where W_i indicates the weights of the i-th convolutional layer and we omit the bias term and activation function for simplicity, $*$ stands for the spatial convolution operator, and $\uparrow_2$ and $\downarrow_2$ correspond to the convolutional layer and deconvolutional layer with stride 2, which perform the upsampling and downsampling with a 2× scale. $[F_1,\ F_2]$ refers to the concatenation of the feature maps.

F_0 is the regulated feature map through the first convolutional layer with a 1×1 kernel. After downsampling twice, F_2 is 4× smaller than F_0 in resolution. To mitigate high-frequency noise from the successive downsampling operation, F_2 is followed by a convolution that does not change the size. F_3 is obtained from a concatenation of F_2 and the current feature map, in consideration of enhancing information flow. After upsampling F_3 2×, an intermediate feature map is obtained. The residual F_4 is the difference between the intermediate feature map and F_1, which are both 2× smaller than the original size of F_0. Afterwards, we upsample the residual F_4 2× again and map it back to the original size. The final output F_5 is obtained by summing the upsampled F_4 and F_0. Notably, all of the layers in our BP-U block are specially designed to retain the original size of the feature maps at its output. Consequently, multiple U-net modules can be stacked without losing resolution, which can be used for dense learning.

Refined block: A refined block is proposed to help refine the local information flow from the BP-U block and alleviate the training difficulty. SRResNet [55] substantially increases the accuracy of the network by assembling a group of residual blocks with identical layout. The residual block in SRResNet is composed of Conv(3,3)-BN-ReLU-Conv(3,3)-BN with an outer residual connection. Following the recent practice in super-resolution, the works of [21], [28] remove all batch normalization layers since they remove flexibility from networks and consume more GPU memory. Taking into

account the above problems, we adopt the modified residual block composition: Conv(3,3)-ReLU-Conv(3,3). At the front of the refined block, we repeat the residual block four times, and the following is four Conv layers stacked with the ReLU activation function.

A group of BP-U blocks, together with a refined block and an upsampling layer, constitute each level in our CSRCNN. As illustrated in Figure 3, the group of BP-U blocks is arranged in dense connections, which have been proven to alleviate the gradient vanishing problem and encourage feature reuse. Notably, the BP-U blocks take effect in LR space to extract local dense features, and after a transposed convolutional layer, the feature maps are upsampled by a scale of 2. The following refined block performs a convolution operation on HR space and does not change the size of feature maps. Because of the different resolution between BP-U blocks and refined blocks at the same level, we impose residual learning on the two blocks separately. Moreover, from the perspective of the whole architecture, we assign more blocks in the lower levels, and the higher levels instead apply a general decrease in block numbers, resulting in the asymmetric structure.

B. Curriculum Learning

Bengio et al. [60] showed that it is easier for humans and animals to understand or learn if the learning samples present in a meaningful order which illustrates gradually more concepts, and gradually more complex ones. Recently, this kind of strategy, called curriculum learning, has been formalized to train CNNs in the context of deep learning. Obviously, in our cascaded framework, the low level corresponding to the small upsampling factor is easy to train. As the number of levels increases, both the difficulty of training and the upsampling factor gradually increase, which allows us to apply curriculum learning in a natural way.

The works of [28], [61] are both progressive with respect to architecture and training. They demonstrate that progressive training strategy not only greatly reduces the total training time but also yields performance gains for all included scales compared with simple multi-scale training. First, starting with the easiest task, we can only train the $2\times$ portion of the network. Then, we gradually blend a new level with the last trained level to proceed to a new training phase in the curriculum, which lessens the interference of the previously trained level. The whole process is like stacking blocks, where the output of the previous trained level is linearly fed to the new level.

C. Loss Functions

In this work, our goal is to learn a mapping function F via a convolutional neural network to generate an HR image $\tilde{I} = F_\theta(x)$ that is close to the ground-truth image I, where x and θ refer to the input LR image and network parameters, respectively. Conventional mean squared error (MSE) is widely used to supervise the training phase for general image restoration as defined below:

$$\mathcal{L}_{MSE} = \frac{1}{N}\sum_{i=1}^{N}\left\|I_i - \tilde{I}\right\|_2^2. \tag{6}$$

Some studies [21], [23], [47] point out that the mean absolute error (MAE), also called $L1$, offers a slight performance gain versus the original model trained with MSE for all scale factors. The MAE is formulated as follows:

$$\mathcal{L}_{MAE} = \frac{1}{N}\sum_{i=1}^{N}\left\|I_i - \tilde{I}\right\|_1. \tag{7}$$

In our cascaded structure with curriculum learning, the overall loss function is defined as:

$$\mathcal{L}(\tilde{I}, I; \theta) = \frac{1}{N}\sum_{i=1}^{N}\sum_{s=1}^{L} f(\tilde{I}_s^{(i)} - I_s^{(i)}), \tag{8}$$

where N represents the number of training images in a patch, while L represents the number of levels in the cascaded structure. During training, we jointly use the $L1$ and $L2$ losses. We first train the network with $L1$ loss and then fine-tune it by alternately using $L2$ and $L1$ losses, empirically showing an improvement in the results. The same phenomenon was reported by [62].

IV. Experiment

In this section, we compare the proposed CSRCNN with some state-of-the-art SR methods on benchmark datasets. Quantitative and qualitative comparisons are given according to the experiments, and the visual quality comparisons in terms of $4\times$ and $8\times$ are shown. We also perform an ablation study to analyze the contributions of different components. Finally, we discuss the limitations of the current method.

A. Datasets

In this work, we train all networks using images from DIV2K [63], which is widely used in recent learning-based SR methods. DIV2K is a newly proposed high-quality image dataset for image restoration tasks. There are 1000 high-resolution images (2K resolution) in the DIV2K dataset, with 800 for training and 100 for validation and testing, respectively. For testing, we use five general benchmark datasets, Set5, Set14, BSD100 [64], Urban100 [10] and Manga109 [65], which contain 5, 14, 100, 100 and 100 images, respectively. The five datasets exhibit various characteristics. Set5, Set14, and BSDS100 consist of natural scenes, Urban100 contains challenging urban scene images with details in different frequency bands, and Manga109 is collected from Japanese manga.

B. Implementation and Training Details

Because of the importance of data augmentation, we augment the training data both online and offline. In the offline augmentation, we mainly expand the training dataset by scaling. The original HR images are downscaled into several sizes (0.9, 0.8, 0.7, and 0.5) to produce the corresponding LR images. To ensure the accuracy of the result, all of these downsampling operations with bicubic interpolation are implemented in MATLAB. By doing so, we expand the training data 5 times. For the online augmentation, we first crop corresponding patches of small size from the expanded

TABLE II
ABLATION INVESTIGATION OF THE BP-U BLOCK, CURRICULUM TRAINING AND ASYMMETRIC ARCHITECTURE. THE RESULTS (PSNR) REFER TO 4× ENLARGEMENT ON THE SET14 DATASET.

Model	FSRCNN	Model1	Model2	Model3	Full
BP-U block	×	√	√	×	√
Curriculum training	×	√	×	√	√
Asymmetric architecture	×	×	√	√	√
PSNR (dB)	27.70	28.78	28.73	28.68	28.87

dataset and then rotate the cropped patches by 90, 180, and 270 degrees and flip them horizontally or vertically with a probability of fifty percent. The size of LR patches for training is 32×32, while the size of HR patches corresponds to the scaling factor. We set the minibatch size of stochastic gradient descent (SGD) to 32, momentum parameter to 0.9, and weight decay to 10^{-4}. It is noteworthy that RGB color channels are used for the input and output images during training. Following most existing networks, only the luminance channel (Y) is used for the evaluation of PSNR and SSIM [66], which are very important image quality assessments (higher means better). The reason is that we are not interested in the color space where the color information is stored in the CbCr channels, but only in their brightness in the Y channel.

In the proposed CSRCNN, each convolutional layer consists of 64 filters with a size of 3×3, and the padding is set to 1 to fix the size of the output feature maps. For weight initialization, we initialize the convolutional filters using the method of He et al. [67], which is suitable for networks utilizing rectified linear units (ReLUs). In our asymmetric architecture, for 4× enlargement, the first level consists of 6 BP-U blocks, and the second level consists of 4 BP-U blocks; for 8× enlargement, the three different levels consist of 6, 4, and 2 BP-U blocks, respectively. Regardless of the upscaling factor, each level only contains one refined block. With so many BP-U blocks and refined blocks stacked in a hybrid manner, we can increase the size of the receptive fields and learn the most complex mapping function. We stopped training when the loss reached a steady state, and approximately four days were required to train the network using an NVIDIA GTX 1080Ti graphics card.

C. Ablation Study

We explore three aspects of our design: the effect of the BP-U block, the effect of curriculum learning, and the effectiveness of an asymmetric architecture. The quantitative results are shown in Table II. Through experiments, the superiority of our design can be justified.

BP-U block: In this work, we propose a novel BP-U block by combining backprojection with the conventional U-net. To verify the power of the BP-U block, based on the structure in Figure 4, we remove the subtraction and concatenation to construct the traditional U block with skip connection. Then, the reconstructed U block takes the place of the BP-U block in the proposed network shown in Figure 3. We train the two kinds of networks equally over 200 epochs. Referring

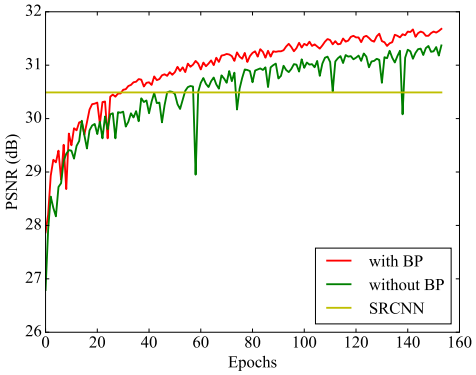


Fig. 5. Convergence analysis for backprojection. The network with backprojection converges faster and achieves the improvement in terms of PSNR.

TABLE III
COMPARISON OF THE DIFFERENT U BLOCKS FOR 4× AND 8× ENLARGEMENTS. T-U BLOCK STANDS FOR THE TRADITIONAL U BLOCK. RED INDICATES THE BEST PERFORMANCE.

		Set5		Set14	
Type	Scale	PSNR (dB)	SSIM	PSNR (dB)	SSIM
T-U block	4×	32.02	0.884	28.63	0.788
BP-U block	4×	32.22	0.896	28.78	0.794
T-U block	8×	27.06	0.781	25.07	0.644
BP-U block	8×	27.15	0.783	25.13	0.648

to Table III, which shows the performance comparison of two blocks in all cases, for 4× enlargement, the proposed BP-U block has gains that are 0.20 dB and 0.15 dB higher than those of the traditional U block on Set5 and Set14, respectively. Regarding the greater enlargement, 8×, the BP-U block performs 0.09 dB and 0.05 dB higher than the traditional U block on Set5 and Set14, respectively. At the same time, the convergence analysis of backprojection is also illustrated in Figure 5. Benefiting from the error feedback mechanism in backprojection, the network integrated with the BP-U block not only converges faster, but also achieves improved performance.

Curriculum learning: The concept of curriculum learning has been used in many fields, for example, [28], [61]. In this work, we employ the idea of curriculum learning in our CSRCNN for HR image generation. Now, we examine the effect of this modification in detail. Referring to Table II, a model is trained in end-to-end manner for 4× upsampling to quantify the benefit of curriculum learning. We find that curriculum learning enables a reconstruction quality gain of 0.14dB. On the other hand, we also visualize the average feature maps from each stage in Figure 6. These results suggest that the first stage of curriculum learning can capture the rough contour features of images. Although the resolution of feature maps is upsampled by continuous learning, average feature maps are gradually enhanced and obtain the features with a relatively clear contour profile. Moreover, curriculum learning also helps to reduce the training time, where the model with curriculum learning achieves higher accuracy than that without curriculum learning for the same epochs. The same evidence is also reported in [28].

Asymmetric architecture: In this study, we show how the

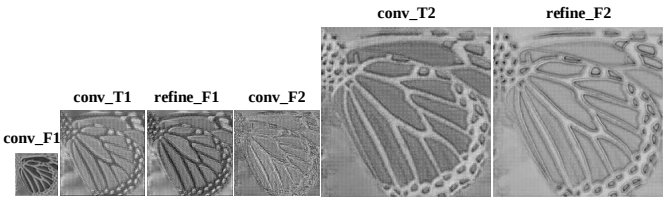


Fig. 6. Visualization of the average feature maps of different levels with the image of a "butterfly" in Set5 as input for 4× enlargement. The annotation above each picture denotes the different stages of our network.

asymmetric architecture benefits the reconstruction accuracy and GPU memory consumption. For a fair comparison, we design two different architectures for 8× unsampling while keeping the total number of BP-U blocks constant. For the asymmetric architecture, from low to high, the three levels have 6, 4, and 2 BP-U blocks, whereas for the symmetric architecture, each level has 4 BP-U blocks. The GPU memory consumption of the asymmetric architecture is approximately half that for the previous symmetric architecture, which is significant to train deeper networks with the limitation of memory. In addition, the time required to super-resolve an image is approximately half that for the previous symmetric architecture. This is because assigning more blocks for the lower levels brings about more computational powers in case of low resolution, which not only reduces the memory consumption but also increases the receptive field with respect to the input image.

D. Study of Model Depth

In the proposed network, we adopt a hybrid block connection at each level comprising the BP-U block and refined block, and the BP-U block is shown in detail in Figure 4. It has been shown that directly utilizing high-dimensional features is advantageous for image reconstruction [28], which is the reason why we add a refined block at the end of each level for features refined on super-resolved feature maps. To verify the superiority of this design and investigate the model depth, we use U_n to denote the number of BP-U blocks, R_n to denote the number of refined blocks in the network, and T to denote the transposed convolution to upsample feature maps $2\times$. We show the quantitative evaluation in Table IV. Due to GPU memory limitations, we did not continue to train deeper networks. It can be seen that the hybrid block connection in each level can achieve the best performance. Although the BP-U block is better than refined blocks in terms of feature extraction, when each level only consists of BP-U blocks while maintaining the total number of blocks, PSRN decreases by a small margin. This result could be because transposed convolutions usually bring about a *checkerboard effect*, and the refined block not only retains the size of each feature but also benefits from residual leaning, which is helpful for feature denoising. On the other hand, upon increasing the number of blocks in each level, the results in the deeper model achieve better performance, which indicates that deeper is still better. By consideration of the tradeoff between the performance and number of network parameters, we only set up one refined block in each level.

TABLE IV
PSNR OF VARIOUS COMBINATIONS OF U AND R. T DENOTES THE TRANSPOSED CONVOLUTION LAYER. THE TESTS ARE CONDUCTED FOR SCALE FACTOR $4\times$ ON SET5 AND SET14.

Model	Parameters	Set5	Set14
$U_2TR_0 \vert U_2TR_0$	$1377K$	31.88	28.26
$U_2TU_1 \vert U_2TU_1$	$2115K$	32.12	28.46
$U_2TR_2 \vert U_2TR_2$	$2854K$	32.19	28.71
$U_3TR_0 \vert U_3TR_0$	$2214K$	32.08	28.59
$U_3TR_1 \vert U_3TR_1$	$2583K$	32.18	28.71
$U_3TR_2 \vert U_3TR_2$	$3322K$	32.23	28.78
$U_4TR_0 \vert U_4TR_0$	$2690K$	32.15	28.66
$U_4TR_1 \vert U_4TR_1$	$3059K$	32.21	28.73
$U_4TR_2 \vert U_4TR_2$	$3798K$	32.24	28.83

E. Comparisons with State-of-the-Art Methods

In this section, to verify the ability of the proposed CSRCNN, we provide extensive quantitative and qualitative comparisons with other approaches. These experiments are carried out on five popular datasets: Set5, Set14, BSD100 [64], Urban100 [10] and Manga109 [65].

More specifically, we evaluate the SR images with two commonly used image quality metrics, namely, the peak signal-to-noise ratio (PSNR) and structural similarity (SSIM), the latter of which measures image similarity in terms of three aspects: brightness, contrast and structure. All measurements used only the luminance channel (Y).

Quantitative Comparison. Table V summarizes the quantitative results on the testing sets mentioned above, with some results cited from prior literature [22], [49]. The proposed method performs favorably against existing methods, especially for 8 × enlargement. In addition, the comparisons of these methods in terms of number of parameters, GPU cost and running time are provided in Table VI. Because our CSRCNN is progressive with respect to both architecture and training, the model becomes deeper in accordance with the enlargement. EDSR [21], RRDB [56] and D-DBPN [49] include a single architecture for all factors. In addition, these three models all suffer from an enormous number of parameters, which results in the substantial capacity for complex data fitting. From Table V we can find that, for the 2× and 4× enlargements, our model outperforms the existing methods except EDSR and D-DBPN. However, the gap among the five datasets is not large. Our network shows its effectiveness for the 8× enlargement, where the CSRCNN outperforms most of the existing methods by a large margin. Referring to Table VI, our method requires less running time and GPU memory during testing than D-DBPN [49] and RRDB [56], and the effect is more prominent under 8× enlargement. Note that, when adopting the recursive learning strategy, DRCN [20] and DRRN [35] contain very few parameters, but the running time and GPU memory are relatively high compared with those of our method with a cascaded architecture. The success for higher upsampling factors suggests that our methodology is reasonable and effective.

Fig. 7. Qualitative comparison of our model with other algorithms for 4× super-resolution. From top to bottom, the two images are "woman" in Set5 and "monarch" in Set14 respectively.

Qualitative comparison. Figure 7 shows the results of 4× enlargement for visual comparison. As we can see, the proposed method produces relatively sharper edges for the patterns owing to progressive reconstruction. Generally, for 4× enlargement, all methods except Bic generated visually pleasant images with clean details, which did not bring about much visual disparity. However, with increasing enlargement, the effect is more prominent. Figure 8 shows the results of different methods for the 8× enlargement. In the top image, we can find that only our method can recover the horizontal line. Referring to the middle image, the patches generated by other SR methods contain serious artifacts and tend to be blurred, whereas our method generated a more visually pleasant image with clean details and sharp edges. For the bottom image, only our result demonstrates clear textures and is as faithful as the ground-truth image. Through the comparison, the results suggest that our methodology is more effective in estimating visual details.

To investigate the effect of images from the real world, we also conduct experiments on two images, as shown in Figure 9. In the first row, the LR image is seriously degraded by the non-uniform motion blur kernel, and the SR images from all methods are not visually friendly and display ghost shadows. The results indicate that the existing SR models are hardly capable of processing images degraded by the non-uniform motion blur kernel and that the ghost shadows will be magnified during SR. To treat this case, we can design a module to first remove the fuzzy information from LR images. In addition, another image from the Internet is given in the second row for the 8× enlargement, where the original HR image is not available. We can find that most methods except bicubic interpolation can recover sharper edges and finer details, which is due to the simple structure of the original LR image.

In addition, Figure 10 shows the PSNR performance of several recent CNN models for 8× enlargement versus the number of parameters, where the red point is our result. We can find that although D-DBPN [49] and EDSR [21] achieve fascinating performance, they suffer from a tremendous number of parameters. However, with an appropriate number of parameters, our model also achieves better performance than state-of-the-art methods for large factors.

F. Limitations

While our model obtains promising results on SR with the 8× factor, it still exhibits some limitations. On the one hand, referring to Figure 9, due to the apparent domain gap between the training dataset and test images, almost all methods trained on synthetic images barely address the situation where the LR image is seriously degraded by the non-uniform motion blur kernel. To address this issue, it is reasonable to design a special module to first remove the fuzzy information in LR images, just as in [68]. In addition, unsupervised image translation between real images and synthetic images is another way to tackle this issue [69]. On the other hand, for video super-resolution, it remains a challenge for the proposed model in terms of speed and visual perception. We will further study these problems in the future.

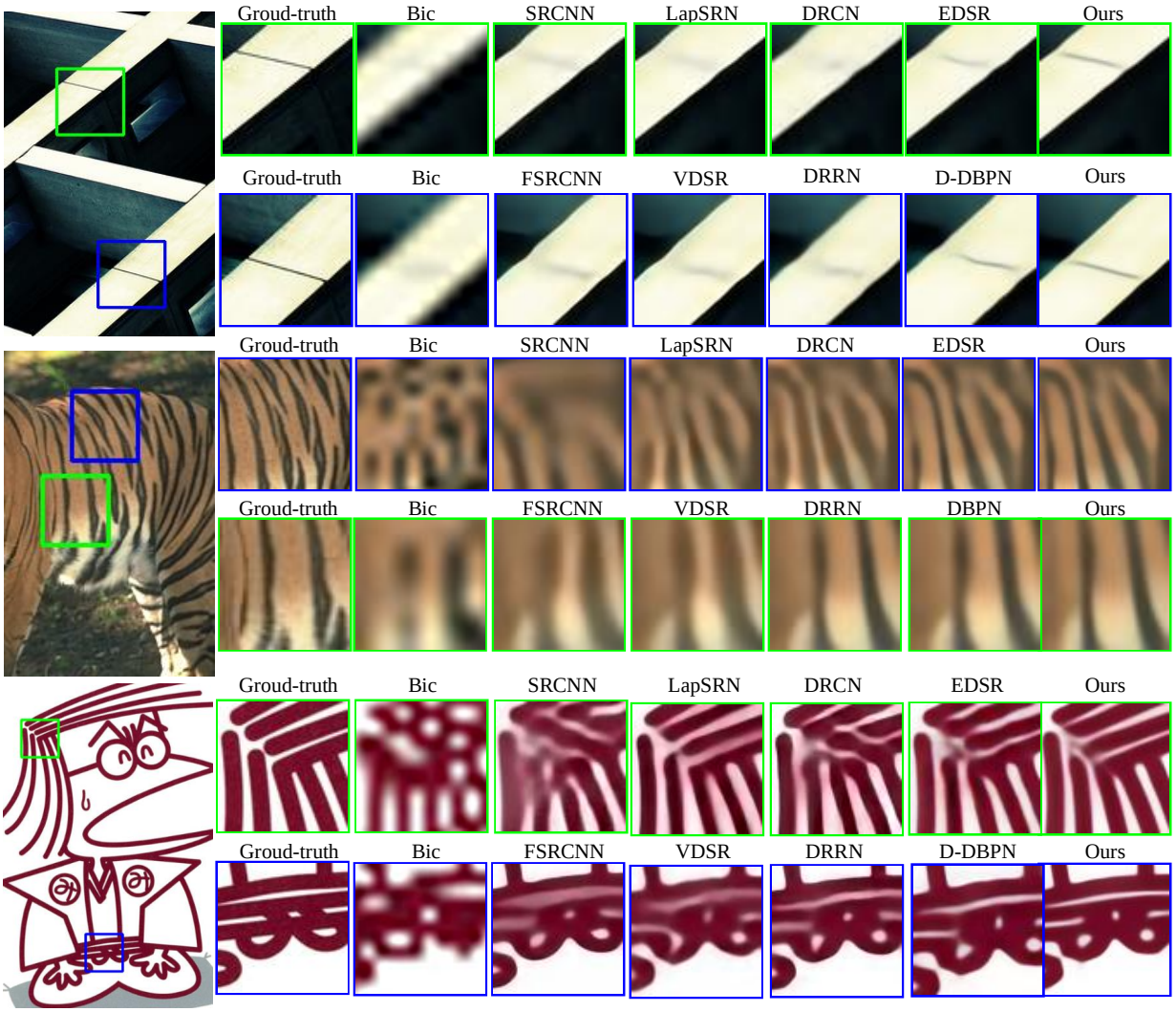


Fig. 8. Qualitative comparison of our model with other algorithms for 8× super-resolution. From top to bottom, the three images are "img043" in Urban100, "108005" in BSDS100 and "Hamlet" in Manga109.

V. Conclusion

This paper proposes a cascaded super-resolution convolutional neural network (CSRCNN) for large factor image super-resolution. Different from the conventional methods that match an LR image to target spatial resolution directly by upsampling layers at the end of network, our model is progressive both in architecture and training. We also propose a novel BP-U block by imposing the conventional U-net with backprojection. In addition, in combination with a refined block, each level adopts the concept of the hybrid block connection. These changes lead to more accurate reconstruction, particularly for large upsampling factors. To further reduce the computational complexity, we propose an asymmetric architecture that includes more blocks in the lower levels. A comprehensive evaluation of various kinds of designs and general benchmarks is presented, and the results show that the proposed CSRCNN is a concise but superior model in terms of GPU cost and running time, particularly for the 8× upsampling factor.

References

[1] C. Yang, C. Ma, and M. Yang, "Single-image super-resolution: A benchmark," in *Computer Vision - ECCV 2014 - 13th European Conference, Proceedings, Part IV*, 2014, pp. 372–386.

[2] W. Shi, J. Caballero, C. Ledig, X. Zhuang, W. Bai, K. K. Bhatia, A. M. S. M. de Marvao, T. Dawes, D. P. O'Regan, and D. Rueckert, "Cardiac image super-resolution with global correspondence using multi-atlas patchmatch," in *MICCAI 2013 - 16th International Conference, Proceedings, Part III*, 2013, pp. 9–16.

[3] L. Li, W. Wang, H. Luo, and S. Ying, "Super-resolution reconstruction of high-resolution satellite ZY-3 TLC images," *Sensors*, vol. 17, no. 5, p. 1062, 2017.

[4] W. W. W. Zou and P. C. Yuen, "Very low resolution face recognition problem," *IEEE Trans. Image Processing*, vol. 21, no. 1, pp. 327–340, 2012.

[5] Y. Romano, J. Isidoro, and P. Milanfar, "RAISR: rapid and accurate image super resolution," *IEEE Trans. Computational Imaging*, vol. 3, no. 1, pp. 110–125, 2017.

[6] S. Farsiu, M. D. Robinson, M. Elad, and P. Milanfar, "Fast and robust multiframe super resolution," *IEEE Trans. Image Processing*, vol. 13, no. 10, pp. 1327–1344, 2004.

[7] J. Xu, M. Li, J. Fan, X. Zhao, and Z. Chang, "Self-learning super-resolution using convolutional principal component analysis and random matching," *IEEE Trans. Multimedia*, vol. 21, no. 5, pp. 1108–1121, 2019.

TABLE V
QUANTITATIVE EVALUATION ON BENCHMARK DATASETS. AVERAGE PSNR AND SSIM VALUES FOR 2×, 4× AND 8× ON DATASETS SET5, SET14, BSDS100, URBAN100 AND MANGA109. RED INDICATES THE BEST PERFORMANCE AND BLUE INDICATES THE SECOND BEST PERFORMANCE ACHIEVED BY ANY METHOD.

Method	Scale	Set5	Set14	BSDS100	Urban100	Manga109
Bicubic	2×	33.69 / 0.931	30.25 / 0.870	29.57 / 0.844	26.89 / 0.841	30.84 / 0.935
SRCNN [13]	2×	36.72 / 0.955	32.51 / 0.908	31.38 / 0.889	29.53 / 0.896	35.33 / 0.967
FSRCNN [18]	2×	37.05 / 0.956	32.66 / 0.909	31.53 / 0.892	29.88 / 0.902	35.72 / 0.968
VDSR [19]	2×	37.53 / 0.958	33.03 / 0.912	31.90 / 0.896	30.77 / 0.914	37.16 / 0.974
DRCN [20]	2×	37.63 / 0.959	33.06 / 0.913	31.85 / 0.895	30.76 / 0.914	37.57 / 0.973
DRRN [35]	2×	37.74 / 0.959	33.23 / 0.914	32.05 / 0.897	31.23 / 0.919	37.92 / 0.976
LapSRN [22]	2×	37.52 / 0.959	33.08 / 0.913	31.80 / 0.895	30.41 / 0.910	37.27 / 0.974
EDSR [21]	2×	38.11 / 0.960	33.92 / 0.919	32.32 / 0.901	32.93 / 0.930	39.10 / 0.977
D-DBPN [49]	2×	38.09 / 0.960	33.85 / 0.919	32.27 / 0.900	33.02 / 0.931	39.32 / 0.978
CSRCNN	2×	37.89 / 0.960	33.81 / 0.918	32.16 / 0.897	32.79 / 0.922	39.06 / 0.977
Bicubic	4×	28.43 / 0.811	26.01 / 0.704	25.97 / 0.670	23.15 / 0.660	24.89 / 0.786
SRCNN [13]	4×	30.50 / 0.863	27.52 / 0.753	26.91 / 0.712	24.53 / 0.725	27.66 / 0.858
FSRCNN [18]	4×	30.72 / 0.866	27.70 / 0.755	26.98 / 0.715	24.62 / 0.728	27.89 / 0.859
VDSR [19]	4×	31.35 / 0.883	28.02 / 0.768	27.29 / 0.726	25.18 / 0.754	28.82 / 0.886
DRCN [20]	4×	31.54 / 0.884	28.03 / 0.768	27.24 / 0.725	25.14 / 0.752	28.97 / 0.886
DRRN [35]	4×	31.68 / 0.888	28.21 / 0.772	27.38 / 0.728	25.44 / 0.764	29.46 / 0.896
LapSRN [22]	4×	31.54 / 0.885	28.19 / 0.772	27.32 / 0.728	25.21 / 0.755	29.09 / 0.889
EDSR [21]	4×	32.46 / 0.897	28.80 / 0.788	27.71 / 0.742	26.64 / 0.803	31.02 / 0.915
D-DBPN [49]	4×	32.47 / 0.898	28.82 / 0.786	27.72 / 0.740	27.08 / 0.795	31.50 / 0.914
CSRCNN	4×	32.26 / 0.898	28.87 / 0.789	27.61 / 0.737	26.09 / 0.785	30.61 / 0.910
Bicubic	8×	24.40 / 0.658	23.10 / 0.566	23.67 / 0.548	20.74 / 0.516	21.47 / 0.649
SRCNN [13]	8×	25.33 / 0.690	23.76 / 0.591	24.13 / 0.566	21.29 / 0.544	22.37 / 0.682
FSRCNN [18]	8×	25.41 / 0.682	23.93 / 0.592	24.21 / 0.567	21.32 / 0.537	22.39 / 0.672
VDSR [19]	8×	25.72 / 0.711	24.21 / 0.609	24.37 / 0.576	21.54 / 0.560	22.83 / 0.707
DRRN [35]	8×	26.18 / 0.738	24.42 / 0.622	24.59 / 0.587	21.88 / 0.583	23.60 / 0.742
LapSRN [22]	8×	26.15 / 0.738	24.35 / 0.620	24.54 / 0.586	21.81 / 0.581	23.39 / 0.735
EDSR [21]	8×	26.97 / 0.775	24.94 / 0.640	24.80 / 0.596	22.47 / 0.620	24.58 / 0.778
D-DBPN [49]	8×	27.21 / 0.784	25.13 / 0.648	24.88 / 0.601	23.25 / 0.622	25.50 / 0.799
CSRCNN	8×	27.22 / 0.786	25.19 / 0.651	24.89 / 0.602	22.64 / 0.625	24.98 / 0.793

[8] 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, 2013.

[9] D. Ferstl, M. Rüther, and H. Bischof, "Variational depth superresolution using example-based edge representations," in *2015 IEEE International Conference on Computer Vision, ICCV 2015*, 2015, pp. 513–521.

[10] J. Huang, A. Singh, and N. Ahuja, "Single image super-resolution from transformed self-exemplars," in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015*, 2015, pp. 5197–5206.

[11] J. Yang, Z. Lin, and S. Cohen, "Fast image super-resolution based on in-place example regression," in *2013 IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 1059–1066.

[12] N. Kumar and A. Sethi, "Fast learning-based single image super-resolution," *IEEE Trans. Multimedia*, vol. 18, no. 8, pp. 1504–1515, 2016.

[13] 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, 2016.

[14] W. Shi, J. Caballero, F. Huszar, J. Totz, A. P. Aitken, R. Bishop, D. Rueckert, and Z. Wang, "Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network," in *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016*, 2016, pp. 1874–1883.

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

[16] 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, 2014.

[17] Y. Zhang, J. Liu, W. Yang, and Z. Guo, "Image super-resolution based on structure-modulated sparse representation," *IEEE Trans. Image Processing*, vol. 24, no. 9, pp. 2797–2810, 2015.

[18] C. Dong, C. C. Loy, and X. Tang, "Accelerating the super-resolution convolutional neural network," in *Computer Vision - ECCV 2016 - 14th European Conference, Proceedings, Part II*, 2016, pp. 391–407.

[19] J. Kim, J. K. Lee, and K. M. Lee, "Accurate image super-resolution using very deep convolutional networks," in *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016*, 2016, pp. 1646–1654.

[20] ——, "Deeply-recursive convolutional network for image super-resolution," in *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016*, 2016, pp. 1637–1645.

[21] B. Lim, S. Son, H. Kim, S. Nah, and K. M. Lee, "Enhanced deep residual networks for single image super-resolution," in *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops*, 2017, pp. 1132–1140.

[22] W. Lai, J. Huang, N. Ahuja, and M. Yang, "Deep laplacian pyramid networks for fast and accurate super-resolution," in *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, 2017, pp. 5835–5843.

[23] X. Yang, H. Mei, J. Zhang, K. Xu, B. Yin, Q. Zhang, and X. Wei, "DRFN: deep recurrent fusion network for single-image super-resolution with large factors," *IEEE Trans. Multimedia*, vol. 21, no. 2, pp. 328–337, 2019.

[24] R. Timofte, V. D. Smet, and L. J. V. Gool, "Anchored neighborhood regression for fast example-based super-resolution," in *IEEE International Conference on Computer Vision, ICCV 2013*, 2013, pp. 1920–1927.

[25] Z. Wang, D. Liu, J. Yang, W. Han, and T. S. Huang, "Deep networks for image super-resolution with sparse prior," in *2015 IEEE International Conference on Computer Vision, ICCV 2015*, 2015, pp. 370–378.

TABLE VI
COMPARISONS OF SR ALGORITHMS IN TERMS OF NUMBERS OF PARAMETERS, GPU COST AND RUNNING TIME FOR 2×, 4× AND 8×, SEPARATELY. DURING TESTING, THE SIZE OF THE LR INPUT IS 48×48 FOR EACH ENLARGEMENT. RED INDICATES THE NUMBER OF PARAMETERS OF OUR METHOD IN EACH ENLARGEMENT. BECAUSE OUR NETWORK IS ASSEMBLED IN A PROGRESSIVE MANNER, EACH NETWORK FOR DIFFERENT ENLARGEMENTS CONTAINS DIFFERENT NUMBERS OF PARAMETERS.

	x2			x4			x8		
Method	#Params. (K)	GPU (MB)	Time (ms)	#Params. (K)	GPU (MB)	Time (ms)	#Params. (K)	GPU (MB)	Time (ms)
SRCNN [13]	57	571	0.351	57	755	0.391	57	811	0.409
DRRN [35]	302	1439	8.023	302	2677	31.268	302	8311	124.157
DRCN [20]	1784	1523	18.705	1784	3305	73.705	1784	10737	292.636
FSRCNN [18]	12	557	1.036	12	688	1.094	12	671	1.154
VDSR [19]	667	599	2.556	667	785	3.235	667	2197	11.751
LapSRN [22]	437	663	1.662	874	851	3.136	1310	1233	4.685
RRDB [56]	16661	1917	63.678	16698	2019	75.990	16735	2209	80.442
D-DBPN [49]	6073	2689	36.814	10288	6363	39.649	22331	19949	110.794
CSRCNN	1549	1043	6.123	3059	1665	11.586	4568	3783	20.083

Fig. 9. Visual results for real-world images. For 4×, the image in the first row is from a camera with serious motion blur. For 8×, the image in the second row is from the Internet and is a native LR image with size of 108×135 and with no corresponding HR image.

[26] X. Mao, C. Shen, and Y. Yang, "Image restoration using very deep convolutional encoder-decoder networks with symmetric skip connections," in *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016*, 2016, pp. 2802–2810.
[27] H. Zhao, O. Gallo, I. Frosio, and J. Kautz, "Loss functions for image restoration with neural networks," *IEEE Trans. Computational Imaging*, vol. 3, no. 1, pp. 47–57, 2017.
[28] Y. Wang, F. Perazzi, B. McWilliams, A. Sorkine-Hornung, O. Sorkine-Hornung, and C. Schroers, "A fully progressive approach to single-image super-resolution," in *2018 IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops*, 2018, pp. 977–986.
[29] A. Odena, V. Dumoulin, and C. Olah, "Deconvolution and checkerboard artifacts," *Distill*, 2016. [Online]. Available: http://distill.pub/2016/deconv-checkerboard
[30] M. Irani and S. Peleg, "Improving resolution by image registration," *CVGIP: Graphical Model and Image Processing*, vol. 53, no. 3, pp. 231–239, 1991.
[31] K. He, X. Zhang, S. Ren, and J. Sun, "Identity mappings in deep residual networks," in *Computer Vision - ECCV 2016 - 14th European Conference, Proceedings, Part IV*, 2016, pp. 630–645.
[32] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems 25: 26th Annual Conference on Neural Information Processing Systems 2012*, 2012, pp. 1106–1114.
[33] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. S. Bernstein, A. C. Berg, and F. Li, "Imagenet large scale visual recognition challenge," *International Journal of Computer Vision*, vol. 115, no. 3, pp. 211–252, 2015.
[34] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. E. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015*, 2015, pp. 1–9.
[35] Y. Tai, J. Yang, and X. Liu, "Image super-resolution via deep recursive residual network," in *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, 2017, pp. 2790–2798.
[36] L. Gao, X. Li, J. Song, and H. T. Shen, "Hierarchical lstms with adaptive attention for visual captioning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 5, pp. 1112–1131, 2020.
[37] J. Chen, J. Shao, and C. He, "Movie fill in the blank by joint learning from video and text with adaptive temporal attention," *Pattern Recognition Letters*, vol. 132, pp. 62–68, 2020.
[38] X. Xu, F. Shen, Y. Yang, H. T. Shen, and X. Li, "Learning discriminative binary codes for large-scale cross-modal retrieval," *IEEE Trans. Image*

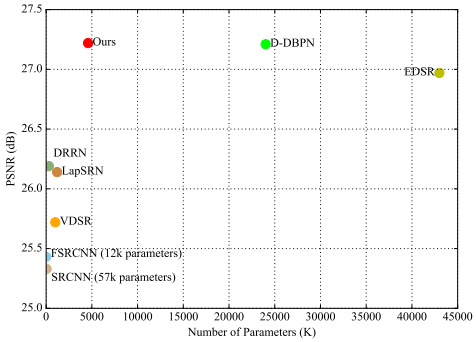


Fig. 10. PSNR and parameters of CNN-based models. The results are evaluated with the Set5 dataset for 8× enlargement.

Processing, vol. 26, no. 5, pp. 2494–2507, 2017.

[39] B. Wang, Y. Yang, X. Xu, A. Hanjalic, and H. T. Shen, "Adversarial cross-modal retrieval," in *Proceedings of the 2017 ACM on Multimedia Conference, MM 2017, Mountain View, CA, USA, October 23-27, 2017*, 2017, pp. 154–162.

[40] X. Xu, T. Wang, Y. Yang, L. Zuo, F. Shen, and H. T. Shen, "Cross-modal attention with semantic consistence for image-text matching," *IEEE Trans. Neural Netw. Learning Syst.*, 2020. [Online]. Available: https://doi.org/10.1109/TNNLS.2020.2967597

[41] D. Zhang, J. Shao, and H. T. Shen, "Kernel attention network for single image super-resolution," *ACM Trans. Multim. Comput. Commun. Appl.*, 2020. [Online]. Available: https://doi.org/10.1145/3398685

[42] Z. Liang, J. Shao, D. Zhang, and L. Gao, "Traffic sign detection and recognition based on pyramidal convolutional networks," *Neural Computing and Applications*, vol. 32, no. 11, pp. 6533–6543, 2020.

[43] A. A. Khan, J. Shao, W. Ali, and S. Tumrani, "Content-aware summarization of broadcast sports videos: An audiocvisual feature extraction approach," *Neural Processing Letters*, 2020. [Online]. Available: https://doi.org/10.1007/s11063-020-10200-3

[44] C. He, J. Shao, and J. Sun, "An anomaly-introduced learning method for abnormal event detection," *Multimedia Tools Appl.*, vol. 77, no. 22, pp. 29 573–29 588, 2018.

[45] J. Sun, X. Wang, N. Xiong, and J. Shao, "Learning sparse representation with variational auto-encoder for anomaly detection," *IEEE Access*, vol. 6, pp. 33 353–33 361, 2018.

[46] J. Sun, J. Shao, and C. He, "Abnormal event detection for video surveillance using deep one-class learning," *Multimedia Tools Appl.*, vol. 78, no. 3, pp. 3633–3647, 2019.

[47] W. Lai, J. Huang, N. Ahuja, and M. Yang, "Fast and accurate image super-resolution with deep laplacian pyramid networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 11, pp. 2599–2613, 2019.

[48] H. Kong, J. Zhao, X. Tu, J. Xing, S. Shen, and J. Feng, "Cross-resolution face recognition via prior-aided face hallucination and residual knowledge distillation," *arXiv preprint*, vol. arXiv:1905.10777, 2019.

[49] M. Haris, G. Shakhnarovich, and N. Ukita, "Deep back-projection networks for super-resolution," in *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018*, 2018, pp. 1664–1673.

[50] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *Computer Vision - ECCV 2016 - 14th European Conference, Proceedings, Part II*, 2016, pp. 694–711.

[51] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *International Conference on Learning Representations, ICLR 2015*, 2015.

[52] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. C. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014*, 2014, pp. 2672–2680.

[53] C. K. Sønderby, J. Caballero, L. Theis, W. Shi, and F. Huszár, "Amortised MAP inference for image super-resolution," in *International Conference on Learning Representations, ICLR 2017*, 2017.

[54] D. Zhang, J. Shao, G. Hu, and L. Gao, "Sharp and real image super-resolution using generative adversarial network," in *Neural Information Processing - 24th International Conference, ICONIP 2017, Proceedings, Part III*, 2017, pp. 217–226.

[55] C. Ledig, L. Theis, F. Huszar, J. Caballero, A. Cunningham, A. Acosta, A. P. Aitken, A. Tejani, J. Totz, Z. Wang, and W. Shi, "Photo-realistic single image super-resolution using a generative adversarial network," in *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, 2017, pp. 105–114.

[56] X. Wang, K. Yu, S. Wu, J. Gu, Y. Liu, C. Dong, Y. Qiao, and C. C. Loy, "ESRGAN: enhanced super-resolution generative adversarial networks," in *Computer Vision - ECCV 2018 Workshops, Proceedings, Part V*, 2018, pp. 63–79.

[57] W. Yang, X. Zhang, Y. Tian, W. Wang, J. Xue, and Q. Liao, "Deep learning for single image super-resolution: A brief review," *IEEE Trans. Multimedia*, vol. 21, no. 12, pp. 3106–3121, 2019.

[58] S. Shah, P. Ghosh, L. S. Davis, and T. Goldstein, "Stacked u-nets: A no-frills approach to natural image segmentation," *arXiv preprint*, vol. arXiv:1804.10343, 2018.

[59] P. Isola, J. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, 2017, pp. 5967–5976.

[60] Y. Bengio, J. Louradour, R. Collobert, and J. Weston, "Curriculum learning," in *Proceedings of the 26th Annual International Conference on Machine Learning, ICML 2009*, 2009, pp. 41–48.

[61] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive growing of gans for improved quality, stability, and variation," in *International Conference on Learning Representations, ICLR 2018*, 2018.

[62] Z. Hui, X. Wang, and X. Gao, "Fast and accurate single image super-resolution via information distillation network," in *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018*, 2018, pp. 723–731.

[63] E. Agustsson and R. Timofte, "NTIRE 2017 challenge on single image super-resolution: Dataset and study," in *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops*, 2017, pp. 1122–1131.

[64] D. R. Martin, C. C. Fowlkes, D. Tal, and J. Malik, "A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics," in *Proceedings of the Eighth International Conference On Computer Vision (ICCV-01) - Volume 2*, 2001, pp. 416–425.

[65] Y. Matsui, K. Ito, Y. Aramaki, A. Fujimoto, T. Ogawa, T. Yamasaki, and K. Aizawa, "Sketch-based manga retrieval using manga109 dataset," *Multimedia Tools Appl.*, vol. 76, no. 20, pp. 21 811–21 838, 2017.

[66] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Trans. Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.

[67] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *2015 IEEE International Conference on Computer Vision, ICCV 2015*, 2015, pp. 1026–1034.

[68] D. Zhang, Z. Liang, and J. Shao, "Joint image deblurring and super-resolution with attention dual supervised network," *Neurocomputing*, 2020. [Online]. Available: https://doi.org/10.1016/j.neucom.2020.05.069

[69] Z. Han, E. Dai, X. Jia, X. Ren, S. Chen, C. Xu, J. Liu, and Q. Tian, "Unsupervised image super-resolution with an indirect supervised path," *arXiv preprint*.

Dongyang Zhang is currently a Ph.D. student with the Center for Future Media, School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu, China. His research interests include image/video restoration and deep learning.

Jie Shao received the B.E. degree in computer science from Southeast University, Nanjing, China, in 2004 and the Ph.D. degree in computer science from The University of Queensland, Brisbane, Australia, in 2009. He is currently a Professor with the Center for Future Media, School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu, China. He worked as a Research Fellow at the University of Melbourne from 2008 to 2011, and at National University of Singapore from 2012 to 2014. His research interests include spatial databases and multimedia information retrieval.

Zhenwen Liang is currently an undergraduate student at the Center for Future Media, School of Computer Science and Engineering, University of Electronic Science and Technology of China. His research interests include artificial intelligence and computer vision.

Lianli Gao received the Ph.D. degree from the University of Queensland, Australia, in 2014. She is currently a Professor with the Center for Future Media, School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu, China, teaching subjects on semantic Web and machine learning theory, etc. Her research interests include data mining, machine learning, multimedia analysis, and semantic Web.

Heng Tao Shen received the B.Sc. (First Class Hons.) and Ph.D. degrees in computer science from the Department of Computer Science at National University of Singapore in 2000 and 2004 respectively. He is currently a Professor, the Dean of School of Computer Science and Engineering, the Executive Dean of Artificial Intelligence Research Institute, and the Director of Center for Future Media at University of Electronic Science and Technology of China. His current research interests include multimedia search, computer vision, artificial intelligence, and big data management. He has published over 250 peer-reviewed papers, and received 7 best paper awards from international conferences, including the Best Paper Award from ACM Multimedia 2017 and Best Paper Award-Honourable Mention from ACM SIGIR 2017. He has served as General Co-chair for ACM Multimedia 2021 and Program Committee Co-Chair for ACM Multimedia 2015, and is an Associate Editor of ACM Transactions of Data Science, IEEE Transactions on Image Processing, IEEE Transactions on Multimedia, and IEEE Transactions on Knowledge and Data Engineering.