Signal Processing: Image Communication 68 (2018) 88–100

Contents lists available at ScienceDirect

Signal Processing: Image Communication

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

Deep unsupervised learning for image super-resolution with generative adversarial network

Guimin Lin [a,b], Qingxiang Wu [a,*], Liang Chen [a], Lida Qiu [a,b], Xuan Wang [a], Tianjian Liu [b], Xiyao Chen [b]

[a] Key Laboratory of Optoelectronic Science and Technology for Medicine of Ministry of Education, College of Photonic and Electronic Engineering, Fujian Normal University, Fuzhou 350007, PR China

[b] Department of Physics and Electronic Information Engineering, Minjiang University, Fuzhou 350108, PR China

ARTICLE INFO

Keywords:
Super-resolution
Deep unsupervised learning
Sub-pixel convolution
Regularizer
Generative adversarial network

ABSTRACT

The aim of Image super-resolution (SR) is to recover high-resolution images from low-resolution ones. By virtue of the great success in numerous computer vision tasks achieved by the convolutional neural networks (CNNs), it is a nice direction to tackle the SR problem using CNNs. Despite progress in accuracy of SR using deeper CNNs, those models are almost trained base upon supervised way. In this paper, we propose a deep unsupervised learning approach for SR with a Generative Adversarial Network (GAN) framework, which is composed of a deep convolutional generator network with dense connections and a discriminator. A sub-pixel convolutional layer is operated on the top of the generator to upscale the inputs, and the standard convolutions are all implemented in the LR space, which leads to a fast restoration. The generator is trained to directly recover the high-resolution image from the low-resolution image. Strided convolution and ReLU activations are employed in the discriminator to distinguish the HR images from the produced HR images. The generator model is optimized with a combination of a data error, a regular term and an adversarial loss, which ensures local–global contents consistency and pixel faithfulness. Note that no labeled training data is employed during the training. Comparisons with several state-of-the-art supervised learn-based methods, experimental results demonstrate that the proposed model achieves a comparable result in terms of both quantitative and qualitative measurements, and it also implies the feasibility and effectiveness of the proposed unsupervised learning-based single-image super-resolution algorithm.

1. Introduction

High-resolution (HR) images often preserve more details and critical information that play a key role in numerous fields, such as medical imaging, surveillance, astronomical imaging and face recognition. Traditional approaches of obtaining HR images mainly depend on increasing the chip size or reducing the pixel size. Nevertheless, increasing the chip size will be followed by a growth in capacitance, and reducing pixel size will lead to an increase in the shot noise. In addition, the high cost of high-precision optical devices and image sensors is also an important concern in many commercial applications regarding HR imaging. Consequently, as an effective technique, which can produce visually pleasing HR images from a low-cost imaging system and limited environmental condition, image super-resolution (SR) has received a lot of attention.

The aim of SR is recovering the original HR image from one or more low-resolution (LR) images by inferring all the missing high-frequency contents, based upon reasonable assumptions or prior knowledge about the imaging process. However, SR is inherently ill-posed because a given LR image may correspond with many HR images due to the degradation factors (e.g. blurring, noising and downsampling). Various methods have been proposed to tackle the ill-posed problem. Traditional SR methods are based on interpolation, such as bicubic interpolation and Lanczos resampling [1], are frequently exploited because of their computational simplicity, but due to use of low order polynomials, this type of SR approaches is lack of fine details in the produced HR images and is prone to produce blurring details in textures and edges. Then several algorithms based on reconstruction constraint or priors are widely studied, including edge-directed priors [2], gradient profile priors [3], and non local self-similarity priors [4]. This kind of reconstruction-based SR methods is particularly effective to preserve

* Corresponding author.
E-mail address: qxwu@fjnu.edu.cn (Q. Wu).

https://doi.org/10.1016/j.image.2018.07.003
Received 22 April 2018; Received in revised form 3 July 2018; Accepted 4 July 2018

geometric structure and to suppress ringing artifacts, however, it fails to insert sufficient novel high-frequency components to the reconstructed HR outputs and is limited in reconstructing the visual complexity of the nature image, especially at high magnification.

Lately, learning-based methods have been extensively explored to model mapping from LR to HR patches. This category of schemes supposes that there exist certain relationships between the LR images and their corresponding HR counterparts and that these relationships can be learned from millions of co-occurrence LR–HR image patches, before they are employed to produce a new HR image. Since the learning-based approaches exploit the information on training images effectively, they have the ability to recover the missing high-resolution details caused by the downsampling and are superior to other SR methods. For example, Yang et al. [5] learned a compact dictionary based on sparse signal representation to generate high-resolution image. Timofte et al. proposed an improved ANR framework (A+) [6] by combining the best qualities of ANR [7] and simple functions. Tang et al. [8] achieved convincing improvement in terms of the reconstruction quality and computational cost by merging an improved structured output regression machine(SORM) and sparse coding.

Recently, motivated by the great success achieved by deep learning [9–12] in various computer vision tasks, researchers begin to exploit convolutional neural networks (CNNs) [13] with deep architecture to improve the reconstruction accuracy for image SR. Recent state-of-the-art methods mostly adopt the CNN-based models, which have provided a new inspiration and direction for the SR problem. Dong et al. [14,15] used convolutional neural networks to address the SR problem (named SRCNN), which draws considerable attention due to its simple network structure and excellent restoration quality. Wang et al. [16] employed deep learning techniques with sparse coding and achieve notable improvement over the generic SC model. Lin et al. [17] proposed a cascade of dilated convolutional neural network(CDCNN), which benefits from the end-to-end training of deep network with a specially designed skip-connections and dilation rate. SRCNN and other methods based on CNNs [16–22] have shown exciting performance.

Although the image super-resolution methods based on CNNs have obtained great success, these methods are almost classified as supervised learning. Deep unsupervised learning for SR problem is seldom addressed. In this paper, we work on exploring the feasible approach to tackle the ill-posed SR problem via deep unsupervised learning. To obtain a well-posed solution, a crucial factor is obtaining an effective regularizer or constraint on a SR reconstruction algorithm. The representative regularization models, such as the Tikhonov regularization [23], Total Variation (TV) [24,25] and bilateral TV (BTV) [26] are effective in removing image noise and outstanding for attractive edge preserving ability. Inspired by the exciting performance of Generative Adversarial Networks (GANs) [11] in unsupervised representation learning, we design a dense-connected CNN generator architecture and a discriminator network, and develop a new regularizer motivated by BTV smoothness prior [26] to well pose the SR problem via unsupervised learning. Our proposed model is competitive with some state-of-the-art methods based upon supervised learning in terms of both recovery accuracy and human perception, as presented in Fig. 1. The model SRCNN-915 [15], trained on 91 images [7] , is composed of 3 convolutional layers, and the kernel size of each layer is 9, 1 and 5 respectively. SRCNN-955 [15] also consists of 3 convolutional layers, however, it is trained on ILSVRC 2013 ImageNet and the kernel size in each layer is 9, 5 and 5 respectively. The images illustrated in the first column are recovered by classical bicubic method and the corresponding region of interesting (ROI), and the number under the images are the PSNR(dB) index. The followed three columns are obtained from SRCNN-915, SRCN-955, and our proposed model, respectively. The last column is the original images. It shows that our proposed approach generates sharp edges with rare artifacts and is most close to the ground truth in terms of subjective visual evaluation even the PSNR index is inferior to SRCNN-955 for butterfly image. In addition, the computational speed of our algorithm is faster than the two SRCNN models.

In short, the contributions of this paper include:

- We propose a novel approach to handle the ill-posed SR problem with deep unsupervised learning. Dense connections are employed in our generator model to combine the local texture information and the global abstraction information. A sub-pixel convolution is used at the top of our generator to upscale the inputs and results in a fast restoration. The generator model is optimized with non-label training data, and achieves a pleasing SR performance.
- Inspired by the BTV regularization, we develop a simple regularizer which combines of the first and second order derivatives of image with separate weight, thus it implies the smoothness of image continual section and preserves the sharp edge.
- We detail the configuration of the discriminator employed in our GAN framework, and demonstrate that deep unsupervised learning is feasible for the problem of image super-resolution, and can achieve good quality.

The rest of this paper is organized as follows. The related work is firstly reviewed in Section 2. In Section 3, the proposed method and some key components are presented. The databases used for evaluation and the experiment results are demonstrated in detail in Section 4. Finally, the main conclusions are presented in Section 5.

2. Related work

There have been numerous publications over the last few years employing learning-based policies on the SR tasks. Compared with conventional SR methods, which depend on hand-crafted features, learning-based approaches may further boost the performance, especially CNN-based methods with deep learning technique.

2.1. Image super-resolution

The goal of image super-resolution is reconstructing an HR image from the LR one. To improve the reconstruction accuracy for image SR, recent state-of-the-art methods mostly employ the supervised learning-based approaches. This type of methods exploits a set of HR images and their corresponding downsampled LR ones to learn dictionaries, regression functions or end-to-end nonlinear mapping between the two. The dictionary-based techniques attempt to create a correspondence map between the LR and HR images by space transformation. Searching in this type of dictionaries is performed via approximate nearest neighbors, as exhaustive search would be prohibitive time cost. In addition, dictionaries quickly grow in size with the amount of training data. Yang et al. [5] proposed a technique to obtain a sparse ”compact dictionary” from the training data to tackle the problem of growing dictionary sizes. He et al. [27] utilized the beta process prior to learn the over-complete dictionary pairs for adding more consistent and accurate mapping between two feature space. Ahmed and Shah [28] learned multiple coupled dictionaries, each containing features along a different direction. The high-resolution patch is reconstructed using a set of directional clustered dictionaries which gives the least sparse representation error. Due to hard capture the statistical variability of face images by only exploiting fixed l_1 norm penalty, Wang et al. [29] proposed a weighted adaptive sparse regularization (WASR) method for face hallucination reconstruction. They also introduced a neighbor embedding (NE) from the low- and high-resolution image manifolds simultaneously and proposed a coupled-layer NE (CLNE) [30] for very low resolution facial image restoration. Rasti et al. [31] just employed separate dictionaries for LR and HR patches and proposed a low complexity approach to generate HR images.

The regression-based methods typically establish regression models to reveal the relationships between the features of LR patches and a single HR patch. Dictionaries can also be leveraged together with regression based approaches to compute projection matrices to produce

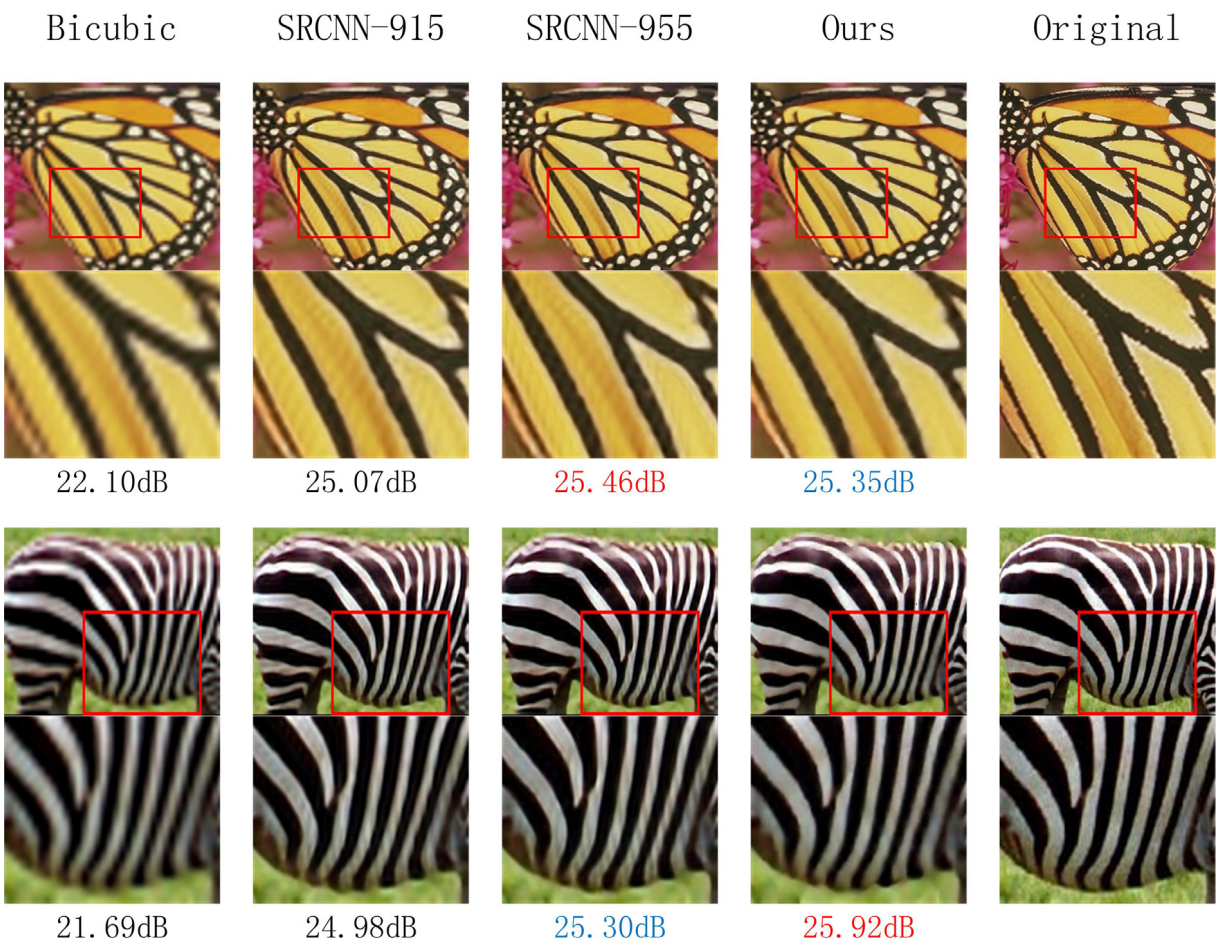

Fig. 1. Comparisons between the bicubic interpolation, the two SRCNN models [15] and our model with upscaling factor $r = 4$.

an HR result. For example, Timofte et al. [7] reduced the image SR to a projection problem from input feature space to HR feature space via the anchored neighborhood regression (ANR). Reconstruction is performed by finding the nearest neighbor of the LR patch and employing the corresponding projection matrix. Recently, they further proposed an improved ANR framework (A+) [6] by combining the best qualities of ANR and simple functions. Tang et al. [8] improved the classical structured output regression machine (SORM) and make it more suitable for training the mapping functions in the sparse coding space, then combine the improved SORM with sparse coding to achieve convincing improvement in terms of the reconstruction quality and computational cost. Other algorithms do not train dictionaries out of the training data, but choose to learn simple operators. Sun et al. [3] learned a gradient profile prior and applies it to perform high resolution image using a gradient field transformation. Tang and Shao [32] treated the SR problem as a problem of learning regression operators in a matrix space, and learn two small sized matrices that are exploited on image patches as left and right multiplication operators and allow fast recovery of the high resolution image. Choi and Kim [33] learned multiple local LR-to-HR linear mappings and a global regressor, which are applied in sequence to generate one refined HR patch from a set of HR candidates.

Recently, convolutional neural network (CNN) [13] has attracted great attentions mainly due to its success in various computer-vision fields, such as non-blind deconvolution [34,35], deblurring [36], image classification [9,37–39], and object detection [40,41]. As CNNs allow an end-to-end training for all the model components between LR input and HR output, CNN-based SR methods have shown excellent performance. Dong et al. [15] trained a deep convolutional neural network (named SRCNN) for establishing the end-to-end mapping between the LR and HR images. SRCNN is a shallow network, which has only three convolution layers. Dong et al. attempt to build deeper models, nevertheless, the deeper models perform worse than the shallow one. Finally, they deduce that deeper CNN models do not always lead to better results. As well known, with the deepening of the network, it can fit more complex mapping between LR and HR images. Therefore, deeper networks are worth trying. In our previous work [17], a 7-layer CNN model and a cascaded model were developed, which combine the skip-connections technique and dilated convolution, resulting in expanding the receptive fields efficiently and achieving noticeable improvement over SRCNN. Kim et al. [42,43] presented two highly accurate SISR methods by training a deeper convolutional network inspired by VGG-net used for ImageNet classification [39] and a deeply-recursive convolutional network. Zhang et al. [21] built a residual dense network (RDN), which makes full use of the hierarchical features from the original LR images and learns global hierarchical features from the shallow features and deep features to reconstruct HR images. Mao et al. [44] employed symmetrical skip-connections between the convolutional and deconvolutional layers to train a very deep Residual Encoder–Decoder Networks for image denoising, image super-resolution, JPEG deblocking and image inpainting. To alleviate the overly smoothing problem caused by the per-pixel loss function, Ledig et al. [20] proposed a generative adversarial network (GAN) [11] with perceptual loss function. Johnson et al. [19] proposed to use a pre-trained 16-layer VGG network [39] as perceptual loss function. Luo et al. [22] proposed an SR method for satellite image with a well-trained CNN model. However, all aforementioned CNN-based SR approaches are via supervised learning. The deep unsupervised learning for SR problems is rarely explored yet. Our goal in this paper is to investigate the suitable approach for producing SR images using unsupervised ways.

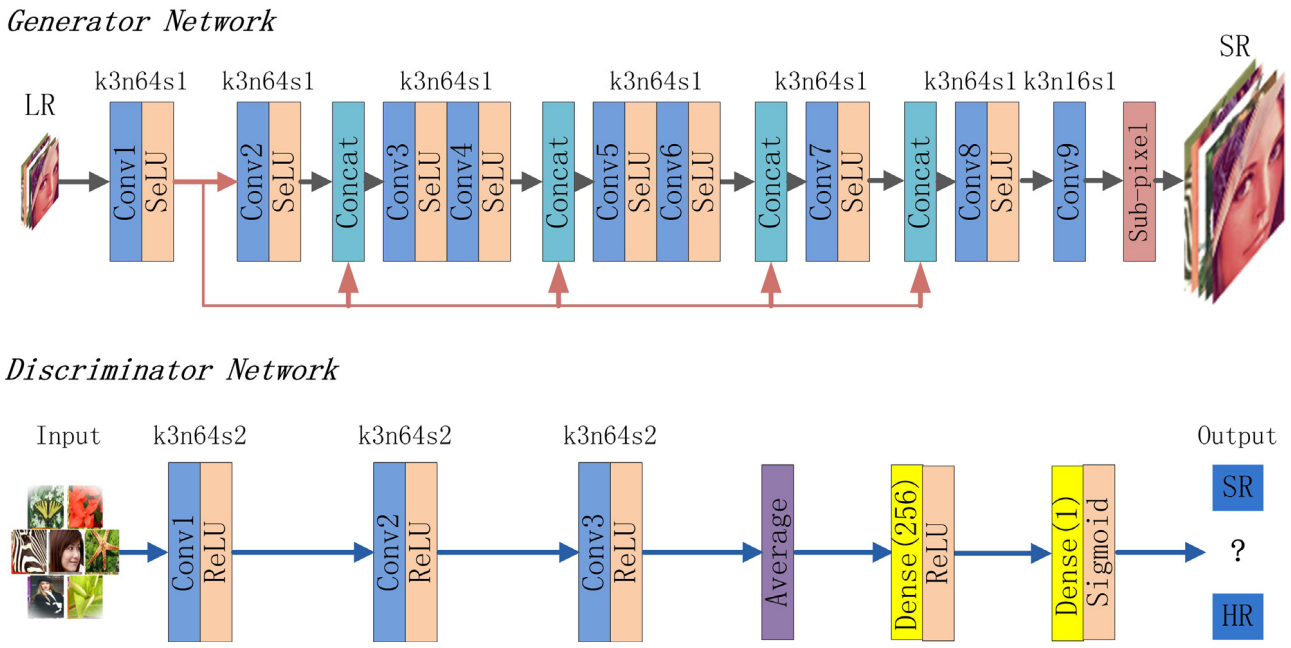

Fig. 2. The proposed model for SR reconstruction, which consists of Generator and Discriminator Network. The kernel size (k), number of feature maps (n) and stride (s) are indicated for each convolutional layer.

2.2. Generative adversarial networks (GANs)

GANs have been vigorously studied since it was proposed by Goodfellow et al. [11]. GANs provide a powerful framework for learning reusable feature representations from large unlabeled datasets and generating visually appealing natural images. However, GANs have been known to be often difficult to train, often resulting in generators that collapse to a parameter setting or produce nonsensical outputs. Many researchers have been devoted to finding ways to make progress towards stable training of GANs. For example, Radford et al. [12] introduced a Deep Convolutional GANs (DCGAN) and some guidelines for improving the stable of GANs' training. Salimans et al. [45] introduced feature matching, minibatch discrimination, etc. to encourage convergence of training GANs. Gulrajani et al. [46] constrained the gradient norm of the critic's output with respect to its input, which enables stable training of a wide variety of GAN architectures with almost no hyper-parameter tuning. Karras et al. [47] proposed a progressive training approach, which incrementally adds layers to generator and discriminator with the training advances, resulting in the stable synthesis in high resolutions and the reduced training time.

A number of attempts on employing GANs for unsupervised learning have been made recently. Radford et al. [12] introduced a deep convolutional GANs and apply GANs as a feature extractor to capture semantic image content enabling vector arithmetic for visual concepts. Salimans et al. [45] exploited GANs to generate a new class of input images, augmenting the total number of training images, further improving the accuracy in the task of image classification. Schlegl et al. [48] performed an unsupervised learning to identify anomalies in imaging data as candidates for markers. Dong et al. [49] proposed a general approach to translate images between different domains using unlabeled images without specifying any correspondence between them. Yuan et al. [50] tackled the SR problem with a Cycle-in-Cycle GAN (CinCGAN) structure trained with unpaired data. In CinCGAN, the noisy and blurry input is firstly mapped to a clean LR space, then the intermediate image is upsampled with a fine-tuned EDSR model [51] and get the HR output.

Our unsupervised objective is inspired by the principle of GANs. In the GAN framework, a generative model and a discriminative model are trained simultaneously with competing goals. The discriminator network is trained to discriminate between natural images and synthetically generated images, while the generator attempts to generate images that are indistinguishable from natural images by the best discriminator. Thus, the GAN procedure encourages the generated high-resolution samples more closer to the natural image.

3. Proposed method

The aim of single image super-resolution is to recover a high-resolution image I^{SR} from a low-resolution input image I^{LR}. Here I^{LR} is downsampled from the corresponding original HR image I^{HR} with a factor of r. In general, the images can have C color channels, thus I^{LR} can be described as a real-valued tensor of size $H \times W \times C$ and I^{HR} is $rH \times rW \times C$. The generation process of a LR image from the original HR image can be presented as

$$I^{LR} = RBI^{HR} + \varepsilon, \tag{1}$$

where B denotes the blurring operation, R is the resampling procedure with a factor r, and ε is the additive Gaussian noise. Therefore, the SR reconstruction is solving the inverse problem to estimate the underlying HR image I^{HR}. It is worth noting that the estimated I^{SR} has the same dimension as the corresponding I^{HR} and is expected to be highly similar to it. However, due to blurring, downsampling and noising, one LR image can correspond with several HR images. As a result, the SR problem is severely undetermined. To make this problem well-posed with unsupervised learning, it is necessary to incorporate effective image priors (denoted as an assembling regularization term) into the reconstruction process.

To handle the image super-resolution problem, we propose a new network architecture, as illustrated in Fig. 2, to reconstructs I^{SR} directly from the corresponding I^{LR} without other pre-processing, e.g. bicubic interpolation. The proposed architecture consists of two networks, a deep generator network G with dense skip connections and a discriminator network D. Specifically, SeLU [52] is utilized as the activation function in generator G, while ReLU [53] activation and global average pooling are employed in discriminator D and strided convolutions are used to reduce the image resolution as well. The generator G is trained as a feed-forward CNN to produce I^{SR}, while the discriminator D as an image prior to encourage generator G generating more vivid high-resolution images. In addition, an assembling prior is

developed and utilized in the training step to keep the smoothness of image in continual section and preserve the sharp edge. It is noticed that the network G is composed of concatenation operation, standard convolutional operation rather than dilated convolutional model, and sub-pixel convolution [54] in the last layer. For two common skip-connection techniques, the performance of the concatenation operation and the element-wise summation are compared in our earlier work [17], and the result suggests that the concatenation operation is more efficient for image SR task. As a consequence, the concatenation operation will be harnessed again in this work. The dilated convolution allows us to expand the receptive fields effectively, while it should introduce more zero padding to maintain the spatial size of feature maps in each convolutional layer. Further, it could bring challenges to recover sharp HR images since padding zero values is equivalent to add extra noise to the input data, especially when the padding operation occurs in LR space. To speed up the process of feed-forward, the spatial size of the input image and the feature maps in the generator G except the last layer is held on to $H \times W$. A sub-pixel convolution layer is employed in the last layer to produce an HR image with required spatial size $r \cdot H \times r \cdot W$ from LR feature maps directly. It is worth noting that our method is distinct from the CinCGAN [50], which is composed of three generators, two discriminators and a pre-trained SR model. The performance of CinCGAN mainly depends on the SR model, which is firstly trained with paired data and then fine-tuned in the Cycle-in-Cycle framework. However, our model is trained from scratch and it is not necessary to prepare LR–HR patch pairs for training.

3.1. Sub-pixel convolution

There are several ways can be used in CNN to upscale an LR image, such as deconvolution layer [55], convolution with fractional stride of $\frac{1}{r}$ in the LR space [56], and sub-pixel convolution [54]. The deconvolution layer can be seen as a multiplication of each pixel in the input feature maps by a filter element-wise with stride r, then summing over the resulting outputs. However, reduction after convolution is still expensive. The second upscaling method is commonly realized by interpolation, perforate [57] or unpooling [58] from LR space to HR space firstly and then followed by a convolution with a stride of 1 in HR space. Because the convolutional operation is involved in HR space, these implementations increase the computational cost by a factor of r^2. In contrast, sub-pixel convolution, which only involves shuffling procedure in LR space, is a more efficient alternative for upscaling an LR image. The operation is a periodic shuffling operator PS, which can be formed as follows:

$$PS(I)_{x,y,c} = I_{\lfloor x/r \rfloor, \lfloor y/r \rfloor, c \cdot r \cdot mod(y,r) + c \cdot mod(x,r)}, \tag{2}$$

where PS rearranges the elements of a tensor to a tensor of shape $r \cdot H \times r \cdot W \times C$. For instance, when a $4 \times 4 \times 2^2$ feature map is passed through the PS function, an output with shape $8 \times 8 \times 1$ can be obtained, as illustrated in Fig. 3. Comparing to reduction or convolution in HR space, the implementation of the periodic shuffling in PS function can be very fast because each operation is independent and is thus trivially parallelizable in one cycle [54].

3.2. Downsampling with Lanczos kernel

To train our proposed generator G with unsupervised learning, it has to resample the SR image estimated by the generator G to LR space and compare with the LR image. The conventional downsampling operators include pooling and convolution with a stride r. However, it would be difficult to guarantee the consistency of geometry between the SR and LR images since numerous detail information will be dropped after pooling operation. For convolution, the filter is a crucial factor which can affect the performance. A nice option is to apply the Lanczos [1] downsampling filter, which can be used as a low-pass filter and has been shown to be particularly useful for visual applications [59]. Such as, it

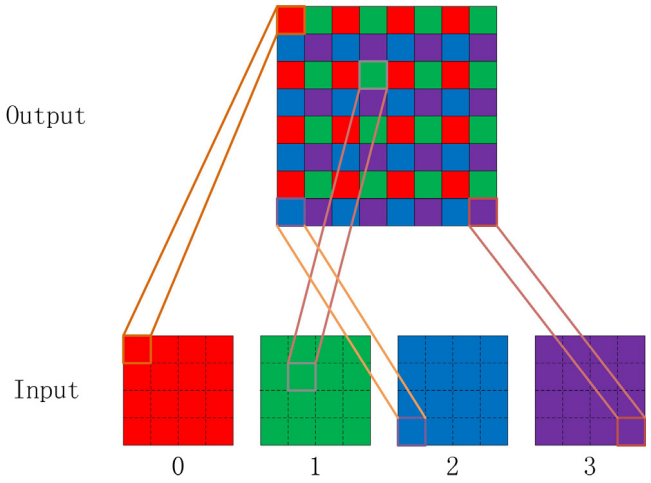

Fig. 3. The sub-pixel convolutional operation on the input feature maps with an upscaling factor of 2.

offers the best compromise in terms of reduction of aliasing, sharpness, and minimal ringing [59], comparing with its counterparts (including nearest neighbor, bilinear, and bicubic). The effect of each input sample on the decimated values is defined by the filter's reconstruction kernel $L(X)$, called the Lanczos kernel. It is the normalized $sinc$ function $sinc(x)$, windowed by the central lobe of a second, longer, $sinc$ function, or the Lanczos window, which is the central lobe of a horizontally stretched $sinc$ function $sinc(x/a)$ for $-a \le x < a$. In the case of one dimension, the Lanczos kernel can be formulated in the following way:

$$L(x) = \begin{cases} sinc(x)sinc(x/a) & \text{if } -a \le x < a, \\ 0 & \text{otherwise.} \end{cases} \tag{3}$$

Equivalently,

$$L(x) = \begin{cases} 1 & \text{if } x = 0, \\ \dfrac{a \sin(\pi x) \sin(\pi x/a)}{\pi^2 x^2} & \text{if } -a \le x < a \text{ and } x \ne 0, \\ 0 & \text{otherwise.} \end{cases} \tag{4}$$

where a is a positive integer, typically 2 or 3, which determines the size of the window. When the Lanczos kernel is used for decimation, a is commonly assigned to 2. For two dimensions, Lanczos filter's kernel is

$$L(x, y) = L(x)L(y). \tag{5}$$

3.3. Assembling regularizer

In the past decades, many regularization models have been proposed, such as the Tikhonov regularization [23], Total Variation (TV) regularization [24,25], and bilateral TV (BTV) regularization [26], among others. Among these models, the TV method is a popular and effective regularization model because of its advantages of preserving edge and detailed information in the SR process. However, the TV model favors a piecewise constant solution, which results in the flat region of the HR image are poor and some staircase artifacts are produced in the flat area. The BTV approach is more robust and can preserve more details than TV regularization method. Nonetheless, it fails to consider the partial smoothness of an image and computationally expensive. To design an effective regularization term as prior, a new regularization operator is proposed, which is highly influenced by the BTV that defines the closeness of two pixels not only based on geometric distance but also based on photometric distance [26]. The proposed regularization operator combines the first and the second order derivative of image with different decaying parameter, which implies the separate weight

for geometric distance and photometric distance. The proposed regularization operator is defined by

$$\begin{aligned}\gamma(I) &= \lambda_1 \|\nabla I\|_2 + \lambda_2 \left\|\nabla^2 I\right\|_2 \\ &\approx \lambda_1 \cdot (\left\|\Delta_x I\right\|_1 + \left\|\Delta_y I\right\|_1) + \lambda_2 \cdot (\left\|\Delta_x{}^2 I\right\|_1 + \left\|\Delta_y{}^2 I\right\|_1) \\ &= \lambda_1 \cdot \Big(\sum_{x=1}^{rH-1}\sum_{y=1}^{rW}|I_{x+1}-I_x| + \sum_{x=1}^{rH}\sum_{y=1}^{rW-1}\left|I_{y+1}-I_y\right|\Big) + \\ &\quad \lambda_2 \cdot \Big(\sum_{x=2}^{rH-1}\sum_{y=1}^{rW}|I_{x+1}-2I_x+I_{x-1}| + \sum_{x=1}^{rH}\sum_{y=2}^{rW-1}\left|I_{y+1}-2I_y+I_{y-1}\right|\Big)\end{aligned} \tag{6}$$

where I is the reconstructed SR image with shape $rH \times rW$, ∇ is the gradient operator, Δ is the forward difference operator, λ_1 and λ_2 are the decaying parameters. For reducing computational cost, the gradient operation is simplified to discrete difference calculation. The first term of the regularization is introduced as gradient homogeneity constraint term to preserve edges in the reconstructed images. The second item is used to constrain the smoothness of the reconstructed images. Thus it implies the smoothness of image continual section and is able to preserve the sharp edge.

3.4. GAN

A GAN is comprised of two adversarial networks, a generator G and a discriminator D. In our setting, the network architecture of the generator G is equivalent to an image transfer that utilizes a stack of standard convolutions and a sub-pixel convolution to magnify the LR image to the corresponding HR image. The discriminator D is a standard CNN that learns the HR image manifold through distinguishing the HR images from the generated SR images. Hence the discriminator D can be employed as a deep image prior to encourage the generator G to recover photo-realistic detail information from the LR images.

The generator G and the discriminator D are trained simultaneously through the following two-player minimax game with value function $V(D,G)$ [11]:

$$\begin{aligned}\min_G \max_D V(D,G) = &\mathrm{E}_{I^{HR}\sim p_{train}(I^{HR})}[\log D(I^{HR})] \\ &+ \mathrm{E}_{I^{LR}\sim p_G(I^{LR})}[\log(1-D(G(I^{LR})))]\end{aligned} \tag{7}$$

where E means "expectation", which is just an average. During the optimization, the minimax game aims at the Nash equilibrium of costs. The discriminator is trained to maximize the probability of identifying I^{HR} training examples as the "real" and samples from $G(I^{LR})$ as the "fake" label, while at the same time the generator G is trained to deceive D via minimizing $V(G) = \log(1-D(G(I^{LR})))$. In the course of adversarial training, the generator improves in generating realistic SR images and the discriminator advances in correctly telling the generated images apart from the HR images.

3.5. Loss function and training

The definition of the loss function $L(\cdot)$ is pivotal for the performance of our generator network with unsupervised learning and thus SR algorithm. $L(\cdot)$ is normally modeled only based on the MSE [15,17,54] in numerous SR algorithms, which is just suitable for supervised learning. In the case of unsupervised learning, we design a loss function that can evaluate the quality of a solution with respect to perceptually relevant characteristics without including the labeled HR images. Our objective function consists of a content loss, a regularization loss and an adversarial loss component that can be specified as follows:

$$L_{Gen}(\theta) = MSE(I^{LR}, R(G_\theta(I^{LR}))) + \gamma(G_\theta(I^{LR})) + \lambda_{Gen}\cdot V(G_\theta(I^{LR})) \tag{8}$$

where θ is the parameters of the generator G, $MSE(\cdot)$ is the data fidelity item, $R(\cdot)$ is the downsampling function, $\gamma(\cdot)$ is the proposed assembling regularization term, $V(\cdot)$ is the adversarial loss, and λ_{Gen} is the adversarial parameter that controls the tradeoff between the data fidelity and the prior items. The data fidelity item stands for the fidelity between the generated SR image and the original LR images. The function $R(\cdot)$ implements the resampling operation via a strided convolution (stride $= r$) with a Lanczos kernel (window size $a = 2$), where r is the scaling factor. The proposed regularization gives a prior model of the produced HR image and is responsible for preserving edge and smoothing the continual section. $V(G_\theta(I^{LR}))$ is the adversarial item and acts as a discriminative image prior, which is equivalent to $\log(1 - D(G_\theta(I^{LR})))$ and D is the discriminator network.

Stochastic gradient descent and its variants (e.g. Adam [60]) have been used to achieve state of the art performance. Optimize the generator G, so as to minimize the loss $L_{Gen}(\theta)$. In the training phase, Adam gradient descent is employed since both the neural network and the downsampling operators are differentiable. The solver parameters are assigned with the default value as recommended in [60]. The learning rate η_{Gen} is fixed to 10^{-4} and the mini-batch size is 128. The biases and the weights of each layer are initialized with a uniform distribution $U[-\frac{1}{\sqrt{n}}, \frac{1}{\sqrt{n}}]$, where n is equal to the size of the previous layer times the kernel size of the current layer. The proposed generator G is trained with a dataset of low-resolution images , where is the non-interpolated low-resolution image. The size of low-resolution patches is set to 8×8 pixels, which could provide a good tradeoff between computational efficiency and diversity of mini-batches. Although the training image size is fixed, the model G can be applied on images of arbitrary sizes during practical applications. That is because the model G is only composed of convolutional operation and non-linear activate functions.

For training the discriminator D, the cost function could be rewritten as:

$$L_{Dis}(\Theta) = \log(1 - D_\Theta(I^{HR})) + \log(D_\Theta(I^{\overline{HR}})) \tag{9}$$

where Θ is the discriminator's parameters. The Adam method is exploited again to optimize the network parameters. The solver parameters' settings are $\beta_1 = 0.5$, $\beta_2 = 0.999$, $\varepsilon = 10^{-8}$. The learning rate is set to 5×10^{-5} and the mini-batch size is 128. The training samples consist of pairs of high-resolution and estimated HR images $I^{HR}, I^{\overline{HR}}$, where I^{HR} is the high-resolution image with shape 32×32 pixels, and $I^{\overline{HR}}$ is produced from 8×8 LR image I^{LR} by the proposed generator G. It is noticed that there is not counterpart relationship between I^{HR} and $I^{\overline{HR}}$. Namely, the LR image I_i^{LR} is not decimated from the I_i^{HR}.

4. Experiments and results

In this section, the 91-images dataset [5] is used as our basic training set, which is augmented with horizontal and vertical flip, and 3 separate angles (90, 180 and 270 degrees) of rotations. The size of training LR images is 8×8, and more than 200,000 LR images can be extracted for a specific upscale factor from the augmented 91-images dataset with a stride of 11. The LR images are generated by downsampling the original images using bicubic kernel with downsampling factor $r = 4$. The HR samples are taken from the Berkeley segmentation dataset [61] BSDS500, which encompasses 500 images and is extended with the same approach as 91-images. The size of training HR images is 32×32, and more than 800,000 HR images can be obtained from the augmented BSDS500 dataset with a stride of 24. The proposed networks are trained on a NVIDIA Quadro M4000 GPU. We evaluate the performance of the proposed models on public benchmark datasets Set5 [62] and Set14 [63] which provides 5 and 14 images. All experiments are performed with a scale factor $r = 4$ between LR and HR images. Following previous works, quantitative assessment is done only on the luminance channel (in YCbCr color space) since humans are more sensitive to illuminance changes and all reported PSNR (dB) and SSIM [64] measures are done on the y-channel images with the same center-cropped method, removal of a 4-pixel wide strip from each border.

In the following, we first detail the hyper-parameters and optimization of three regular models and their performance on Set5 dataset. Next, the impacts on the restoration performance of the proposed method

are analyzed from three aspects: training samples, field of view and skip connection. To handle the arbitrary upscaling factor problem, a cascaded DCNN model is designed and inspected. At last, we compare our approach with some state-of-the-arts learning-based methods using the same database and display several visual comparison results.

4.1. Efficiency of assembling regularization terms

The image prior is an important factor for the performance of unsupervised learning-based SR algorithms. The TV prior is one of the commonly used image priors. The bilateral TV (BTV) is an improved version of TV. Both of them are helpful in removing image noise and excellent for edge preservation. However, the TV model usually introduces some staircase artifacts in the flat region of the HR image, and the BTV technique fails to consider the partial smoothness of an image. To compare the performance of different regularizer, the generator G is trained by minimized a loss function, which is only composed of data error term $MSE(\cdot)$ and regular term $\gamma(\cdot)$, as defined in the following:

$$L(\theta) = MSE(I^{LR}, R(G_\theta(I^{LR}))) + \gamma(G_\theta(I^{LR})) \tag{10}$$

$$\gamma_{TV}(I) = \lambda_{TV} \cdot \|\nabla I\|_1 \tag{11}$$

$$\gamma_{BTV}(I) = \lambda_{BTV} \cdot \underbrace{\sum_{l=-P}^{P} \sum_{m=0}^{p}}_{l+m\geq 0} \alpha^{|m|+|l|} \left\| I - S_x^l S_y^m I \right\|_1 \tag{12}$$

For Eq. (10), the regular item $\gamma(\cdot)$ can be selected from Eqs. (6), (11) and (12) , which standards for our proposed regularizer, TV regularizer and BTV regularizer, respectively. In Eq. (12), operators S_x^l, and S_y^m shift I by l , and m pixels in horizontal and vertical directions respectively [26], the scalar weight is fixed to 0.7 in this setting. During the training phase, the parameter $\lambda_{TV} = 5\times10^{-8}$, $\lambda_{BTV} = 2\times10^{-8}$, $P = 2$, $\lambda_1 = 5\times10^{-8}$ and $\lambda_2 = 2 \times 10^{-8}$. The rescaling function $R(\cdot)$ is implemented by a strided convolution operator with a 16×16 Lanczos kernel. These models are optimized on the same training samples using a Quadro M4000 GPU with Adam [60] algorithm for 800 epochs. The performance of TV, BTV and our prior are evaluated on Set5, as shown in Fig. 4. The bicubic interpolation is used as the baseline. From the result of comparison in PSNR index, it implies that our proposed regularizer is simple but effective for improve the performance of generator G. A visual comparison is presented in Fig. 4(b). The original HR image and its region of interest (ROI) are laid on the top, while the ROIs of four reconstructed HR images with PSNR are displayed on the bottom. It is shown that the ROI super-resolved by our method is relatively more perceptually satisfying than TV and BTV, meanwhile our regularizer spends less computational cost than BTV. The staircase artifacts emerge obviously in the HR images obtained by TV regular, especial in the branch section. The BTV regular appears overly smooth and fails to recover the texture of the branch. Our designed regular can achieve a trade-off in terms of artifact and smoothness and result in better performance. So in the following experiments, the regular term will be fixed to Eq. (6).

4.2. Downsampling

The goal of our proposed generator G is to reconstruct a corresponding HR image from a LR image. However, it is a challenge to train the generator G without counterpart ground truth HR images to produce visually pleasing HR images. A reasonable approach is to decimate the recovered HR image I^{SR} to the LR domain space and evaluate the error between the original LR image I^{LR} and the downsampled LR image, just as the aforementioned data error term in Eq. (10). So an effective rescaling method is crucial for achieving the optimization of generator G. A popular technique for decimation is pooling operator, such as maximum and average pooling, which are commonly exploited in CNNs to reduce the spatial size of feature maps. To better understand the maximum and average pooling operator, a flow chart of pooling with kernel = 2 and stride = 2 is illustrated in Fig. 5(a). In an extreme case, a 2×2 LR matrix can be obtained by maximum pooling from a 4×4 HR matrix, where it only contains zeros and the elements appeared in the LR matrix. However, the texture details in two matrices would be decidedly different in perception. So it is impossible to attain pleasing HR images from generator G trained via maximum pooling. And this deduction is confirmed by our testing result, as the red curve marked in Fig. 5(b). During the training with maximum pooling, the performance of generator G is boosted in the first few epochs and then dramatically deteriorated. The average pooling sounds reasonable because there exists similar texture information between the corresponding HR matrix and the LR matrix, demonstrated in the bottom of Fig. 5(a). In the training phase, the performance of generator G can be steadily improved and get 29.35 dB on the Set5 dataset after 800 epochs.

Strided convolution is an alternative approach for downsampling, but the kernel is an important component to define. A learnable kernel is explored in our trials, however, it cannot satisfy the visual requirement and is disregarded. Lanczos filter is a low-pass filter and can provide compromise in terms of reduction of aliasing, sharpness, and minimal ringing [59]. A 16×16 Lanczos kernel is constructed and normalized by picking up values in x and y directions according to Eq. (5). A strided convolution with the Lanczos kernel is employed to decimate the produce HR image to LR space and the test convergence curve with blue color is plotted in Fig. 5(b). It indicates that downsampling via strided convolution with Lanczos kernel is a better choice for our networks.

4.3. Efficiency of activation functions

Skip-connections has successfully used for several vision tasks [38,56,65], which is not only a means to combine local information across different layers to refine the performance of CNNs, but also allows for faster convergence during training. A similar idea is employed in our proposed generator network. The feature maps of the 1st convolution layer are concatenated with the feature maps from 2nd layer, 4th layer, 6th layer and 7th layer via skip-connections technique in the feed-forward generator network, as presented in Fig. 2. The filter size in each convolutional layer is 3×3. The number of filters is 64 except the 9th layer because a sub-pixel convolution is connected after it. According to the principle of sub-pixel convolution, the channels of feature maps before sub-pixel operator must equal to $C\cdot r^2$, where C standards for the channels of sub-pixel output and r is the upscaling factor. Here we only consider the high magnification $r = 4$ and $C = 1$ since only a luminance channel is concerned.

After convolutional operations, a non-linear activation function is applied in place to each convolutional layer excluding the 9th layer. ReLU is the most frequently utilized activation function and commonly achieves a nice performance in numerous vision fields. However, a problem with the ReLU function is that the unit is almost closed and cannot update its input parameters when an input over the zero boundary. PReLU is a kind of modified activation functions, which improves model fitting with nearly zero extra computational cost and little overfitting risk especially when the coefficient α is shared by all channels of one layer. SeLU activation can automatically converge towards zero mean and unit variance without the help of batch normalization and well handle the vanishing and exploding gradients problem. These three activation functions are also evaluated during our trials. The Set5 dataset is employed as the validation set during the training stage, and performance (e.g., PSNR) is evaluated only on the luminance channel. The test convergence curves for 3 different activation functions with one upscaling factor $r = 4$ are depicted in Fig. 6. The curve shown in blue color denotes the SeLU, red color is the PReLU, and black color is the ReLU. From the comparisons among three activation functions, it expresses that SeLU is more efficient than the other two in our proposed architectures and there is no obvious difference between PReLU with channel-shared coefficient and ReLU.

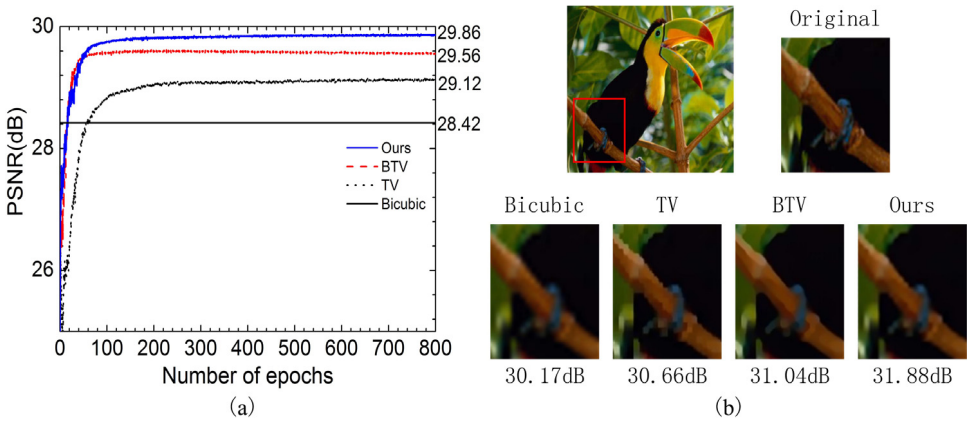

Fig. 4. Comparison among generators trained with different regularization for an upscaling factor $r = 4$. (a) The average PSNR curve testing on the Set5. (b) Super-resolution results of "bird" from Set5.

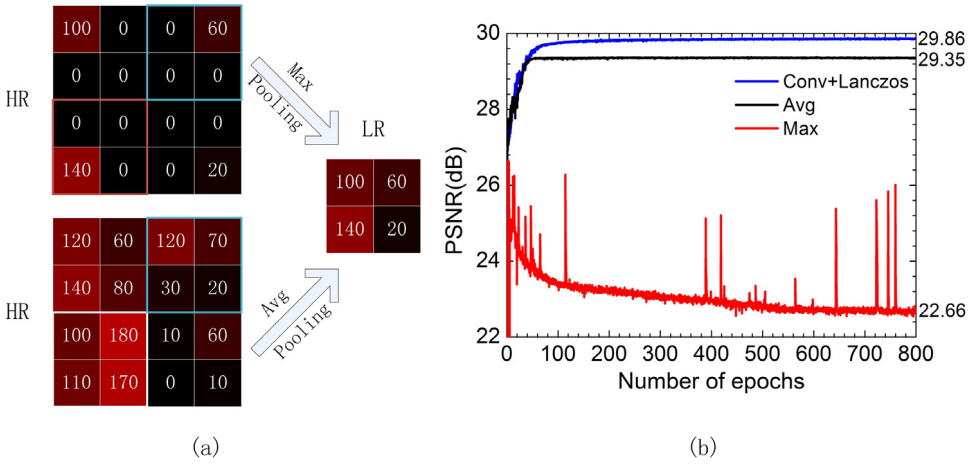

Fig. 5. (a)Illustration of Max pooling and Average pooling operate on matrices.(b)Test convergence curves for generators trained with three different downsampling methods on the Set5 dataset. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

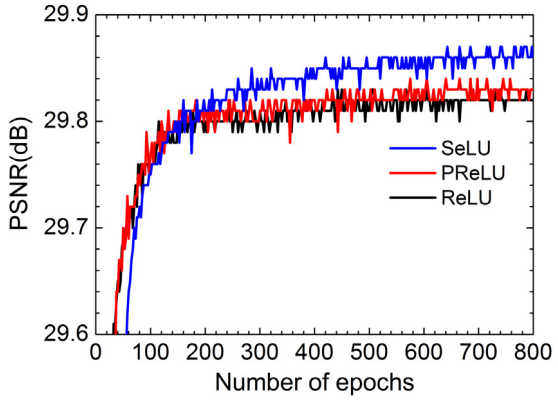

Fig. 6. The test convergence curves of the proposed generator trained with 3 activation functions. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

4.4. Benefits of GAN

Just as the guidelines summarized by Radford et al. [12], most prior GAN implementations [12,20,45] employ batch normalization in both the generator and the discriminator to help stabilize training. At the same time, batch normalization is able to significantly improve the classification accuracy of discriminator network [66]. However, ReLU activations and batch normalization are not adopted in our proposed generator network since SeLU activations have provided the power of normalization and achieve better results in this setting.

It is well known that generation of high-resolution images is difficult due to its ill-posed characteristic. So, for a discriminator, it is easier to tell the generated images apart from the ground truths for higher resolution. The goal of GAN is to help the generator to produce refined HR images via a discriminator. Nevertheless, it will rashly label the estimated SR image as SR image when the discriminator has skilled at distinguishing HR images from SR images. In such case, it could not obtain any help to improve the generator from the discriminator because the gradient from the discriminator has vanished and the parameters of the discriminator will not update. To resolve this, batch normalization is simply omitted and LeakyReLU is replaced with ReLU in our discriminator model, which will weaken its discriminatory power and make it suitable for an image prior learning from the HR images and SR images.

The contribution of two discriminator models is compared in the test. The model using LeakyReLU activation followed by batch-normalization layers is simply denoted as "BN" in Fig. 7, the other one just employing ReLU activation is referred as "No BN". The loss curve of "BN" model is plotted in red color, as shown in Fig. 7(a). The curve drastically declines to near zero after 100 epochs, which implies the gradient vanishing and results in a non-GAN training. Hence, the "BN" model cannot effectively boost the power of estimating high-frequency details for the generator. The "No BN" model fails to achieve a balance between HR images and

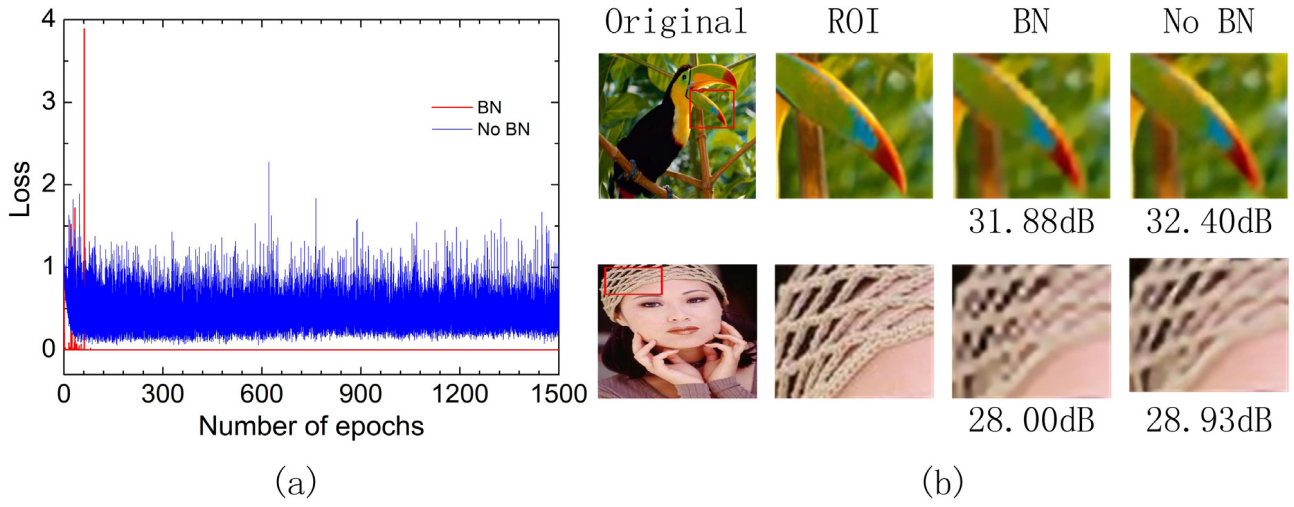

Fig. 7. (a) Loss curves of two discriminator models. (b)Comparison between two generators trained with "BN" and "No BN" models for an upscaling factor $r = 4$. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 1
Quality Assessment Results of PSNR(dB) and SSIM on Set5 dataset. In the "Average" row, the best scores are highlighted in bold red and the second best in blue.

Images	Methods				
	Bicubic	A+	SRCNN-915	SRCNN-955	Ours
Baby	31.78/ 0.8567	33.28/0.8842	32.98/ 0.8779	33.13/ 0.8824	33.03/ 0.8752
Bird	30.18/ 0.8729	32.54/0.9131	31.98/ 0.9018	32.50/ 0.9112	32.39/ 0.9078
Butterfly	22.10/ 0.7369	24.42/0.8390	25.07/ 0.8416	25.46/ 0.8566	25.35/ 0.8591
Head	31.59/ 0.7536	32.52/0.7828	32.19/ 0.7726	32.44/ 0.7801	32.41/ 0.7735
Woman	26.46/ 0.8318	28.65/0.8820	28.21/ 0.8710	28.90/ 0.8837	28.93/ 0.8814
Average	**28.42/ 0.8104**	**30.28/0.8603**	**30.09/ 0.8530**	**30.49/ 0.8628**	**30.42/ 0.8594**

SR images, as the blue curve displayed in Fig. 7(a), but it seems to successfully learn the HR image manifold and could act as an image prior to refine the produced HR images using the objective function described in Eq. (8). The hyper-parameter λ_{Gen} is set to 3×10^{-8} during the training phase. For the proposed GAN architecture, the generator and the discriminator are trained for 1500 epochs because of the difficulty in training GAN. In Fig. 7(b), two images from the Set5 dataset are recovered with our proposed generator network. The 1st column is the original images, the ROIs cut from the original images are listed in the 2nd column. The "BN" column is the ROIs extracted from the SR images produced by the generator trained with "BN" model. And the last column is generated by the generator trained with "No BN" model. It is obvious that the SR images produced by the generator trained with "No BN" model are more pleasing than the one trained with "BN" model in terms of both quantitative and qualitative measurements.

4.5. Comparison with other methods

We compare our algorithm with three state-of-the-art algorithms, including: Adjusted anchored neighborhood regression (A+) [6], SRCNN's standard model 9-1-5 and its improved model 9-5-5 [15]. Considering that human eyes are more sensitive to the luminance component than to the chrominance components, the comparison is performed only in the luminance channel. Tables 1 and 2 show the PSNR(dB) and structural similarity (SSIM) [64] for bicubic, A+ , two SRCNN models and our approach evaluated on each image from the Set5 and Set14 dataset. The average values of the two indices are calculated in the last row for each table. Note that our model used here is optimized via the GAN framework and the activation functions exploited after convolution operators in our discriminator are ReLU not followed by batch normalization. From the quality assessment results, it can be seen that our unsupervised learning approach performs comparably with previous supervised learning methods in PSNR and SSIM.

To further evaluate the SR performance of the proposed generator, four images from Set5 and Set14 are performed and their corresponding results are displayed in Figs. 8–11. The results of previous state-of-the-art methods are processed by the codes from authors' websites. For a fair comparison, all test images are downscaled using the same bicubic kernel.

The first column of these figures is the original images with a red box, in which is the region of interesting (ROI) patch, and the ROI sub-image is placed below the corresponding HR image. The SR results reconstructed by compared methods and their ROIs are depicted in orders from left to right. It can be seen that the proposed generator recover more pleasing details without obvious artifacts across the images, but other methods are prone to produce some artifacts along the edges.

From the visual results, it can be seen that our method generates sharp edges with rare artifacts and is more close to the ground truth in subjective visual evaluation even not best in the PSNR index. As depicted in Fig. 8, the fingers in the result of "Ours" have sharp edges and less blurred artifacts. In Fig. 9, although there contain plentiful high-frequency details in the whiskers of baboon, a comparable result can also be produced by our model. In Fig. 10, the words in the ROI produced with our model is more distinct. In the enlarged ROI patch, the results of other methods are less realistic while our result shows clearly contour of the gaps between tiles, as shown in Fig. 11. According to the comparison, we can see that though our proposed method obtains lower PSNR and SSIM values than others, the reconstructed images have sharp edges and fewer artifacts than others.

4.6. Running time

The running time comparisons of several state-of-the-art methods along with their restoration performance on Set14 is illustrated in

Table 2
Quality Assessment Results of PSNR(dB) and SSIM on Set14 dataset. In the "Average" row, the best scores are highlighted in bold red and the second best in blue.

Images	Methods				
	Bicubic	A+	SRCNN-915	SRCNN-955	Ours
Baboon	22.44/0.4522	22.74/0.5033	22.70/0.4956	22.73/0.5029	22.66/0.4860
Barbara	25.15/0.6863	25.74/0.7289	25.70/0.7225	25.76/0.7293	25.68/0.7205
Bridge	23.15/0.5403	23.77/0.5971	23.66/0.5892	23.76/0.5990	23.71/0.5845
Coastguard	25.48/0.5222	25.98/0.5581	25.93/0.5470	26.04/0.5563	26.02/0.5471
Comic	21.69/0.5831	22.59/0.6596	22.53/0.6513	22.70/0.6658	22.55/0.6500
Face	31.55/0.7519	32.44/0.7805	32.12/0.7698	32.38/0.7779	32.30/0.7712
Flowers	25.52/0.7215	26.90/0.7746	26.84/0.7660	27.14/0.7791	27.00/0.7683
Foreman	29.41/0.8663	32.24/0.9091	31.47/0.8971	32.14/0.9080	32.05/0.9025
Lena	29.84/0.8139	31.41/0.8454	31.20/0.8394	31.40/0.8436	31.23/0.8366
Man	25.70/0.6756	26.78/0.7276	26.65/0.7184	26.89/0.7300	26.79/0.7189
Monarch	27.46/0.8808	29.39/0.9122	29.89/0.9126	30.22/0.9181	30.05/0.9114
Pepper	30.60/0.8363	32.87/0.8655	32.34/0.8565	32.98/0.8648	32.92/0.8603
Ppt3	21.98/0.8126	23.64/0.8764	23.84/0.8670	24.80/0.8928	24.31/0.8859
Zebra	24.08/0.6831	25.94/0.7495	25.97/0.7455	26.09/0.7505	26.16/0.7339
Average	**26.00/0.7019**	**27.32/0.7491**	**27.20/0.7413**	**27.50/0.7513**	**27.39/ 0.7412**

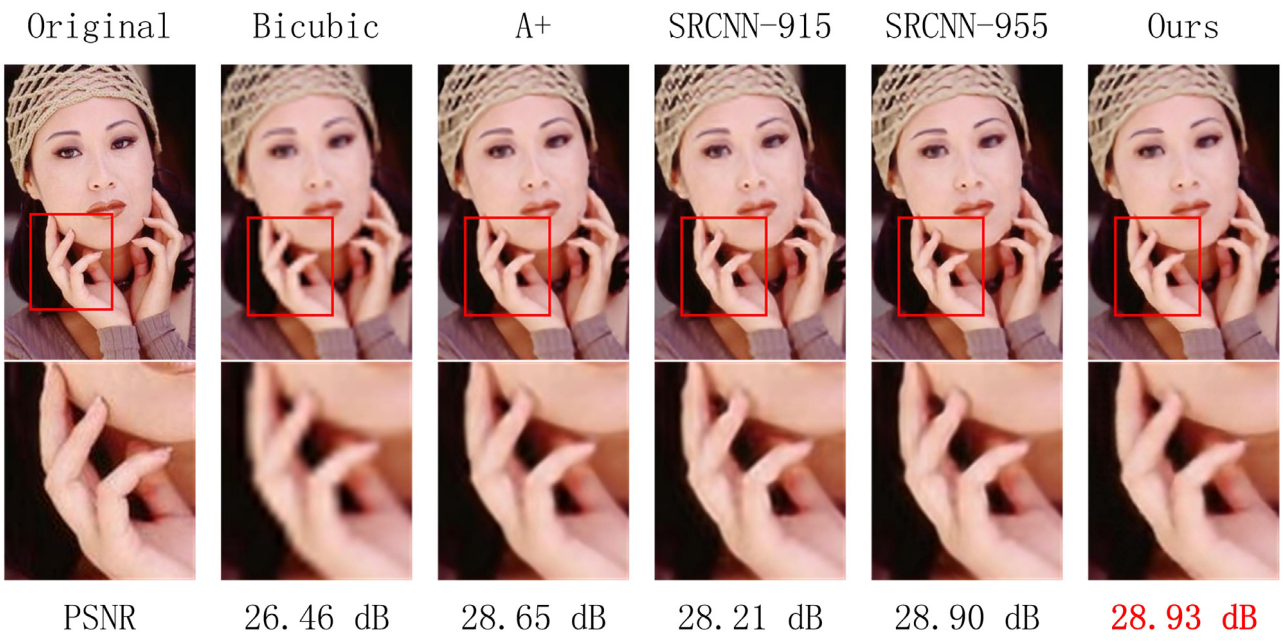

Fig. 8. Super-resolution results of "woman" from Set5. The edge of fingers in ground truth is clearly recovered in our model, whereas it is restored with some blurred artifacts in the results of other methods.

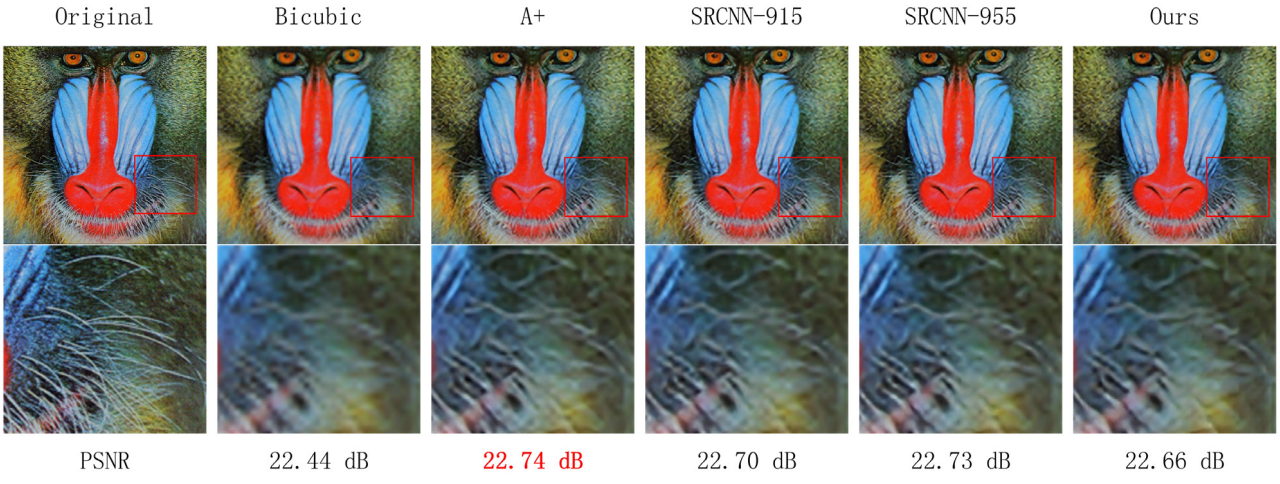

Fig. 9. Super-resolution results of "baboon" from Set14. There is a substantial high-frequency detail in the ROI, a pleasing result can also be recovered via our model.

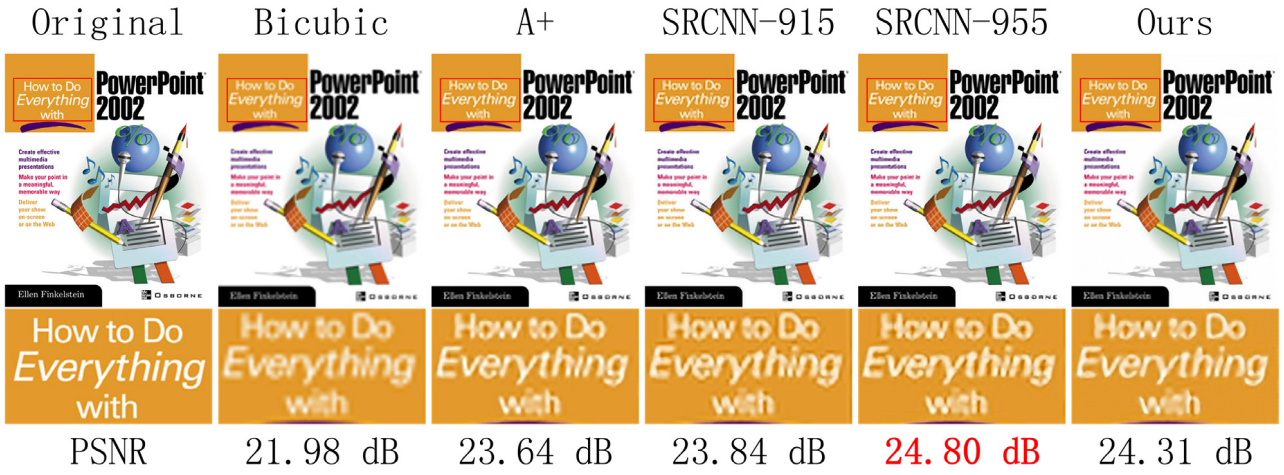

Fig. 10. Super-resolution results of "ppt3" from Set14. The words in the ROI produced with our method are sharper than others even its PSNR index is lower than SRCNN-955.

Fig. 11. Super-resolution results of "foreman" from Set14. There are less staircase artifacts in the ROI produced with our generator.

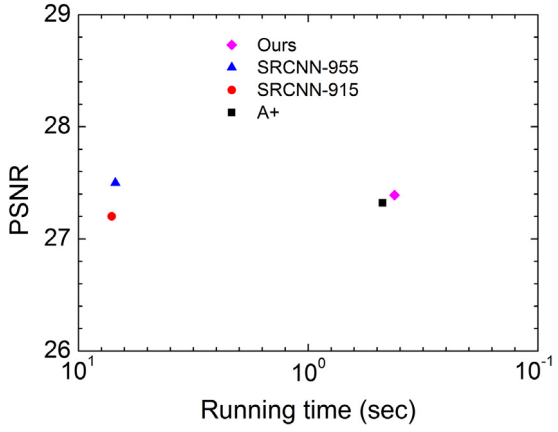

Fig. 12. The proposed approach maintains high and competitive speed in comparison with some supervised learn-based methods. The chart is based on Set14 results summarized in Table 2.

Fig. 12. The running time is obtained by averaging from the implementation on Set14 ten times for each method. All baseline methods are implemented from the publicly available code provided by the authors. All the experiments are implemented on an Intel i7-6700 3.4 GHz CPU. It clearly shows that the two SRCNN models are computational cost since all their convolutions work in the HR space. There are 9 convolutional layers in our generator, while it still achieves faster than the A + because they all operate in the LR domain and lead to fast restoration.

5. Conclusion

In summary, we propose a novel image super-resolution method using deep unsupervised learning, which is comparable with recent supervised state-of-the-arts methods. These goals are achieved mainly due to our designed assembling regularizer, the effective GAN framework and the combination of low-level local information and high-level abstraction information through skip-connections. Quantitative and qualitative assessments on the benchmark test images suggest the effectiveness of the proposed approach. However, the problem of unstable training for GAN still presents in our case. In the future, one can design a better regularizer or image prior to boost the SR performance, and using the GAN as an effective image prior is still worth exploring.

Acknowledgments

The authors would like to thank the questions and suggestions of the anonymous reviewers that helped to improve this document. This work was supported in part by the Program for Changjiang Scholars and Innovative Research Team in University (Grant No. IRT_15R10), Special Funds of the Central Government Guiding Local Science and Technology Development (2017L3009), and the Natural Science Foundation of Fujian Province of China under Grant 2017J01560.

References

[1] C.E. Duchon, Lanczos filtering in one and two dimensions, J. Appl. Meteorol. 18 (1979) 1016–1022.

[2] Y.-W. Tai, S. Liu, M.S. Brown, S. Lin, Super resolution using edge prior and single image detail synthesis, in: Computer Vision and Pattern Recognition, CVPR, IEEE, 2010, pp. 2400–2407.

[3] J. Sun, Z. Xu, H.-Y. Shum, Gradient profile prior and its applications in image super-resolution and enhancement, IEEE Trans. Image Process. 20 (2011) 1529–1542.

[4] M. Protter, M. Elad, H. Takeda, P. Milanfar, Generalizing the nonlocal-means to super-resolution reconstruction, IEEE Trans. Image Process. 18 (2009) 36–51.

[5] J. Yang, J. Wright, T.S. Huang, Y. Ma, Image super-resolution via sparse representation, IEEE Trans. Image Process. 19 (2010) 2861–2873.

[6] R. Timofte, V. De Smet, L. Van Gool, A+: Adjusted anchored neighborhood regression for fast super-resolution, in: Asian Conference on Computer Vision, Springer, 2014, pp. 111–126.

[7] R. Timofte, V. De, L. Van Gool, Anchored neighborhood regression for fast example-based super-resolution, in: IEEE International Conference on Computer Vision, ICCV, IEEE, 2013, pp. 1920–1927.

[8] Y. Tang, W. Gong, Q. Yi, W. Li, Combining sparse coding with structured output regression machine for single image super-resolution, Inform. Sci. 430 (2018) 577–598.

[9] A. Krizhevsky, I. Sutskever, G.E. Hinton, ImageNet classification with deep convolutional neural networks, in: International Conference on Neural Information Processing Systems, NIPS, 2012, pp. 1097–1105.

[10] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, Y. Lecun, OverFeat: Integrated recognition, localization and detection using convolutional networks, in: International Conference on Learning Representations, ICLR, 2014.

[11] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial nets, in: Advances in Neural Information Processing Systems, NIPS, 2014, pp. 2672–2680.

[12] A. Radford, L. Metz, S. Chintala, Unsupervised representation learning with deep convolutional generative adversarial networks, in: International Conference on Learning Representations, ICLR, 2016.

[13] Y. LeCun, B. Boser, J.S. Denker, D. Henderson, R.E. Howard, W. Hubbard, L.D. Jackel, Backpropagation applied to handwritten zip code recognition, Neural Comput. 1 (1989) 541–551.

[14] C. Dong, C.C. Loy, K. He, X. Tang, Learning a deep convolutional network for image super-resolution, in: European Conference on Computer Vision, Springer, 2014, pp. 184–199.

[15] C. Dong, C.C. Loy, K. He, X. Tang, Image super-resolution using deep convolutional networks, IEEE Trans. Pattern Anal. Mach. Intell. 38 (2016) 295–307.

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

[17] G. Lin, Q. Wu, L. Qiu, X. Huang, Image super-resolution using a dilated convolutional neural network, Neurocomputing 275 (2018) 1219–1230.

[18] J. Kim, J.K. Lee, K.M. Lee, Deeply-recursive convolutional network forimage super-resolution, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR, IEEE, 2015, pp. 1637–1645.

[19] J. Johnson, A. Alahi, F.F. Li, Perceptual losses for real-time style transfer and super-resolution, in: European Conference on Computer Vision, 2016, pp. 694–711.

[20] C. Ledig, L. Theis, F. Huszar, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, Photo-realistic single image super-resolution using a generative adversarial network, in: Computer Vision and Pattern Recognition, 2017. CVPR 2017, IEEE, 2017, pp. 4681–4690.

[21] Y. Zhang, Y. Tian, Y. Kong, B. Zhong, Y. Fu, Residual dense network for image super-resolution, in: 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR, IEEE, 2018.

[22] Y. Luo, L. Zhou, S. Wang, Z. Wang, Video satellite imagery super resolution via convolutional neural networks, IEEE Geosci. Remote Sens. Lett. 14 (2017) 2398–2402.

[23] A.N. Tikhonov, Solution of incorrectly formulated problems and the regularization method, Sov. Math. Dokl. 4 (1963) 1035–1038.

[24] L.I. Rudin, S. Osher, E. Fatemi, Nonlinear total variation based noise removal algorithms, Physica D 60 (1992) 259–268.

[25] S.D. Babacan, R. Molina, A.K. Katsaggelos, Variational Bayesian blind deconvolution using a total variation prior, IEEE Trans. Image Process. 18 (2009) 12–26.

[26] S. Farsiu, M.D. Robinson, M. Elad, P. Milanfar, Fast and robust multiframe super resolution, IEEE Trans. Image Process. 13 (2004) 1327–1344.

[27] L. He, H. Qi, R. Zaretzki, Beta process joint dictionary learning for coupled feature spaces with application to single image super-resolution, in: Computer Vision and Pattern Recognition, 2013, pp. 345–352.

[28] J. Ahmed, M.A. Shah, Single image super-resolution by directionally structured coupled dictionary learning, EURASIP J. Image Video Process. 2016 (2016) 36.

[29] Z. Wang, R. Hu, S. Wang, J. Jiang, Face hallucination via weighted adaptive sparse regularization, IEEE Trans. Circuits Syst. Video Technol. 24 (2014) 802–813.

[30] J. Jiang, R. Hu, Z. Wang, Z. Han, J. Ma, Facial image hallucination through coupled-layer neighbor embedding, IEEE Trans. Circuits Syst. Video Technol. 26 (2016) 1674–1684.

[31] P. Rasti, K. Nasrollahi, O. Orlova, G. Tamberg, C. Ozcinar, T.B. Moeslund, G. Anbarjafari, A new low-complexity patch-based image super-resolution, IET Comput. Vis. 11 (2017) 567–576.

[32] Y. Tang, L. Shao, Pairwise operator learning for patch based single-image super-resolution, IEEE Trans. Image Process. 26 (2017) 994–1003.

[33] J.-S. Choi, M. Kim, Single image super-resolution using global regression based on multiple local linear mappings, IEEE Trans. Image Process. 26 (2017) 1300–1314.

[34] C.J. Schuler, H.C. Burger, S. Harmeling, B. Scholkopf, A machine learning approach for non-blind image deconvolution, in: IEEE Conference on Computer Vision and Pattern Recognition, 2013, pp. 1067–1074.

[35] L. Xu, J.S. Ren, C. Liu, J. Jia, Deep convolutional neural network for image deconvolution, in: Advances in Neural Information Processing Systems, 2014, pp. 1790–1798.

[36] M. Hradiš, J. Kotera, P. Zemčík, F. Šroubek, Convolutional neural networks for direct text deblurring, in: Proceedings of the British Machine Vision Conference, BMVC, BMVA Press, 2015, pp. 6.1–6.13.

[37] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, Going deeper with convolutions, in: IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2014, pp. 1–9.

[38] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Computer Vision and Pattern Recognition, CVPR, 2016, pp. 770–778.

[39] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, in: International Conference on Learning Representations, 2015.

[40] W. Ouyang, X. Wang, X. Zeng, S. Qiu, DeepID-Net: Deformable deep convolutional neural networks for object detection, in: Computer Vision and Pattern Recognition, 2015, pp. 2403–2412.

[41] N. Zhang, J. Donahue, R. Girshick, T. Darrell, Part-based R-CNNs for fine-grained category detection, in: European Conference on Computer Vision, Springer, 2014, pp. 834–849.

[42] J. Kim, J. Kwon Lee, K. Mu Lee, Accurate image super-resolution using very deep convolutional networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2016, pp. 1646–1654.

[43] J. Kim, J. Kwon Lee, K. Mu Lee, Deeply-recursive convolutional network for image super-resolution, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 1637–1645.

[44] X. Mao, C. Shen, Y. Yang, Image restoration using very deep convolutional encoder-decoder networks with symmetric skip connections, Neural Inf. Process. Syst. (2016) 2802–2810.

[45] T. Salimans, I.J. Goodfellow, W. Zaremba, V. Cheung, A. Radford, X. Chen, Improved techniques for training GANs, in: In Advances in Neural Information Processing Systems, 2016, pp. 2226–2234.

[46] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, A.C. Courville, Improved training of wasserstein gans, in: Advances in Neural Information Processing Systems, 2017, pp. 5769–5779.

[47] T. Karras, T. Aila, S. Laine, J. Lehtinen, Progressive growing of gans for improved quality, stability, and variation, in: International Conferenceon Learning Representations, ICLR, 2018.

[48] T. Schlegl, P. Seeböck, S.M. Waldstein, U. Schmidt-Erfurth, G. Langs, Unsupervised anomaly detection with generative adversarial networks to guide marker discovery, in: International Conference on Information Processing in Medical Imaging, Springer, 2017, pp. 146–157.

[49] H. Dong, P. Neekhara, C. Wu, Y. Guo, Unsupervised image-to-image translation with generative adversarial networks, (2017) arXiv preprint. arXiv:1701.02676.

[50] Y. Yuan, S. Liu, J. Zhang, Y. Zhang, C. Dong, L. Lin, Unsupervised image super-resolution using cycle-in-cycle generative adversarial networks, in: The IEEE Conference on Computer Vision and Pattern Recognition, CVPR Workshops, 2018, pp. 814–823.

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

[52] G. Klambauer, T. Unterthiner, A. Mayr, S. Hochreiter, Self-normalizing neural networks, in: Advances in Neural Information Processing Systems, NIPS, 2017, pp. 972–981.

[53] V. Nair, G.E. Hinton, Rectified linear units improve restricted Boltzmann machines, in: Proceedings of the 27th International Conference on Machine Learning, ICML-10, 2010, pp. 807–814.

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

[55] M.D. Zeiler, G.W. Taylor, R. Fergus, Adaptive deconvolutional networks for mid and high level feature learning, in: IEEE International Conferenceon Computer Vision, ICCV, IEEE, 2011, pp. 2018–2025.

[56] J. Long, E. Shelhamer, T. Darrell, Fully convolutional networks for semantic segmentation, in: Computer Vision and Pattern Recognition, CVPR, 2015, pp. 3431–3440.

[57] C. Osendorfer, H. Soyer, P.V. Der Smagt, Image super-resolution with fast approximate convolutional sparse coding, in: International Conferenceon Neural Information Processing, 2014, pp. 250–257.

[58] M.D. Zeiler, R. Fergus, Image super-resolution with fast approximate convolutional sparse coding, in: International Conference on Neural Information Processing, 2014, pp. 250–257.

[59] K. Turkowski, Filters for common resampling tasks, Graph. Gems 16 (1990) 147–165.
[60] D.P. Kingma, J.L. Ba, Adam: A method for stochastic optimization, in:International Conference on Learning Representations, ICLR, 2015.
[61] D. Martin, C. Fowlkes, D. Tal, J. Malik, A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics, in: Proceedings of Eighth IEEE International Conference on Computer Vision, 2001. ICCV 2001, IEEE, 2001, pp. 416–423.
[62] M. Bevilacqua, A. Roumy, C. Guillemot, A. Morel, Low-complexity single-image super-resolution based on nonnegative neighbor embedding, in: Proc. British Machine Vision Conference, 2012, pp. 1–10.
[63] R. Zeyde, M. Elad, M. Protter, On single image scale-up using sparse-representations, in: International Conference on Curves and Surfaces, 2010, pp. 711–730.
[64] Z. Wang, A.C. Bovik, H.R. Sheikh, E.P. Simoncelli, Image quality assessment: From error visibility to structural similarity, IEEE Trans. Image Process. 13 (2004) 600–612.
[65] G. Huang, Z. Liu, K.Q. Weinberger, L. van der Maaten, Densely connected convolutional networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, p. 3.
[66] S. Ioffe, C. Szegedy, Batch normalization: Accelerating deep network training by reducing internal covariate shift, in: International Conference on International Conference on Machine Learning, ICML, 2015, pp. 448–456.