

Deep Edge Guided Recurrent Residual Learning for Image Super-Resolution

Wenhan Yang¹, Jiashi Feng², Jianchao Yang³, Fang Zhao², Jiaying Liu¹,
Zongming Guo¹, Shuicheng Yan²

¹Peking University, ²National University of Singapore, ³Snapchat Inc.

Abstract. In this work, we consider the image super-resolution (SR) problem. The main challenge of image SR is to recover high-frequency details of a low-resolution (LR) image that are important for human perception. To address this essentially ill-posed problem, we introduce a Deep Edge Guided REcurrent rEsidual (DEGREE) network to progressively recover the high-frequency details. Different from most of existing methods that aim at predicting high-resolution (HR) images directly, DEGREE investigates an alternative route to recover the difference between a pair of LR and HR images by recurrent residual learning. DEGREE further augments the SR process with edge-preserving capability, namely the LR image and its edge map can jointly infer the sharp edge details of the HR image during the recurrent recovery process. To speed up its training convergence rate, by-pass connections across multiple layers of DEGREE are constructed. In addition, we offer an understanding on DEGREE from the view-point of sub-band frequency decomposition on image signal and experimentally demonstrate how DEGREE can recover different frequency bands separately. Extensive experiments on three benchmark datasets clearly demonstrate the superiority of DEGREE over well-established baselines and DEGREE also provides new state-of-the-arts on these datasets.

1 Introduction

Image super-resolution (SR) aims at recovering a high resolution (HR) image from low resolution (LR) observations. Although it has seen wide applications, such as surveillance video recovery [1], face hallucination [2], medical image enhancement [3], the SR problem, or more concretely the involved inverse signal estimation problem therein, is essentially ill-posed and still rather difficult to solve. In order to relieve ill-posedness of the problem, most of recent SR methods propose to incorporate various prior knowledge about natural images to regularize the signal recovery process. This strategy establishes a standard maximum *a posteriori* (MAP) image SR framework [4,5], where an HR image is estimated by maximizing its fidelity to the target with kinds of *a priors*.

Most of existing MAP based image SR methods [6,7] associate the data fidelity term with the mean squared error (MSE), in order to ensure consistency between the estimated HR image and the ground truth when learning model parameters. However, solely considering minimizing MSE usually fails to recover the sharp or high-frequency details such as textures and edges. This phenomenon is also observed in much previous literature [8,9,10,11]. To address this

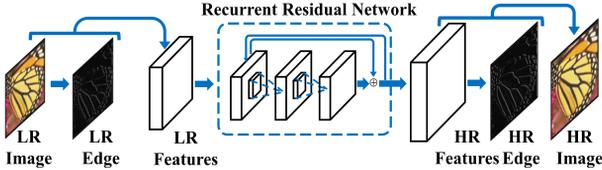


Fig. 1. The framework of the proposed DEGREE network. The recurrent residual network recovers sub-bands of the HR image features iteratively and edge features are utilized as the guidance in image SR for preserving sharp details.

problem, bandpass filters – that are commonly used to extract texture features – were employed to preserve sharp details in the image SR process [8,11,12,13]. The bandpass filters decompose an LR image into several sub-band images and build hierarchical fidelity terms to steer recovery of those sub-band images. The hierarchical fidelity consideration is shown to be able to help preserve moderate-frequency details and thus improve quality of the produced HR images.

Besides data fidelity, another important aspect for MAP based image SR methods is priors on HR images, which are effective in relieving ill-posedness of the problem. Commonly used priors describing natural image properties include sparseness [14,15], spatial smoothness [16,17] and nonlocal similarity [18], which help produce more visually pleasant HR images. Among those priors, the edge prior [19,20,21] is a very important one. In contrast to textures that are usually difficult to recover after image degradation, edges are much easier to detect in LR images and thus more informative for recovering details of HR images. Thus, separating edges from the image signal and modeling them separately would benefit image SR substantially.

Recently, several deep learning based SR methods have been developed, in order to utilize the strong capacity of deep neural networks in modeling complex image contents and details. The image super-resolution CNN (SRCNN) [22] is the seminal work that has introduced a deep convolutional network model to image SR. The proposed SRCNN consists of three convolutional layers and is equivalent to performing a sparse reconstruction to generate HR images. Benefiting from being end-to-end trainable, SRCNN improves the quality of image SR significantly. However, SRCNN only aims at minimizing the MSE loss without exploiting natural image priors and suffers from losing sharp details. Following SRCNN, several recent works [23,24] propose to embed sparsity priors into the deep networks for image SR, offering more visually pleasant results. However, much domain knowledge and extra effort are needed for designing a suitable architecture to model the sparsity priors. A simple and adaptive method to embed various priors into standard CNN networks for image SR is still absent.

Motivated by the fact that edge features can provide valuable guidance for image SR and the success of deep neural network models, we propose a Deep Edge Guided REcurrent rESidual (DEGREE) network to progressively perform image SR with properly modeled edge priors. Instead of trying to predict HR images from LR ones directly, the DEGREE model takes an alternative route and focuses on predicting the *residue* between a pair of LR and HR images, as well as the edges in HR images. Combining these predictions together give a recovered

HR image with high quality and sharp high-frequency details. An overview on the architecture of the DEGREE model is provided in Figure 1. Given an LR image, DEGREE extracts its edge features and takes the features to predict edges of the HR image via a deep recurrent network. To recover details of an HR image progressively, DEGREE adopts a recurrent residual learning architecture that recovers details of different frequency sub-bands at multiple recurrence stages. Bypass connections are introduced to fuse recovered results from previous stages and propagate the fusion results to later stages. In addition, adding bypass connections enables a deeper network trained with faster convergence rate.

In summary, our contributions to image SR can be summarized as:

1. We introduce a novel DEGREE network model to solve image SR problems. The DEGREE network models edge priors and performs image SR recurrently, and improves the quality of produced HR images in a progressive manner. DEGREE is end-to-end trainable and effective in exploiting edge priors for both LR and HR images. To the best of our knowledge, DEGREE is the first recurrent network model with residual learning for recovering HR images.
2. We provide a general framework for embedding natural image priors into image SR, which jointly predicts the task-specific targets and feature maps reflecting specific priors. It is also applicable to other image processing tasks.
3. We demonstrate that the recurrent residual learning with bypass structures, designed under the guidance of the sub-band signal reconstruction, is more effective in image SR than the standard feed forward architecture used in the modern CNN model. DEGREE outperforms well-established baselines significantly on three benchmark datasets and provides new state-of-the-arts.

2 Related Work

Many recent works have exploited deep learning for solving low level image processing problems including image denoising [25], image completion [26] and image super-resolution [27]. Particularly, Dong *et al.* [28] proposed a three layer CNN model for image SR through equally performing sparse coding. Instead of using a generic CNN model, Wang *et al.* [24] incorporated the sparse prior into CNN by exploiting a learned iterative shrinkage and thresholding algorithm (LISTA), which provided better reconstruction performance.

To address the high-frequency information loss issue in purely minimizing the MSE, sub-band decomposition based methods propose to recover information at different frequency bands of the image signal separately [8,10,11,12,13]. In [12], interpolation to high-frequency sub-band images by discrete wavelet transform (DWT) was performed for image SR. In [11], Song *et al.* proposed a joint sub-band-based neighbor-embedding SR with a constraint on each sub-band, achieving more promising SR results.

Some works also explore how to preserve edges in application of image SR, denoising and deblurring. Total variation (TV) [29,30], focusing on modeling the intensity change of image signals, was proposed to guide the SR recovery by suppressing the excessive and possibly spurious details in the HR estimation.

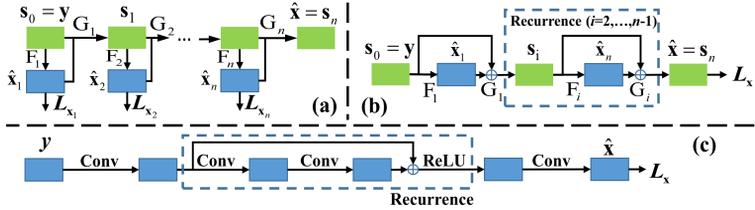


Fig. 2. (a) The flowchart of the sub-band reconstruction for image super-resolution. (b) A relaxed version of (a). G_i is set as the element-wise summation function. In this framework, only the MSE loss is used to constrain the recovery. (c) The deep network designed with the intuition of (b). G_i is the element-wise summation function and F_i is modeled by two layer convolutions.

Bilateral TV (BTV) [31,32] was then developed to preserve sharp edges. Sparsity prior [33,34] constraining the transformation coefficients was introduced to enhance salient features. As a kind of sparsity prior, the gradient prior [35,36,37] was proposed to enforce the gradient distribution of the denoised image to fit distribution estimated from the original image. By embedding these regularizations, sharper and finer edges of HR images are restored.

3 Deep Recurrent Residual Learning for Image SR

In this section we first review the sub-band reconstruction methods [12,8] for image SR. Then we illustrate how to build a recurrent residual network that can learn to perform sub-band reconstruction and recover HR images progressively.

3.1 Sub-Band Reconstruction for Image SR

In most cases, quality degradation of an HR image \mathbf{x} to an LR image \mathbf{y} is caused by blurring and down sampling, and the degradation process can be modeled as

$$\mathbf{y} = D\mathbf{H}\mathbf{x} + v, \quad (1)$$

where H and D depict the blurring and down-sampling effects respectively. The additive noise in the imaging process is denoted as v . Given an observed LR image \mathbf{y} , image SR aims at estimating the original HR \mathbf{x} . Most of image SR methods obtain an estimation of HR by solving the following MAP problem:

$$\hat{\mathbf{x}} = \arg \min_{\mathbf{x}} \|D\mathbf{H}\mathbf{x} - \mathbf{y}\|_2^2 + p(\mathbf{x}), \quad (2)$$

where $p(\cdot)$ is a regularization term induced by priors on \mathbf{x} . However, directly learning a one-step mapping function from \mathbf{y} to \mathbf{x} usually ignores some intrinsic properties hidden in different frequency bands of \mathbf{x} , such as the high-frequency edge and textural details. This is because the recovery function needs to fit the inverse mapping from the low-frequency component of the LR image to that of the HR one. It by nature neglects some high-frequency details with small energy.

To address this problem, a sub-band based image reconstruction method is proposed to recover images at different frequency bands separately. It separates the image signal into multiple components of different intrinsic frequencies, which are called sub-bands, and models them individually. In this way, the sub-bands with small energy can still gain sufficient ‘‘attention’’ and sharper image details can be preserved during image SR. Formally, let \mathbf{y}_i be the i -th sub-band of the LR image \mathbf{y} out of in total n sub-bands, *i.e.*, $\mathbf{y} = \sum_{i=1}^n \mathbf{y}_i$. \mathbf{y}_i is used for

estimating the i -th corresponding sub-band \mathbf{x}_i of the HR image \mathbf{x} . The sub-band based method recovers different sub-bands individually and outputs the recovered HR image as follows,

$$\hat{\mathbf{x}}_i = \arg \min_{\mathbf{x}_i} \|DH\mathbf{x}_i - \mathbf{y}_i\|_2^2 + p(\mathbf{x}_i), i = 1, 2, \dots, n; \hat{\mathbf{x}} = \sum_{i=1}^n \hat{\mathbf{x}}_i. \quad (3)$$

However, recovering each sub-band separately in (3) neglects the dependency across sub-bands. To fully model the dependencies both in the corresponding sub-bands and across sub-bands, we relax (3) into a progressive recovery process. It performs an iterative sub-band recovery implicitly and utilizes the useful information from lower-frequency sub-bands to recover higher-frequency ones.

For ease of explanation, we introduce an auxiliary signal \mathbf{s}_i that approximates the signal \mathbf{x} up to the i -th sub-band, *i.e.*, $\mathbf{s}_i = \sum_{j=1}^i \hat{\mathbf{x}}_j$. Then, the sub-band image \mathbf{x}_i and HR image \mathbf{x} can be estimated through recovering $\hat{\mathbf{x}}_i$ and \mathbf{s}_i progressively. We here use $F_i(\cdot)$ and $G_i(\cdot)$ to denote the generating functions of \mathbf{s}_i and $\hat{\mathbf{x}}_i$ respectively, *i.e.*,

$$\hat{\mathbf{x}}_i = F_i(\mathbf{s}_{i-1}), \mathbf{s}_i = G_i(\hat{\mathbf{x}}_i, \mathbf{s}_{i-1}), \quad (4)$$

where $\mathbf{s}_0 = \mathbf{y}$ is the input LR image and \mathbf{s}_n eventually re-produces the HR image \mathbf{x} . Figure 2(a) gives an overall illustration on this process. The functions F_i and G_i usually take linear transformations as advocated in [8,10,11]. F_i learns to recover high frequency detail, estimating the i -th sub-band component based on the accumulated recovered results from previous $(i - 1)$ sub-bands. G_i fuses $\hat{\mathbf{x}}_i$ and \mathbf{s}_{i-1} in order to balance different sub-bands. In the figure, $\mathbf{L}_{\mathbf{x}_i}$ is the loss term corresponding to the data fidelity in (3). The progressive sub-band recovery can be learned in a supervised way [8,9], where the ground truth sub-band signal \mathbf{x}_i is generated by applying band filters on \mathbf{x} . In our proposed method, we choose the element-wise summation function to model G_i in the proposed network, following the additive assumption for the sub-bands of the image signal that is generally implied in previous methods [12,8].

3.2 Learning Sub-Band Decomposition by Recurrent Residual Net

The sub-band paradigm mentioned above learns to recover HR images through minimizing a hierarchical loss generated by applying hand-crafted frequency domain filters, as shown in Figure 2(a). However, this paradigm suffers from following two limitations. First, it does not provide an end-to-end trainable framework. Second, it suffers from the heavy dependence on the choice of the frequency filters. A bad choice of the filters would severely limit its capacity of modeling the correlation between different sub-bands, and recovering the HR \mathbf{x} .

To handle these two problems, by employing a summation function as G_i , we reformulate the recover process in (4) into:

$$\mathbf{s}_i = \mathbf{s}_{i-1} + F_i(\mathbf{s}_{i-1}). \quad (5)$$

In this way, the intermediate estimation $\hat{\mathbf{x}}_i$ is not necessary to estimate explicitly. An end-to-end training paradigm can then be constructed as shown in Figure 2(b). The MSE loss $\mathbf{L}_{\mathbf{x}}$ imposed at the top layer is the only constraint on $\hat{\mathbf{x}}$ for the HR prediction. Motivated by (5) and Figure 2(b), we further propose a recurrent residual learning network whose architecture is shown in Figure 2(c). To increase the modeling ability, F_i is parameterized by two layers of convolutions.

To introduce nonlinearities into the network, G_i is modeled by an element-wise summation connected with a non-linear rectification. Training the network to minimize the MSE loss gives the functions F_i and G_i adaptive to the training data. Then, we stack n recurrent units into a deep network to perform a progressive sub-band recovery. Our proposed recurrent residual network follows the intuition of gradual sub-band recovery process. The proposed model is equivalent to balancing the contributions of each sub-band recovery. Benefiting from the end-to-end training, such deep sub-band learning is more effective than the traditional supervised sub-band recovery. Furthermore, the proposed network indeed has the ability to recover the sub-bands of the image signal recurrently, as validated in Section 5.4.

4 DEGREE Network for Edge Preserving SR

We have presented how to construct a recurrent residual network to perform deep sub-band learning. In this section, we proceed to explain how to embed the edge prior into the recurrent residual network, in order to predict high-frequency details better for image SR.

4.1 Edge Extraction

An HR image \mathbf{x} can be separated into low-frequency and high-frequency components, as $\mathbf{x} = \mathbf{x}_L + \mathbf{x}_H$, where the high-frequency component \mathbf{x}_H contains subtle details of the image, such as edges and textures. Patterns contained in \mathbf{x}_H are usually irregular and have smaller magnitude compared with \mathbf{x}_L . Thus, in image degradation, the component of \mathbf{x}_H is more fragile and easier to be corrupted, which is also difficult to recover. To better recover \mathbf{x}_H , we propose to extract extra prior knowledge about \mathbf{x}_H from the LR image \mathbf{y} as a build-in component in the deep recurrent residual network to regularize the recovery process. Among all the statistical priors about natural images, edge is one of the most informative priors. Therefore, we propose to model edge priors and develop a deep edge guided recurrent residual network, which is introduced in the following section. However, our proposed network architecture can also embed other statistical priors extractable from LR inputs for image SR. To extract edges, we first apply an off-the-shelf edge detector (such as the Sobel one) on \mathbf{y} and \mathbf{x} to get its high-frequency component \mathbf{y}_H and \mathbf{x}_H . Then we train the model to predict \mathbf{x}_H based on both \mathbf{y} and \mathbf{y}_H . Please note that \mathbf{x}_H is the high-frequency residual of \mathbf{x} .

4.2 DEGREE Network

We propose an end-to-end trainable deep edge guided recurrent residual network (DEGREE) for image SR. The network is constructed based on the following two intuitions. First, as we have demonstrated, a recurrent residual network is capable of learning sub-band decomposition and reconstruction for image SR. Second, modeling edges extracted from the LR image would benefit recovery of details in the HR image. An overview on the architecture of the proposed DEGREE network is given in Figure 3. As shown in the figure, DEGREE contains following components. **a) LR Edge Extraction.** An edge map of the input LR image is extracted by applying a hand-crafted edge detector and is

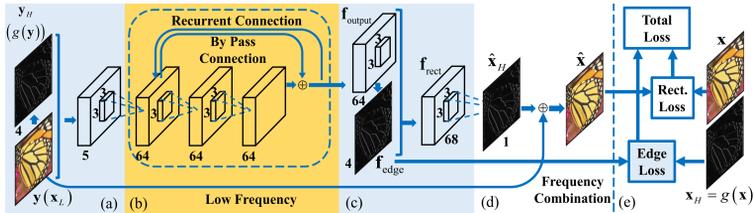


Fig. 3. The architecture of the DEGREE network for image SR. (a) The LR edge maps $y_H(g(y))$ of the LR image are part of the input features. (b) Recurrent residual learning network for sub-band recovery. (c) Part of the feature maps f_{edge} in the penultimate layer aim at generating HR edges. (d) Combining the high-frequency estimation and the LR image by $\hat{x} = x_L + \hat{x}_H$. (e) The total loss is the combination of the edge loss and reconstruction loss, which constrain the recovery of HR edges and HR images respectively. Our main contributions, the edge guidance and recurrent residual learning, are highlighted with blue and orange colors.

fed into the network together with the raw LR image, as shown in Figure 3(a). **b) Recurrent Residual Network.** The mapping function from LR images to HR images is modeled by the recurrent residual network as introduced in Section 3.2. Instead of predicting the HR image directly, DEGREE recovers the residual image at different frequency sub-bands progressively and combine them into the HR image, as shown in Figure 3(b). **c) HR Edge Prediction.** DEGREE produces convolutional feature maps in the penultimate layer, part of which (f_{edge}) are used to reconstruct the edge maps of the HR images and provide extra knowledge for reconstructing the HR images, as shown in Figure 3(c). **d) Sub-Bands Combination For Residue.** Since the LR image contains necessary low-frequency details, DEGREE only focuses on recovering the high-frequency component, especially several high-frequency sub-bands of the HR image, which are the differences or *residue* between the HR image and the input LR image. Combining the estimated residue with sub-band signals and the LR image gives an HR image, as shown in Figure 3(d). **e) Training Loss.** We consider the reconstruction loss of both the HR image and HR edges simultaneously for training DEGREE as shown in Figure 3(e). We now explain each individual part of the proposed network in details.

Recurrent Residual Network The recurrent residual network aims to refine SR images progressively through producing the residue image at different frequency. We follow the notations in Section 3. To provide a formal description, let \mathbf{f}_{in}^k denote the input feature map for the recurrent sub-network at the k -th time step. The output feature map $\mathbf{f}_{\text{out}}^k$ of the recurrent sub-network is progressively updated as follows,

$$\mathbf{f}_{\text{out}}^k = \max\left(0, \mathbf{W}_{\text{mid}}^k * \mathbf{f}_{\text{mid}}^k + \mathbf{b}_{\text{mid}}^k\right) + \mathbf{f}_{\text{in}}^k, \text{ with } \mathbf{f}_{\text{mid}}^k = \max\left(0, \mathbf{W}_{\text{in}}^k * \mathbf{f}_{\text{in}}^k + \mathbf{b}_{\text{in}}^k\right), \quad (6)$$

where $\mathbf{f}_{\text{in}}^k = \mathbf{f}_{\text{out}}^{k-1}$ is the output features by the recurrent sub-network at $(k-1)$ -th time step. Please note the by-pass connection here between \mathbf{f}_{in}^k and $\mathbf{f}_{\text{out}}^k$. In the context of sub-band reconstruction, the feature map $\mathbf{f}_{\text{out}}^k$ can be viewed as the recovered k -th sub-band of the image signal. Let K be the total recurrence

number of the sub-networks. Then, the relation between \mathbf{f}_{in}^1 , $\mathbf{f}_{\text{out}}^K$ and the overall network is

$$\begin{aligned}\mathbf{f}_{\text{in}}^1 &= \max(0, \mathbf{W}_{\text{input}} * \mathbf{f}_{\text{input}} + \mathbf{b}_{\text{input}}), \\ \mathbf{f}_{\text{output}} &= \mathbf{f}_{\text{out}}^K,\end{aligned}\quad (7)$$

where $\mathbf{W}_{\text{input}}$ and $\mathbf{b}_{\text{input}}$ denote the filter parameter and basis of the convolution layer before the recurrent sub-network. Thus, $\mathbf{f}_{\text{output}}$ is the output features of the recurrent residual network, which are used to reconstruct both the HR features and images.

Edge Modeling We here illustrate how to embed the edge information into the proposed deep network. This can also generalize to modeling other natural image priors. In particular, the proposed network takes edge features extracted from the LR image as another input, and aims to predict edge maps of the HR image as a part of its output features which are then used for recovering the HR image.

The input feature $\mathbf{f}_{\text{input}}$ to the network is a concatenation of the raw LR image \mathbf{y} and its edge map $g(\mathbf{y})$,

$$\mathbf{f}_{\text{input}} = [\mathbf{y}, g(\mathbf{y})]. \quad (8)$$

To recover the HR image, DEGREE outputs two types of features at its penultimate layer. One is for HR image recovery and the other one is for edge prediction in the HR image. More specifically, let $\mathbf{f}_{\text{output}}$ denote the features used to reconstruct HR images and let \mathbf{f}_{edge} denote the edge feature computed by

$$\mathbf{f}_{\text{edge}} = \max(0, \mathbf{W}_{\text{edge}} * \mathbf{f}_{\text{output}} + \mathbf{b}_{\text{edge}}), \quad (9)$$

where \mathbf{W}_{edge} and \mathbf{b}_{edge} are the filter and the bias of the convolution layer to predict the HR edge map. Thus, the features \mathbf{f}_{rect} in the penultimate layer for reconstructing the HR image with the edge guidance are given as follows,

$$\mathbf{f}_{\text{rect}} = [\mathbf{f}_{\text{output}}, \mathbf{f}_{\text{edge}}]. \quad (10)$$

Sub-Bands Combination In sub-band based image SR methods, the low-frequency and high-frequency components of an image signal are usually extracted at different parts in a hierarchical decomposition of the signal. DEGREE network also models the low-frequency and high-frequency components of an image jointly. Denote the high-frequency and low-frequency components of an HR image \mathbf{x} as \mathbf{x}_H and \mathbf{x}_L respectively. We have $\mathbf{x} = \mathbf{x}_H + \mathbf{x}_L$. Here, we use the notation \mathbf{y} to denote both the original LR image and its up-scaled version of the same size as \mathbf{x} , if it causes no confusion. Obviously, \mathbf{y} is a good estimation of the low frequency component \mathbf{x}_L of the HR image \mathbf{x} . The retained high-frequency component \mathbf{y}_H of \mathbf{y} , *i.e.*, the edge map of \mathbf{y} , is estimated by applying an edge extractor (we use Sobel) onto \mathbf{y} . In our proposed DEGREE network, as shown in Figure 3, the low-frequency component $\mathbf{x}_L \approx \mathbf{y}$ is directly passed to the last layer and combined with the predicted high-frequency image $\widehat{\mathbf{x}}_H$ to produce an estimation $\widehat{\mathbf{x}}$ of the HR image \mathbf{x} : $\widehat{\mathbf{x}} = \mathbf{x}_L + \widehat{\mathbf{x}}_H$. Here, $\widehat{\mathbf{x}}_H$, an estimation of the high-frequency component \mathbf{x}_H , is generated by

$$\widehat{\mathbf{x}}_H = \max(0, \mathbf{W}_{\text{rect}} * \mathbf{f}_{\text{rect}} + \mathbf{b}_{\text{rect}}), \quad (11)$$

where \mathbf{f}_{rect} is the features learned in the penultimate layer to reconstruct \mathbf{x}_H . The filters and biases involved in the layer are denoted as \mathbf{W}_{rect} and \mathbf{b}_{rect} .

Training Let $\mathbf{F}(\cdot)$ represent the learned network for recovering the HR image \mathbf{x} based on the input LR image \mathbf{y} and the LR edge map \mathbf{y}_H . Let $\mathbf{F}_{\text{edge}}(\cdot)$ denote the learned HR edge predictor which outputs \mathbf{f}_{edge} . We use Θ to collectively denote all the parameters of the network,

$$\Theta = \{\mathbf{W}_{\text{input}}, \mathbf{b}_{\text{input}}, \mathbf{W}_{\text{in}}, \mathbf{b}_{\text{in}}, \mathbf{W}_{\text{mid}}, \mathbf{b}_{\text{mid}}, \mathbf{W}_{\text{edge}}, \mathbf{b}_{\text{edge}}, \mathbf{W}_{\text{rect}}, \mathbf{b}_{\text{rect}}\}. \quad (12)$$

Given n pairs of HR and LR images $\{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^n$ for training, we first extract the high-frequency components of LR and HR images, $\{\mathbf{y}_{i,H}\}$ and $\{\mathbf{x}_{i,H}\}$, by applying Sobel operator on the image \mathbf{x}_i and \mathbf{y}_i respectively. We adopt the following joint mean squared error (MSE) to train the network parameterized by Θ such that it can jointly estimate the HR images and HR edge maps:

$$L(\Theta) = \frac{1}{n} \sum_{i=1}^n (\|\mathbf{F}(\mathbf{y}_i, \mathbf{y}_{i,H}, \mathbf{x}_i, \mathbf{x}_{i,H}; \Theta) - \mathbf{x}_i\|^2 + \lambda \|\mathbf{F}_{\text{edge}}(\mathbf{y}_i, \mathbf{y}_{i,H}, \mathbf{x}_i, \mathbf{x}_{i,H}; \Theta) - \mathbf{x}_{i,H}\|^2).$$

Here λ is a trade-off parameter that balances importance of the data fidelity term and the edge prior term. We empirically set λ as 1 throughout the paper because we observe that our method performs similarly for different values of λ in a large range, as mentioned in Section 5 and validated in supplementary material.

5 Experiments

Datasets Following the experimental setting in [43] and [44], we compare the proposed method with recent SR methods on three popular benchmark datasets: Set5 [38], Set14 [39] and BSD100 [40] with scaling factors of 2, 3 and 4. The three datasets contain 5, 14 and 100 images respectively. Among them, the Set5 and Set14 datasets are commonly used for evaluating traditional image processing methods, and the BSD100 dataset contains 100 images with diverse natural scenes. We train our model using a training set created in [6], which contains 91 images. For fair comparison with other methods [24], we do not train our models with a larger dataset. We either do not use any ad-hoc post-processing.

Baseline Methods We compare our DEGREE SR network (DEGREE) with Bicubic interpolation and the following six state-of-the-art SR methods: ScSR (Sparse coding) [33], A+ (Adjusted Anchored Neighborhood Regression) [41], SRCNN [28], TSE-SR (Transformed Self-Exemplars) [42], CSCN (Deep Sparse Coding) [24] and JSB-NE (Joint Sub-Band Based Neighbor Embedding) [11]. It is worth noting that CSCN and JSB-NE are the most recent deep learning and sub-band recovery based image SR methods respectively.

Implementation Details We evaluate our proposed model with 10 and 20 layers respectively. The bypass connections are set with an interval of 2 convolution layers, as illustrated in Figure 3. The number of channels in each convolution layer is fixed as 64 and the filter size is set as 3×3 with a padding size of 1. All these settings are consistent with the one used in [43]. The edge extractor is applied along four directions (up-down, down-up, left-right and right-left) for extracting edge maps. Following the experimental setting in [28], we generate LR images by applying Bicubic interpolation on the HR images. The training and validation images are cropped into small sub-images with a size of 33×33 pixels. We use flipping (up-down and left-right) and clockwise rotations ($0^\circ, 90^\circ, 180^\circ$

and 270°) for data augmentation. For each training image, 16 augmented images are generated. The final training set contains around 240,000 sub-images. The weighting parameter λ for balancing the losses is empirically set as 1. We empirically show that the DEGREE network is robust to the choice of λ in the supplementary material and the best performance is provided by setting $\lambda \leq 1$. Following the common practice in many previous methods, we only perform super-resolution in the luminance channel (in YCrCb color space). The other two chrominance channels are bicubically upsampled for displaying the results. We train our model on the Caffe platform [44]. Stochastic gradient descent (SGD) with standard back-propagation is used for training the model. In particular, in the optimization we set momentum as 0.9, the initial learning rate as 0.0001 and change it to 0.00001 after 76 epochs. We only allow at most 270 epochs.

5.1 Objective Evaluation

We use DEGREE-1 and DEGREE-2 to denote two versions of the proposed model when we report the results. DEGREE-1 has 10 layers and 64 channels, and DEGREE-2 has 20 layers and 64 channels. The quality of the HR images produced by different SR methods is measured by the Peak Signal-to-Noise Ratio (PSNR) [45] and the perceptual quality metric Structural SIMilarity (SSIM) [46], which are two widely used metrics in image processing. The results of our proposed DEGREE-1 and DEGREE-2 as well as the baselines are given in Table 1.

Table 1. Comparison among different image SR methods on three test datasets with three scale factors ($\times 2$, $\times 3$ and $\times 4$). The bold numbers denote the best performance and the underlined numbers denote the second best performance. The performance gain of DEGREE-2 over the best baseline results is shown in the last row.

Dataset		Set5			Set14			BSD100		
Method	Metric	$\times 2$	$\times 3$	$\times 4$	$\times 2$	$\times 3$	$\times 4$	$\times 2$	$\times 3$	$\times 4$
Bicubic	PSNR	33.66	30.39	28.42	30.13	27.47	25.95	29.55	27.20	25.96
	SSIM	0.9096	0.8682	0.8105	0.8665	0.7722	0.7011	0.8425	0.7382	0.6672
ScSR	PSNR	35.78	31.34	29.07	31.64	28.19	26.40	30.77	27.72	26.61
	SSIM	0.9485	0.8869	0.8263	0.8990	0.7977	0.7218	0.8744	0.7647	0.6983
A+	PSNR	36.56	32.60	30.30	32.14	29.07	27.28	30.78	28.18	26.77
	SSIM	0.9544	0.9088	0.8604	0.9025	0.8171	0.7484	0.8773	0.7808	0.7085
TSE-SR	PSNR	36.47	32.62	30.24	32.21	29.14	27.38	31.18	28.30	26.85
	SSIM	0.9535	0.9092	0.8609	0.9033	0.8194	0.7514	0.8855	0.7843	0.7108
JSB-NE	PSNR	36.59	32.32	30.08	32.34	28.98	27.22	31.22	28.14	26.71
	SSIM	0.9538	0.9042	0.8508	0.9058	0.8105	0.7393	0.8869	0.7742	0.6978
CNN	PSNR	36.34	32.39	30.09	32.18	29.00	27.20	31.11	28.20	26.70
	SSIM	0.9521	0.9033	0.8530	0.9039	0.8145	0.7413	0.8835	0.7794	0.7018
CNN-L	PSNR	36.66	32.75	30.49	32.45	29.30	27.50	31.36	28.41	26.90
	SSIM	0.9542	0.9090	0.8628	0.9067	0.8215	0.7513	0.8879	0.7863	0.7103
CSCN	PSNR	36.88	33.10	30.86	32.50	29.42	27.64	31.40	28.50	27.03
	SSIM	0.9547	0.9144	0.8732	0.9069	0.8238	0.7573	0.8884	0.7885	0.7161
DEGREE-1	PSNR	<u>37.29</u>	<u>33.29</u>	<u>30.88</u>	<u>32.87</u>	<u>29.53</u>	<u>27.69</u>	<u>31.66</u>	<u>28.59</u>	<u>27.06</u>
	SSIM	<u>0.9574</u>	<u>0.9164</u>	<u>0.8726</u>	<u>0.9103</u>	<u>0.8265</u>	<u>0.7574</u>	0.8962	<u>0.7916</u>	0.7177
DEGREE-2	PSNR	37.40	33.39	31.03	32.96	29.61	27.73	31.73	28.63	27.07
	SSIM	0.9580	0.9182	0.8761	0.9115	0.8275	0.7597	<u>0.8937</u>	0.7921	0.7177
Gain	PSNR	0.52	0.29	0.17	0.46	0.19	0.09	0.26	0.13	0.04
	SSIM	0.0033	0.0038	0.0029	0.0046	0.0037	0.0025	0.0053	0.0036	0.0016

From the table, it can be seen that the our proposed DEGREE models consistently outperform those well-established baselines with significant performance gains. DEGREE-2 performs the best for all the three scaling factors on the three datasets, except for the setting of $\times 2$ on BSD100 in terms of SSIM, where

DEGREE-1 performs the best. Comparing the performance of DEGREE-1 and DEGREE-2 clearly demonstrates that increasing the depth of the network indeed improves the performance, but we observe that further increasing the depth leads to no performance gain. We also list the concrete performance gain brought by the proposed DEGREE model over the state-of-the-art (CSCN). One can observe that when enlarging the image by a factor of 2, our proposed method can further improve the state-of-the-art performance with a margin up to 0.52 (PSNR) and 0.0033 (SSIM) on Set5. For other scaling factors, our method also consistently provides better performance. For example, on the Set5 dataset, DEGREE-2 improves the performance by 0.29 and 0.17 for $\times 3$ and $\times 4$ settings respectively. Our models are more competitive for a small scale factor. This might be because edge features are more salient and are easier to be predicted in small scaling enlargements. This is also consistent with the observation made for the gradient statistics in the previous edge-guided SR method [21].

5.2 Subjective Evaluation

We also present some visual results in Figures 6, 9 and 10 to investigate how the methods perform in terms of visual quality. These results are generated by our proposed network with 20 layers, *i.e.* DEGREE-2. Since our method is significantly better than baselines for the scaling factor of 2, here we in particular focus on comparing the visual quality of produced images with larger scaling factors. Figure 6 displays the SR results on the image of *Butterfly* from Set5 for $\times 4$ enlargement. From the figure, one can observe that the results generated by A+, SRCNN and JSB-NE contain artifacts or blurred details. CSCN provides fewer artifacts. But there are still a few remained, such as the corners of yellow and white plaques as shown in the enlarged local result in Figure 6. Our method generates a more visually pleasant image with clean details and sharp edges. For the image *86000* from BSD100, as shown in Figure 9, our method produces an image with the cleanest window boundary. For the image *223061* from BSD100 in Figure 10 that contains a lot of edges and texture, most of methods generate the results with severe artifacts. Benefiting from explicitly exploiting the edge prior, our method produces complete and sharp edges as desired. Note that more visual results are presented in the supplementary material due to space limitation.

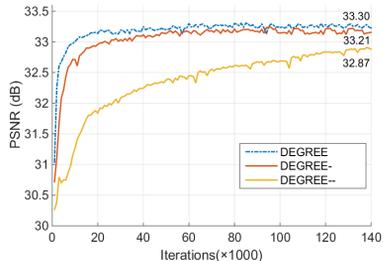
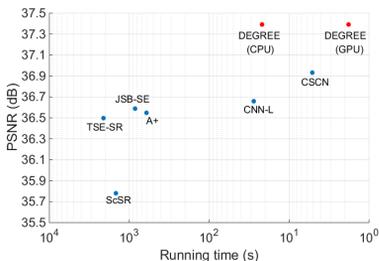


Fig. 4. The performance of our method compared with state-of-the-art methods, including the effectiveness and time complexity, in $2\times$ enlargement on dataset Set5.

Fig. 5. The comparison of three versions of the proposed method in $3\times$ enlargement on dataset Set5.

5.3 Running Time

We report time cost of our proposed model and compare its efficiency with other methods. Figure 4 plots their running time (in secs.) against performance (in PSNR). All the compared methods are implemented using the public available codes from the authors. We implement our method using Caffe with its Matlab wrapper. We evaluate the running time of all the algorithms with following machine configuration: Intel X5675 3.07GHz and 24 GB memory. The GPU version of our method costs 1.81 seconds for performing SR on all the images of Set5, while other methods are significantly slower than ours in orders. The CPU version of our method is comparable to other deep learning-based SR methods, including CSCN and CNN-L.

5.4 Discussions

We further provide additional investigations on our model in depth, aiming to give more transparent understandings on its effectiveness.

Ablation Analysis We here perform ablation studies to see the individual contribution of each component in our model to the final performance. We first observe that, *without* by-pass connections, the training process of our proposed model could not converge within 40,000 iterations, for various learning rates (0.1, 0.01 and 0.001). This demonstrates that adding by-pass connections indeed speeds up the convergence rate of a deep network. In the following experiments, we always keep the bypass connections and evaluate the performance of the following three variants of our model: a vanilla one without edge prior or frequency combination (denoted as DEGREE--), the one without frequency combination (denoted as DEGREE-) and the full model. Figure 5 shows their training performance (plotted in curves against number of iterations) and testing performance (shown in digits) in PSNR on the dataset Set5, for $\times 3$ enlargement. From the results, one can observe that modeling the edge prior boosts the performance significantly and introducing frequency combination further improves the performance.

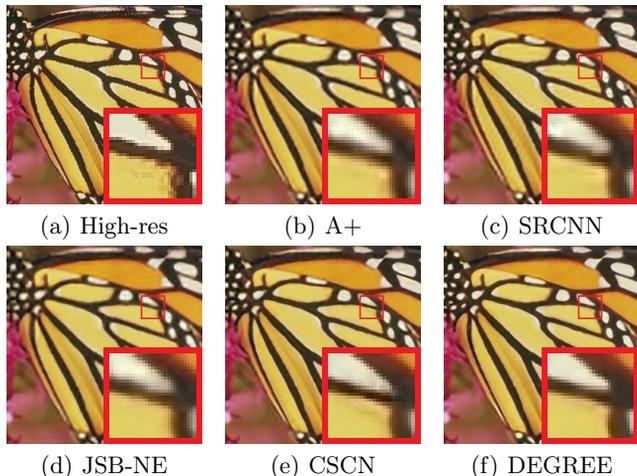


Fig. 6. Visual comparisons between different algorithms for the image *butterfly* ($4\times$). The DEGREE avoids the artifacts near the corners of the white and yellow plaques.

Model Size We investigate how the size of the model, including number of layers and size of channels within each layer, influences the final performance. We compare performance of our model with different pairs of ($\#$ layers, $\#$ channels) in Figure 7. It can be seen that a large model with more than $(20, 32) \times 10^5$ and $(8, 64) \times 10^5$ parameters (shown as yellow points) is necessary for achieving reasonably good performance. The combination of $(20, 8) \times 10^4$ (the purple point) results in a model with the same size of SCN64 (the green point where its dictionary size is equal to 64) and achieves almost the same performance. Further increasing the model size to $(20, 16) \times 10^4$ (the higher purple point) gives a better result than SCN128 (with a dictionary size of 64), whose model size is slightly smaller.

Visualization of Learned Sub-Bands We also visualize the learned features from the bottom feature extraction layer (denoted as 1L) and four recurrent time steps (denoted as \cdot R). The results are produced by a network with 10 layers for the $\times 2$ testing case. The reconstructed results of *Butterfly* at different layers are shown in Figure 8. One can observe that the proposed model captures details at different frequencies, similar to sub-band decomposition. The 1L layer extracts and enhances the edge features remarkably but brings some artifacts. The 1R layer enhances edges and makes up some false enhancements. In 2R and 3R, the sub-bands contain textures. The 4R layer fixes details. In all, for the whole network, previous layers’ sub-bands contain edge features. Later ones include texture features. The sub-band of the last layer models the “residual signal”. **More visual results are presented in the supplementary material.**

Application in JPEG Artifacts Reduction It is worth mentioning that the DEGREE network is a general framework in which the prior knowledge is embedded, by properly setting $g(\mathbf{y})$ in $\mathbf{f}_{\text{input}}$ and replacing \mathbf{f}_{edge} with the feature maps representing specific priors. For example, for JPEG artifacts reduction, we take as input the edge maps of the compressed image and the block map of DCT transformation, *i.e.* $g(\mathbf{y})$, a part of preliminary feature maps. Then we let the network predict \mathbf{f}_{edge} consisting of the general edge maps and the edges only overlapped with the block boundary of the high-quality one, which are a part of feature maps of the penultimate layer. The block and feature maps in fact impose the priors about the blockness and edges on the network. **Results about DEGREE on JPEG artifacts reduction are presented in the supplementary material.**

6 Conclusions

In this paper, we proposed a deep edge guided recurrent residual network for image SR. The edge information is separated out from the image signal to guide the recovery of the HR image. The extracted LR edge maps are used as parts of the input features and the HR edge maps are utilized to constrain the learning of parts of feature maps for image reconstruction. The recurrent residual learning structure with by-pass connections enables the training of deeper networks. Extensive experiments have validated the effectiveness of our method for producing HR images with richer details. Furthermore, this paper presented a general

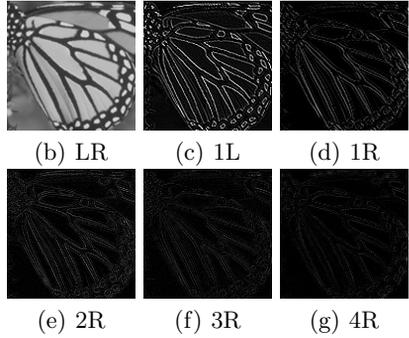
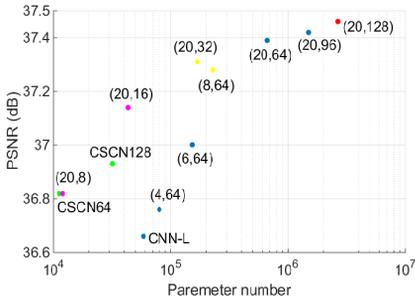


Fig. 7. PSNR for $2\times$ SR on Set5 with various parameter numbers, compared with CSCN and CNN.

Fig. 8. The visualization of the learned sub-bands in the recovery on *Butterfly*.

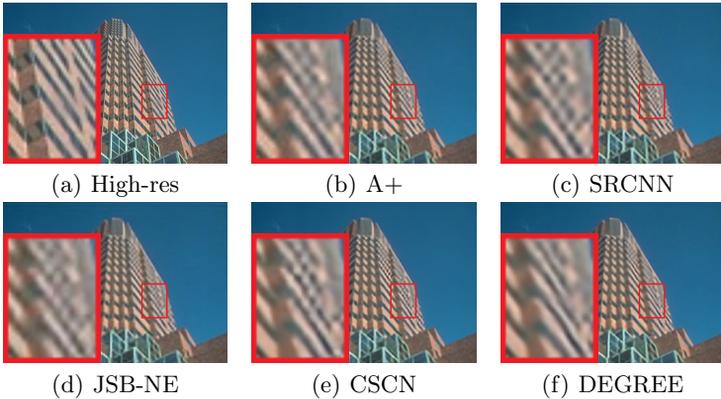


Fig. 9. Visual comparisons between different algorithms for the image 86000 ($3\times$). The DEGREE presents less artifacts around the window boundaries.

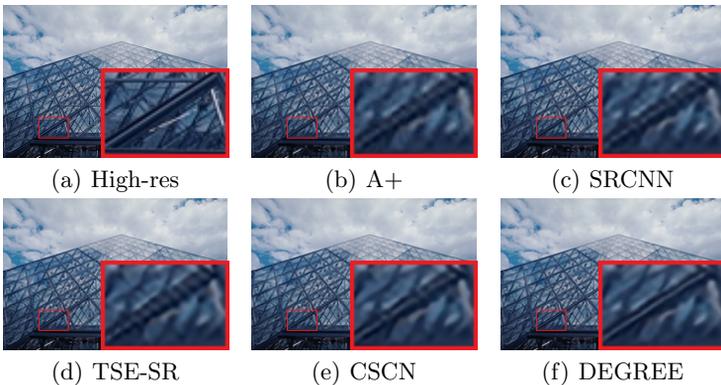


Fig. 10. Visual comparisons between different algorithms for the image 223061 ($3\times$). The DEGREE produces more complete and sharper edges.

framework for embedding various natural image priors into image processing tasks.

References

1. Zhang, L., Zhang, H., Shen, H., Li, P.: A super-resolution reconstruction algorithm for surveillance images. *Signal Processing* **90**(3) (2010) 848 – 859
2. Liu, C., Shum, H.Y., Freeman, W.T.: Face hallucination: Theory and practice. *Int'l Journal of Computer Vision* **75**(1) (October 2007) 115–134
3. Greenspan, H.: Super-resolution in medical imaging. *Comput. J.* **52**(1) (January 2009) 43–63
4. Belekos, S.P., Galatsanos, N.P., Katsaggelos, A.K.: Maximum a posteriori video super-resolution using a new multichannel image prior. *IEEE Transactions on Image Processing* **19**(6) (June 2010) 1451–1464
5. Pickup, L.C., Capel, D.P., Roberts, S.J., Zisserman, A.: Bayesian image super-resolution, continued. In: *Proc. Annual Conference on Neural Information Processing Systems*. (2006) 1089–1096
6. Yang, J., Wright, J., Huang, T., Ma, Y.: Image super-resolution via sparse representation. *IEEE Transactions on Image Processing* **19**(11) (Nov 2010) 2861–2873
7. Sun, J., Sun, J., Xu, Z., Shum, H.Y.: Image super-resolution using gradient profile prior. In: *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition, IEEE* (2008) 1–8
8. Singh, A., Ahuja, N.: Sub-band energy constraints for self-similarity based super-resolution. In: *Proc. IEEE Int'l Conf. Pattern Recognition*. (Aug 2014) 4447–4452
9. Chatterjee, P., Namboodiri, V.P., Chaudhuri, S.: Super-resolution using sub-band constrained total variation. In: *Proceedings of the first International Conference on Scale Space and Variational Methods in Computer Vision*. (2007) 616–627
10. Singh, A., Ahuja, N.: Super-resolution using sub-band self-similarity. In: *Proc. IEEE Asia Conf. Computer Vision*. (2014) 552–568
11. Song, S., Li, Y., Liu, J., Zongming, G.: Joint sub-band based neighbor embedding for image super resolution. In: *Proc. IEEE Int'l Conf. Acoustics, Speech, and Signal Processing*. (2016)
12. Gholamreza, A., Hasan, D.: Image super resolution based on interpolation of wavelet domain high frequency subbands and the spatial domain input image. *ETRI Journal* **32**(3) (June 2010) 390–394
13. Chatterjee, P., Namboodiri, V.P., Chaudhuri, S.: Super-Resolution Using Sub-band Constrained Total Variation. In: *In Proceedings of Scale Space and Variational Methods in Computer Vision: First International Conference*. (2007) 616–627
14. Daubechies, I., Defrise, M., De Mol, C.: An iterative thresholding algorithm for linear inverse problems with a sparsity constraint. *Communications on Pure and Applied Mathematics* **57**(11) (2004) 1413–1457
15. Tropp, J., Wright, S.: Computational methods for sparse solution of linear inverse problems. *Proceedings of the IEEE* **98**(6) (June 2010) 948–958
16. Marquina, A., Osher, S.J.: Image super-resolution by tv-regularization and bregman iteration. *J. Sci. Comput.* **37**(3) (December 2008) 367–382
17. Aly, H., Dubois, E.: Image up-sampling using total-variation regularization with a new observation model. *IEEE Transactions on Image Processing* **14**(10) (Oct 2005) 1647–1659
18. Peyr, G., Bougleux, S., Cohen, L.: Non-local regularization of inverse problems. In: *Proc. IEEE European Conf. Computer Vision*. Volume 5304. (2008) 57–68
19. Zhou, Q., Chen, S., Liu, J., Tang, X.: Edge-preserving single image super-resolution. In: *ACM Trans. Multimedia*. (2011) 1037–1040

20. Dai, S., Han, M., Xu, W., Wu, Y., Gong, Y.: Soft Edge Smoothness Prior for Alpha Channel Super Resolution. In: Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition. (2007)
21. Tai, Y.W., Liu, S., Brown, M., Lin, S.: Super resolution using edge prior and single image detail synthesis. In: Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition. (June 2010) 2400–2407
22. Dong, C., Loy, C.C., He, K., Tang, X.: Image super-resolution using deep convolutional networks. In: Proc. IEEE European Conf. Computer Vision. (2014)
23. Osendorfer, C., Soyer, H., Smagt, P.: Image Super-Resolution with Fast Approximate Convolutional Sparse Coding. In: Proceedings of International Conference Neural Information Processing. (2014) 250–257
24. Wang, Z., Liu, D., Yang, J., Han, W., Huang, T.: Deep networks for image super-resolution with sparse prior. In: Proc. IEEE Int'l Conf. Computer Vision. (June 2015)
25. Vincent, P., Larochele, H., Lajoie, I., Bengio, Y., Manzagol, P.A.: Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *Journal of Machine Learning Research* (2010)
26. Xie, J., Xu, L., Chen, E.: Image denoising and inpainting with deep neural networks. In: Proc. Annual Conference on Neural Information Processing Systems. (2012)
27. Dong, C., Loy, C., He, K., Tang, X.: Image super-resolution using deep convolutional networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2015)
28. Dong, C., Loy, C., He, K., Tang, X.: Learning a deep convolutional network for image super-resolution. In Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T., eds.: Proc. IEEE European Conf. Computer Vision. Volume 8692. (2014) 184–199
29. Aly, H., Dubois, E.: Image up-sampling using total-variation regularization with a new observation model. *IEEE Transactions on Image Processing* **14**(10) (Oct 2005) 1647–1659
30. Marquina, A., Osher, S.J.: Image super-resolution by tv-regularization and bregman iteration. *J. Sci. Comput.* **37**(3) (December 2008) 367–382
31. Farsiu, S., Robinson, D., Elad, M., Milanfar, P.: Robust Shift and Add Approach to Super-Resolution. In: Proceedings the International Society for Optical Engineering. Volume 5203. (2003) 121–130
32. Ye, G., Wang, Y., Xu, J., Herman, G., Zhang, B.: A practical approach to multiple super-resolution sprite generation. In: Proc. IEEE Workshop on Multimedia Signal Processing. (Oct 2008) 70–75
33. Yang, J., Wright, J., Huang, T., Ma, Y.: Image super-resolution via sparse representation. *IEEE Transactions on Image Processing* **19**(11) (Nov 2010) 2861–2873
34. Dong, W., Zhang, D., Shi, G., Wu, X.: Image deblurring and super-resolution by adaptive sparse domain selection and adaptive regularization. *IEEE Transactions on Image Processing* **20**(7) (July 2011) 1838–1857
35. Zuo, W., Zhang, L., Song, C., Zhang, D., Gao, H.: Gradient histogram estimation and preservation for texture enhanced image denoising. *IEEE Transactions on Image Processing* **23**(6) (June 2014) 2459–2472
36. Levin, A., Weiss, Y., Durand, F., Freeman, W.: Efficient marginal likelihood optimization in blind deconvolution. In: Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition. (June 2011) 2657–2664
37. Sun, J., Sun, J., Xu, Z., Shum, H.Y.: Image super-resolution using gradient profile prior. In: Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition, IEEE (2008) 1–8

38. Bowden, R., Collomosse, J.P., Mikolajczyk, K., eds.: British Machine Vision Conference, BMVA Press (2012)
39. Zeyde, R., Elad, M., Protter, M.: On single image scale-up using sparse-representations. In: Proceedings of the 7th International Conference on Curves and Surfaces, Berlin, Heidelberg, Springer-Verlag (2012) 711–730
40. Martin, D., Fowlkes, C., Tal, D., Malik, J.: A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In: Proc. IEEE Int’l Conf. Computer Vision. Volume 2. (July 2001) 416–423
41. Timofte, R., DeSmet, V., VanGool, L.: A+: Adjusted anchored neighborhood regression for fast super-resolution. In: Proc. IEEE Asia Conf. Computer Vision. Volume 9006. (2015) 111–126
42. Huang, J.B., Singh, A., Ahuja, N.: Single image super-resolution from transformed self-exemplars. In: Proc. IEEE Int’l Conf. Computer Vision and Pattern Recognition. (2015) 5197–5206
43. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. Arxiv: **1512.03385** (2015)
44. Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S., Darrell, T.: Caffe: Convolutional architecture for fast feature embedding. ArXiv: 1408.5093 (2014)
45. Huynh-Thu, Q., Ghanbari, M.: Scope of validity of psnr in image/video quality assessment. Electronics Letters **44**(13) (June 2008) 800–801
46. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: From error visibility to structural similarity. IEEE Transactions on Image Processing **13**(4) (April 2004) 600–612