International Journal of Automation and Computing
X(X), X X, X-X
DOI: XXX

Deep Learning Based Single Image Super-resolution: A Survey

Viet Khanh Ha[1] Jinchang Ren[2,1] Xinying Xu[2] Sophia Zhao [1] Gang Xie [3] Valentin Masero Vargas [4] Amir Hussain [5]

[1]Department of Electronic and Electrical Engineering, University of Strathclyde, Glasgow, UK
[2]College of Electrical and Power Engineering, Taiyuan University of Technology, Taiyuan, China
[3] College of Electronic Information Engineering, Taiyuan University of Science and Technology, Taiyuan, China
[4] Dept. of Computer Systems and Telematics Engineering, Universidad de Extremadura, Badajoz, Spain
[5] School of Computing, Edinburgh Napier University, Edinburgh, UK

Abstract: Single image super-resolution has attracted increasing attention and has a wide range of applications in satellite imaging, medical imaging, computer vision, security surveillance imaging, remote sensing, objection detection, and recognition. Recently, deep learning techniques have emerged, blossomed, and have produced the-state-of-the-art in many domains. Due to the capability in feature extraction and mapping, it is very helpful to predict the high-frequency details lost in the low-resolution image. In this paper, we give an overview to recent advances on deep learning based models and methods that have been applied for single image super-resolution task. We also summarize, compare and discuss various models from the past, present for comprehensive understanding and finally provide open problems and the possible directions for future research.

Keywords: Image super-resolution, convolutional neural network, high-resolution image, low-resolution image, deep learning.

1 Introduction

Single image super-resolution (SISR) aims to obtain high-resolution (HR) images from a low-resolution (LR) image. It has practical applications in many real-world problems, where certain restrictions present in image or video such as bandwidth, pixel size, scene details, and other factors. Since multiplicity solution exist for a given input LR image, SISR is to solve an ill-posed inverse problem. There are various techniques to solve a SISR problem, which can be classified into three categories, i.e. interpolation-based, reconstruction-based, and example-based methods. The interpolation-based methods are quite straightforward, but they can not provide any additional information for reconstruction and therefore the lost frequency cannot be restored. Reconstruction-based methods usually introduce certain knowledge priors or constraints in an inverse reconstruction problem. The representative priors can be local structure similarity, non-local means, or edge priors. Example-based methods attempt to reconstruct the prior knowledge from a massive amount of internal or external LR-HR patch pairs, in which deep learning techniques have shined new light on SISR.

This survey focuses mainly on deep learning based methods and aims to make a comprehensive introduction to the field of SISR.

The remaining of the paper are organized as follows: Section 2 provides the background and covers different types of example-based SISR algorithms, followed by recent advances in deep learning related models in Section 3. Section 4 compares CNN-based SISR algorithms. Section 5 presents in-depth discussions, followed by open questions for future research in Section 6. Finally, the paper is concluded in Section 7.

Manuscript received date; revised date: 11-4-2019

2 Background

2.1 Early example-based methods

Example-based algorithms aim to enhance the resolution of LR images by learning from other LR-HR patch pair examples. The relationship between LR and HR was applied to un-observed LR image to recover the most likely HR version. According to learning source, example-based methods can be classified into two types: internal learning and external learning based methods.

2.1.1 Internal learning based methods

The natural image has self-similarity property, which tends to recur many times within both the same scale or across different scales inside the image.

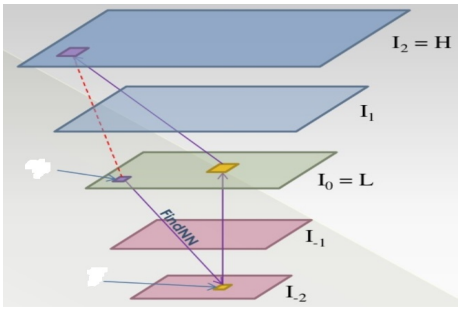

Fig. 1 Pyramid model [1] for SISR. From the bottom, when a similar patch found in a down-scale patch (dark green, dark red), its parent (light green, light red) is copied to unknown HR image with appropriate gap in scale and support of different kernels.

To determine the similarity, Glasner et al [1] made a test by comparing the original image and multiple cascade of images of decreasing resolutions. After that, a scale space pyramid to match LR and HR pairs was proposed as shown in Fig.1 [1]. Since dictionary is limited on the given LR-HR

patch pairs, Huang et al. [2] extended the search space to both planar perspective and affine transform of patches to exploit abundant feature similarity. However, the most important limitation lies in the fact that self-similarity based methods lead to high complexity of computation due to huge numbers of searching and the accuracy of algorithms is highly variant according to natural properties of images.

2.1.2 External learning based methods

The external learning based methods attempt to search the similar information from other images or patches instead. It was first introduced to estimate an underlying scene X with the given image data Y [3]. The algorithm aimed to learn the posterior probability $P(X|Y) = \frac{1}{P(Y)}P(X,Y)$, by adding image patches X and its corresponding scenes Y as nodes in a Markov network. It was then applied for generating super-resolution images, where the input image is LR and the scene to be estimated is replaced by HR image [4].

Locally linear embedding (LLE) is one of the manifold learning algorithms, based on the idea that the high dimensionality may be represented as a function of a few underlying parameters. LLE begins by finding a set of nearest neighbors of each point that can best describe that point as a linear combination of its neighbors. It is then determized to find the low-dimensional embedding of points, such that each point is still represented by the same linear combination of its neighbors. However, one of the disadvantages is that LLE handles non-uniform sample density poorly because the feature represented by the weights varied according to regions in sample densities. The concept of LLE were also applied in SISR neighbor embedding [5], where the features are learned in the LR space before being applied to estimate HR images. There were several other studies based on Locally linear regression such as: ridge regression [6], anchored neighborhood regression [7, 8], random forest [9], and manifold embedding [10].

Another group of algorithms that has received attention is sparsity-based methods. In the sparse representation theory, the data or images can be described as a linear combination of sparse elements chosen from appropriately over-complete dictionary. Let $D \in R^{n\times K}$ be an over-complete dictionary ($K \gg n$), we can to build a dictionary for most scenarios of inputs and then any new image (patch) $X \in R^n$ can be represented as $X = D \times \alpha$, where α is a set of sparse coefficients. Hence, there were dictionary learning problems and sparse coding problems to optimize D and α, respectively. The objective function for standard sparse coding is:

$$\arg\min_{D} \sum_{i=1}^{N} \arg\min_{\alpha_i} \frac{1}{2} \parallel x_i - D\alpha_i \parallel^2 + \lambda\|\alpha_i\| \qquad (1)$$

Unlike standard sparse coding, SISR sparsity-based method works with two dictionaries to learn the compact representation for these patch pairs. Assuming that the observed low-resolution image Y is blurred and a downsampled version of the high-resolution X:

$$Y = S.H.X \qquad (2)$$

where H represents a blurring filter and S the downsampling operation. Under mild conditions, the sparest α_0 can be unique for both dictionaries because the dictionary is over-complete or very large. Hence, the joint sparse coding can be represented as:

$$\arg\min_{D_x,D_y} \sum_{i=1}^{N} \arg\min_{\alpha_i} \frac{1}{2} \parallel x_i - D_x\alpha_i \parallel^2 + \frac{1}{2} \parallel y_i - D_y\alpha_i \parallel^2 + \lambda\|\alpha_i\| \qquad (3)$$

The two dictionaries of high-resolution D_h and low-resolution D_l are co-trained to find the compact coefficients $\alpha_h = \alpha_l = \alpha$ [11], such that sparse representation of high-resolution patch is the same as the sparse representation of the corresponding low-resolution patch. A dictionary D_l was first learned to best fit the LR patches, then the D_h dictionary was learned that worked best with α_l. When these steps were completed, α_l was then used to recover high-resolution image based on high-resolution dictionary D_h.

One of the major drawbacks of this method is that two dictionaries do not always linearly connected. Another problem is that HR images are unknown in the testing phase, hence the equivalence constraint on the HR sparse representation does not guarantee as it has been done in the training phase. Yang et al. [12] suggested a coupled dictionary learning process to pose constraints for two spaces of LR and HR. The main disadvantage of this method is that both dictionaries are assumed to be strictly aligned to achieve alignment between α_h and α_l or simplifying assumption $\alpha_h = \alpha_l$. To relax this invariance assumption, Peleg et al. [13] learn α_h, α_l differently, connecting them via a statistical parametric model. Wang et al. [14] proposed Semi-couple dictionary learning, in which two dictionaries are not fully coupled. It was based on an assumption that there exists an mapping in sparse domain f(.): $\alpha_l \rightarrow \alpha_h$ or $\alpha_h = f(\alpha_l)$. Therefore, the objective function has one additional error term $\|\alpha_h - f(\alpha_l)\|^2$ and other regularization terms. Beta process joint dictionary learning was proposed in [15], which enables to decompose these sparse coefficients to the element multiplication of dictionary atom indicators and coefficient values, providing the much needed flexibility to fit each feature space. Finally, sparsity-based algorithms have remaining limitations in feature extraction and mapping, which are not always adaptive or optimal for generating HR images.

3 Deep Learning related models

3.1 CNNs-based models

The CNNs have been developed rapidly in the last two decades. The first CNNs model to solve the SISR problems is introduced by Dong et al. [16, 17], named Super-Resolution Convolutional Neural Network (SRCNN). Given a training set of LR and corresponding HR images x^i, y^i, i = 1...N, the objective is to find an optimal model f, which will then be applied to accurately predict Y = f(X) on unobserved examples X. The SRCNN [16, 17] consists of the following step, as shown in Fig. 2 [16]:

1. Preprocessing: Upscale the LR image to desired HR image using bicubic interpolation.

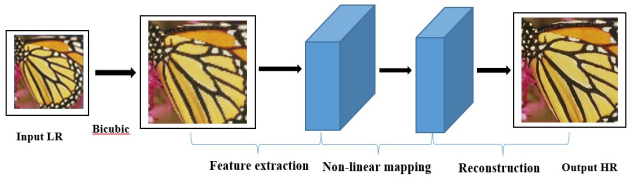

Fig. 2 SRCNN model for SISR

2. Feature extraction: Extract a set of feature map from the upscaled LR image.

3. Non-linear mapping: Maps the features between LR and HR patches.

4. Reconstruction: Produce the HR image from HR patches.

Interestingly, although only three layers have been used, the result significantly outperforms those non-deep learning algorithms discussed previously. However, it seems possible that the accuracy cannot be improved further based on this simple model. This led to observation that whether "the deeper the better" is or not the case in SR. Inspired by the success of very deep networks, Kim et al. [18, 19] proposed two models named Very Deep Convolutional Networks (VDSR) [18] and Deeply Recursive Convolutional Network [19] (DRCN) [18], which both stack 20 convolutional layers, as shown in Fig. 3 (a, b). The VDSR is trained with a very high learning rate (10^{-1} instead of 10^{-4} in SRCNN) in order to accelerate the convergence speed and whilst gradient clipping was used to control the explosion problem.

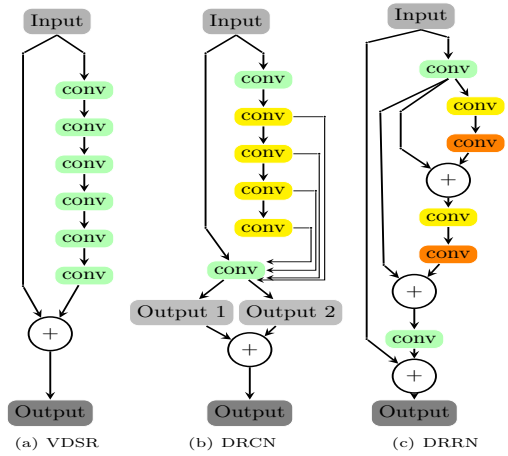

Fig. 3 VDSR, DRCN, DRRN model for SISR. The same color of yellow or orange indicates the sharing parameters.

Instead of predicting the whole image like as did in SRCNN, residual connection was used to force the model to learn the difference between inputs and outputs. The zeros were padding at borders to avoid the problem of quickly reducing feature maps through deep network. In order to gain more benefits from residual learning, Tai et al. [20] used both global residual connection and local residual connection in Deeply Recursive Residual Network (DRRN). The global residual learning is used in the identity branch and recursive learning in local residual branch, as illustrated in Fig. 3 (c) [18]. Mao et al.[21] proposed a 30-layer convolutional auto-encoder network namely very deep Residual Encoder-Decoder Network (RED30). The convolutional layers work as feature extractor, encode image content, while the de-convolutional layers decode and recover image details. Unlike other methods as mentioned above, encoder reduces the feature map to encode the most important features. By doing in this way, noise/corruption can be efficiently eliminated. Hence, this model has made extended test for several tasks of image restoration such as image de-noising, JPEG de-blocking, non-blind de-blurring and image in-painting [21].

Recent advances in CNN architecture such as DenseNet, Network in Network, and Residual Network has been exploited for SISR applications [22, 23]. Among them, RCAN and SRCliqueNet have recently been the-state-of-the-art (up to 2018) in term of pixel-wise measurement, as shown in Table 2, section 4.

Channel attention: Each of the learned filters operates with a local receptive field and the interdependence between channels is entangled with spatial correlation. Therefore, the transformation output is unable to exploit information such as interrelationship between channels outside the region. The RCAN [24] has been the deepest model (about 400 layers) for SISR task. It integrated channel attention mechanism inside the residual block, as shown in Fig. 4 [24]: The input with shape of a H x W x C is squeezed into the channel descriptor by averaging through a spatial dimension of H x W to generate the output shape of 1 x 1 x C. This channel descriptor is put through gate activation of sigmoid f and element-wise product with the input in order to control how much information from each channel is passed up to the next layer in the hierarchy.

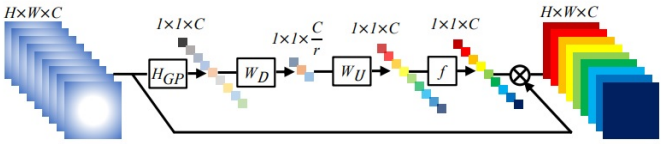

Fig. 4 Channel attention block [24].

Joint sub-band learning with clique structure - SR-CliqueNet [25]: CliqueNet is newly proposed convolutional network architecture where any pair of layers in the same block are connected bilaterally, as shown in Fig. 5. This architecture encourages the features to be refined, which provides more discriminative and leads to a better performance.

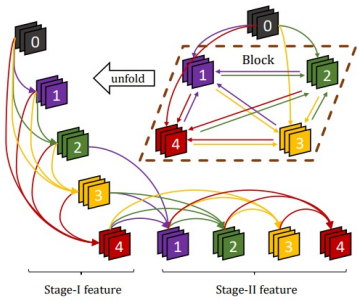

Fig. 5 Clique block with two stages updated. Four layers 1, 2, 3, 4 in blocks are stacked in the order of 1, 2, 3, 4, 1, 2, 3, 4 and bilaterally connected by the residual shortcut. It has more skip connection compared with the Densenet block.

Zhong et al. [25] proposed Super-Resolution CliqueNet, which applied this architecture to jointly learned wavelet sub-band in both the feature extraction stage and sub-band refinement stage.

Concatenation for feature fusion rather than summation- RDN [26]: As model goes deeper, the feature in each layers would be hierarchical with different receptive fields. The information from each layer may not be fully used by recent methods. Zhang et al. [26] proposed concatenated operation on the DenseNet to build hierarchical features from all layers, as shown in Fig. 6 .

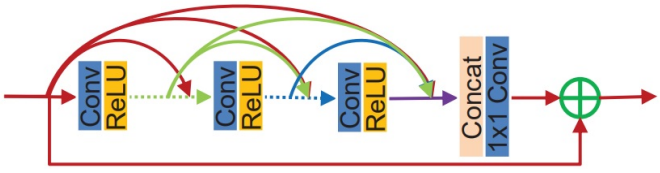

Fig. 6 Residual dense block [26]. All previous feature are concatenated to build hierarchical features.

Wide activation in residual block - WDSR [27]: The efficiency and higher accuracy image resolution can be achieved with less parameters than that of EDSR by expanding the number of channels by a factor of $\sqrt{r}$ before RELU activation in residual block. As such, the residual identity mapping path slimmed as a factor $\sqrt{r}$ to maintain constant output channels.

Cascading Residual to incorporate the features from multiple layers - CARN [28]: The most interesting finding was that there are a similar mechanism in MemNet (Section 3.2), RDN and CARN models. In addition to the ResNet architecture, they all use 1 x 1 convolution as a fusion module to incorporate multiple features from previous layers. Their results boost the performance effectively and can be considered in model design.

Information Distillation Network - IDN [29]: The IDN model uses the distillation block, which combines an enhancement unit with a compression unit. In this block, the information is distilled inside block before pass to next level.

When we use neural network to generate images, it usually involves up-sampling from low resolution to high resolution. One of the problems with the use of interpolation based methods is that it is predefined and there is nothing that the network can learn about. This method is also being criticized for high computational complexity while computing in HR space without additional information. On the other hand, transposed convolution and PixelShuffle concepts has learnable parameters for optimally up-sampling the input. It provides flexible up-sampling and can be inserted at any place in the architecture. Lai et al. [30] proposed Laplacian Pyramid Super-Resolution Network (Lap-SRN) to reconstruct image progressively. In general, the Laplacian Pyramid scheme decomposes an image as a series of high-pass bands and low-pass bands. At each level of reconstruction, a transposed convolution was used to up-sample the image in both high-pass branch and low-pass branch. Beside the Laplace decomposition, Wavelet transform (WT) has been shown to be an efficient and highly intuitive tool to represent and store images in a multi-resolution way. WT can describe the contextual and textural information of an image at different scales. WT for super-resolution has been applied successfully to the multi-frame SR problem. However, conventional discrete wavelet transformation reduces the image size by a factor of 2^n, which is inconvenient when testing images are in certain size. It is proposed by Rohini et al. [31] to reduce the image to any (variable scale) size, using discrete wavelet transformation.

For comparison, most SISR algorithms have been performed on the LR image, which was downsampled with scaling factors of 2x, 3x, 4x from the HR image. Otherwise, features available in the LR space have not sufficed for learning. It is suggested that a training model for high upscaling factor can be benefited from pre-trained model on lower upscaling factor [32]. In other words, it can be described as a transfer learning. Wang et al. [33] proposed a Progressive Asymmetric Pyramidal Structure to adapt with multiple upscaling factors and up to a large scaling factor of 8x. Also, a Deep Back Projection Network [34] using mutually connected up- and down-sampling stages has been used for reaching such high up-scaling factor. These experiments support recommendation to use progressive up-sampling or iterative up and down-sampling when reconstructing SR images under larger scaling factors.

When assuming a low-resolution image is downsampled from the corresponding high-resolution image, CNN-based methods ignored the true degradation such as noise in real world applications. Zhang et al. [35] proposed Super-Resolution Multiple Degradation (SRMD) training on LR images, synthesizing with three kinds of degradations: a blur kernel, bicubicly downsampling followed by Additive White Gaussian noise (AWGN). Obviously, to learn invariant features, this model had to use large training datasets of approximate 6,000 images. Shocher et al. [36] observed strong internal data repetition in the natural images, which is similar to that in [1]. The information for tiny object, for example, is better to be found inside the image, other than in any external database of examples. A "Zero Shot" SR (ZSSR) was then proposed without relying on any prior image examples or prior training. It exploits cross-scale internal recurrence of image-specific information, where the test image itself is trained before fed again to resulting trained network. Because of few research has been focused on variant degradations of SISR, more evaluations and comparisons are required and further investigations would be of great help.

3.2 RNN-CNN-based models

A ResNet with weight sharing can be interpreted as an unrolled single-state Recurrent Neural Network (RNN) [37]. A Dual-State Recurrent Network (DSRN) [38] allows that both the LR path and HR path caption information at different spaces and connected at every step in order to contribute jointly to the learning process, as shown in Fig. 7 [38]. However, the average of all recovered SR images at each stage may have result deteriorated result. Another reason is that the down-sampling operation at every stage can lead to information loss at the final reconstruction layer.

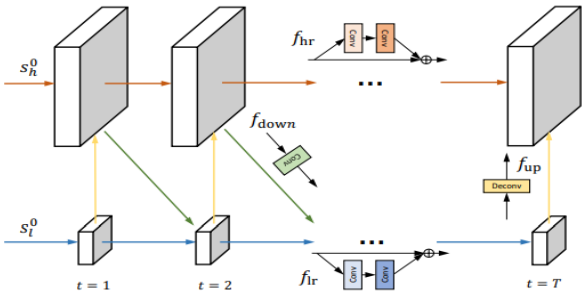

Fig. 7 Dual State Model [38]. The top branch operates on the HR space, where the bottom branch works on the LR space. A connection from LR to HR using de-convolution operation; a delayed feedback mechanism is to connect previous predicted HR to LR at the next stage.

In the view of memory in RNNs, CNNs can be interpreted as: *Short-term memory.* The conventional plain CNNs adopts a single path feed-forward architecture, in which a latter feature influenced by a previous state. *Limited long-term memory:* When the skip connection is introduced, one state is influenced by a previous state and specific point prior state. To enable the latter state can see more prior states and decide whether the information should be kept or discarded, Tai et al. [39] proposed Memory Network (MemNet), which uses recursive layers followed by a memory unit to allow the combination of short and long-term memory for image reconstruction, as shown in Fig. 8 [39]. In this model, a gate unit controls information from the prior

recursive units, which extracts features at different levels.

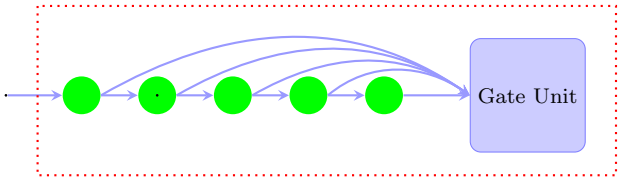

Fig. 8 Memory block in MemNet[39] includes multiple Recursive units and a Gate UnitMemNet Model.

Unlike convolutional operations, which captures features by repeatedly processing local neighborhoods of pixels, the non-local operation describes a pixel as a combination of weighted distance to all other pixels, regardless of their positional distance or channels. Non-local means to provide an efficient procedure for image noise reduction; however, the local and non-local based methods are treated separately, thereby not taking account of their advantages. The non-local block was introduced in [40], enabling integrate non-local operation into end-to-end training with local operation based models such as CNNs. Each pixel at point i in an image can be described as:

$$y_i = \frac{1}{C(x)} \sum_{j \in \Omega} f(x_i, x_j) g(x_j) \tag{4}$$

where $f(x_i, x_j) = e^{\Theta(x_i)^T \varphi(x_j)}$ is a weighted function, measuring how closely related the image at point i is to the image at point j. Thus, by choosing $\Theta(x_i) = W_\Theta x_i$, $\varphi(x_j) = W_\varphi x_j$ and $g(x_j) = W_g x_j$, the self-similarity can be jointly learned in embedding the space by following blocks, as shown in Fig. 9 [40].

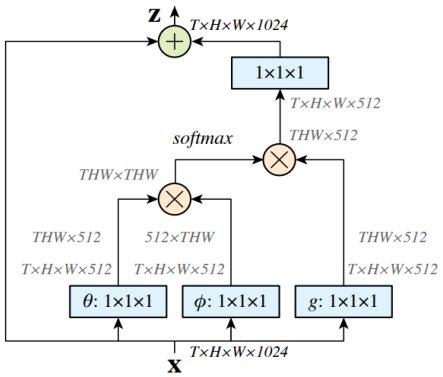

Fig. 9 A non-local block[40].

For SISR tasks, Li et al. [41] incorporated this model into the RNN network by maintaining two paths: a regular path, that contains convolution operation on image, and the other path that maintains non-local information at each step as input branches in the regular RNNs structure. However, non-local means it has disadvantage that remarkable denoising results are obtained at a high expense of computational cost due to the enormous amount of weighting computations.

3.3 GAN-based models

Generative Adversarial Network (GAN) was first introduced in [42], targeting the minimax game between a discriminative network D and a generative network G. The generative network G takes the input z ∼ p(z) as a form of random noise, then outputs new data G(z), whose distribution p_g is supposed to be close to that of the data distribution p_{data}. The task of the discriminative network D is to distinguish a generated sample G(z) ∼ p_g(G(z)) and the ground truth data sample x ∼ p_{data}(x). In other word, the discriminative network determines whether the given images are natural looking images or they look like artificial created images. As the models are trained through alternative optimization, both networks are improved until reach a point called Nash Equilibrium that fake images are indistinguishable from real images. The objective function is represented as:

$$\begin{aligned} &\min_G \max_D E_{x \sim p_{\text{data}}}[\log D(x)] + E_{z \sim p_z}[log(1 - D(G(z)))] \\ &= \min_G \max_D E_{x \sim p_{\text{data}}}[\log D(x)] + E_{x \sim p_z}[log(1 - D(x))] \end{aligned} \tag{5}$$

This concept is consistent with the problem solving in image super resolution. Ledig et al. [43] introduced the Super-Resolution Generative Adversarial Network (SRGAN) model, of which a generative network upsamples LR images to super resolution (SR) images and the discriminative network is to distinguish the ground truth HR images and SR images. Pixel-wise quality assessment metric has been critical of showing poorly to human perception. By incorporating newly adversarial loss, the GAN-based algorithms have solved the problem and produced highly perceptive, naturalistic images, as can be seen from Fig. 10 [43].

The GAN-based SISR model has been developed further in [44, 45], which has resulted in an improved SRGAN by fusion of pixel-wise loss, perceptual loss, and newly proposed texture transfer loss. Park et al. [46] proposed SRFeat, employed an additional discriminator work in feature domain. The generator is trained through two phases: pre-training and adversarial training. In the pre-training phase, the generator is trained to obtain high PSNR by minimizing MSE loss. The training procedure focuses on improving perceptual quality using perceptual similarity loss (section 5.2.2), GAN loss in pixel domain and GAN loss in feature domain. Perhaps the most serious disadvantage of GAN-based SISR methods is difficulties in the training models, which will be further discussed in Section 5.2.

4 Comparison of SISR Algorithms

In order to provide a brief overview of current performance of deep learning-based SISR algorithms, we compare some recent work in Table 1 and Table 2. Two image quality metrics have been used for performance evaluation: A Peak Signal-to-Noise Ratio (PSNR) and a Structural SIMlarity (SSIM) index. The higher the PSNR and SSIM, the better quality of the image being reconstructed.

$$\text{PSNR} = 10 \log_{10} \frac{255^2}{\text{MSE}} \tag{6}$$

where MSE is Mean Squared Error between two images of I_1 and I_2:

$$\text{MSE} = \frac{\sum_{\text{M, N}} [I_1((\text{m, n}) - I_2(\text{m,n})]^2}{\text{M} \times \text{N}} \tag{7}$$

Here, M and N are the number of rows and columns in the input images, respectively. Equation (6) shows that minimizing L_2 loss tends to maximizing the PSNR value.

Table 1 summarizes the detailed performance comparison of some typical deep learning based SISR models, including SRCNN [17], VDSR [18], DRCN [19], DRRN [20], RED30 [21], MemNet [39], EDSR [32], LapSRN [30], Zero Shot [36], IDN [29], CARN [28], RDN [26], SRCliqueNet [25], and RCAN [24]. The detailed performance comparison of those models is presented in Table 2. The four standard benchmark datasets are used including SET5 [47], SET14 [48], B100 [49], URBAN 100 [50] which are popularly used for comparison of SR algorithms. The down-sampling scale factor used include 2x, 3x, and 4x, and missing information that was not provided by the authors is marked by [-]. All quantitative results are duplicated from the original papers.

From Table 1, Table 2 and Fig. 11, CARN stand out through their high accuracy using small model. SRCliqueNet+ and RCAN+ achieved higher accuracy in comparison with EDSR in term of PSNR/MMSI measurement and require smaller model size. GAN-based models are in favour of perceptual reconstruction, which do not include in Table 2 and Fig. 11.

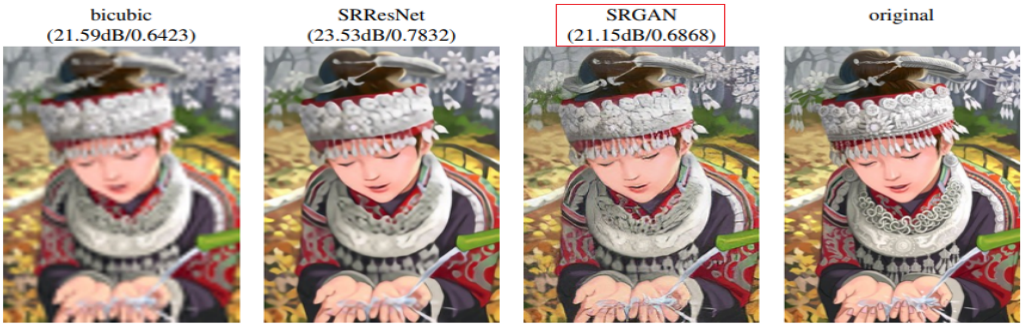

Fig. 10 From left to right, image is reconstructed by bicubic interpolation, deep residual network (SRResNet) measured by MSE, SRGAN optimize more sensitive to human perception, original image. Corresponding PSNR and SSIM are provided on top.

5 Discussion on Optimization Objectives

Generally, when a random variable X has been observed, the attempt is to predict the random variable Y as the output of the network. Let g(X) be the predictor, clearly we would like to choose g so that g(X) tends to be close to Y via the Maximum Likelihood Estimation (MLE). One possible criterion for closeness is to choose g to minimize $E[(Y-g(X))^2]$, thus the optimal predictor of Y becomes $g(X)=E[Y|X]$ as the mean conditional expectation of Y given X. Most of the objective functions originally comes from MLE and we will show that the typical objective functions below are special cases of MLE.

5.1 Content loss

By using CNNs, the mapping between a pair of corresponding LR and HR images is non-linear. The classical content loss function for the regression problem are LAD (Least Absolutes Deviations) (or L_1) and LSE (Least Squared Errors) (or L_2) defined as:

$$L_1=\sum_{i=1}^{n}|\hat{y}-y| \tag{8}$$

$$L_2=\sum_{i=1}^{n}(\hat{y}-y)^2 \tag{9}$$

where the estimation of y can be defined as $y=W^Tx$ and $\hat{y}$ is the ground truth. This objective function is to minimize the cost function regard to the weight matrix W. If we could write regression target as $\hat{y}=y+\xi$ and model regression target as a Gaussian random variable $y\sim N(\mu,\sigma^2)$ with $\mu=y=W^Tx$, the prediction model is:

$$P(\hat{y}|x,W)=N(\hat{y}|W^Tx,\sigma^2)=\frac{1}{\sqrt{2\sigma^2\pi}}\exp(-\frac{(\hat{y}-W^Tx)^2}{2\sigma^2}) \tag{10}$$

then the optimum W can be determined by using the Maximum Likelihood Estimation (MLE):

$$\begin{aligned}W_{\text{MLE}}&=\arg\max_W N(\hat{y}|W^Tx,\sigma^2)\\&=\arg\max_W \exp(-\frac{(\hat{y}-W^Tx)^2}{2\sigma^2})\end{aligned} \tag{11}$$

Optimized with log likelihood and let says $\sigma=1$, we have:

$$W_{\text{MLE}}=\arg\min_W \frac{1}{2}(\hat{y}-W^Tx)^2 \tag{12}$$

which is equal to minimum the loss function L_2 in (9). In other words, Least Square Estimate is actually the same as the Maximum Likelihood Estimate under a Gaussian model. We have replace the L_2 loss function with L_1 loss: $E[(Y-g(X)]$ as mentioned previously, the solution is $g(x)=median(Y|X)$, which is also a solution for MLE. It is important to bear in mind that the assumption is for uni-modal distribution with a single peak, which will not work well to predict multi-modal distributions.

5.2 Perceptual loss

5.2.1 Adversarial loss

A key relationship between images and statistics is that we can interpret images as samples from a high-dimensional probability distribution. The probability distribution goes over the pixels of images and is what we use to define whether an image is natural or not. This is when A Kullback-Leibler Divergence measurement comes into place. It measures the difference between two probability distributions, which is different from the Euclidean distance, , i.e. L_1, L_2 loss. It may be tempting to think as a distance metric, but we cannot use KL Divergence to measure distance between two distributions because it is not symmetric. Given two distribution P_{data} and P_{model}, the forward KL Divergence can be computed as follow:

$$\begin{aligned}D_{KL}[P_{x|data}||P_{x|model}]&=E_{x\sim P_{data}}\log\frac{P_{x|data}}{P_{x|model}}\\&=E_{x\sim P_{data}}[\log P_{x|data}]-E_{x\sim P_{data}}[\log P_{x|model}]\end{aligned} \tag{13}$$

The left term is entropy of $P_{x|data}$ which is dependent on model and thus can be ignored. If we sample N of $x\in P_{x|data}$ when N goes to infinity, following by the law of large numbers we have:

$$-\frac{1}{N}\sum_i^n\log P(x_i|model)=-E_{x\sim P_{x|data}}[P(x|model)] \tag{14}$$

where the right term is negative log-likelihood. The Minimum Kullback-Leibler Divergence is also equivalent to the Maximum the Log Likelihood.

When $P_{model}=P_{data}$ the KL Divergence comes to the minimum 0. It is assumed that human observers have learn p_{data} as a natural distribution or a kind of prior belief. The GAN-based model is to encourage reconstructed images to have similar distribution as the ground truth images, which refer to adversarial loss as part of the perceptual loss in SRGAN [43]. Adversarial learning is actually useful when facing with the complicated manifold distributions in natural images. However, training GANs-based model is elusive due to several drawbacks:

1) *Hard to achieve Nash Equilibrium* [51]: According to game theory, the GANs-based model converges when the discriminator and generator reach a Nash Equilibrium. However, updating each model with no respect to each other cannot guarantee the convergence. Both models can reach a state when the action of each model does not matter to each other.

Table 1 The comparison of different SISR models

Models	Input	Type of network	No of params	MultAdds	Reconstructions	Train data	Loss function
SRCNN	LR + Bicubic	Supervised	8K	52.7G	Direct	Yang91	L2(MSE)
VDSR	LR + Bicubic	Supervised	666K	612G	Direct	G200+Yang91	L2
DRCN	LR + Bicubic	Supervised	1,775K	17,974G	Direct	Yang91	L2
DRRN	LR + Bicubic	Supervised	297K	6,796G	Direct	G200+Yang91	L2
RED30	LR + Bicubic	Supervised	4,2M	-	Direct	BSD300	L2
LapSRN	LR	Supervised	812K	29.9G	Progressive	G200+Yang91	Charbonnie
MemNet	LR + Bicubic	Supervised	677K	2,662G	Direct	G200+Yang91	L2
Zero-Shot	LR + Bicubic	Unsupervised	225K	-	Direct	-	L1(MAE)
Dual State	LR + Bicubic	Supervised	1,2M	-	Progressive	Yang91	L2
SRGAN	LR	Supervised	-	-	Direct	ImageNet	L2 + Perceptual loss
EDSR	LR	Supervised	43M	2890G	Direct	DIV2K	L1
IDN	LR	Supervised	677K	-	Direct	G200+Yang91	L1
CARN	LR	Supervised	1,6M	222G	Direct	DIV2K+Yang91+B200	L1
RDN	LR	Supervised	22.6M	1300G	Direct	DIV2K	L1
SRCliqueNet+	LR	Supervised	-	-	Direct	DIV2K+Flickr	L1 + L2
RCAN+	LR	Supervised	16M	-	Direct	DIV2K	L1

2) *Vanishing problem* [52]*:* As given in (5), when the discriminator learn better we can assume that $D(x) = 1, \forall x \in p_{data}$ and $D(x) = 0, \forall x \in p_{p_z}$ and the loss function falls to 0 and ends up with a vanishing gradient. As a result, the learning is super slow and even jammed. Conversely, when the discriminator behaves badly, the generator does not give accurate feedback.

3) *Mode collapse* [53]*:* a generator generates a limited diversity of samples, or even the same sample regardless the input. We have demonstrated that L1 and L2 loss are special cases of MLE and further KLD is equivalent of MLE. This finding leads to a question whether there exists another effective representation of MLE which is a better representation for image super resolution.

5.2.2 MSE in feature space

The MSE in feature space is to compare two images based on high-level representations from pre-trained Convolutional Neural Networks (trained on Image Classification tasks, for example the ImageNet Dataset).

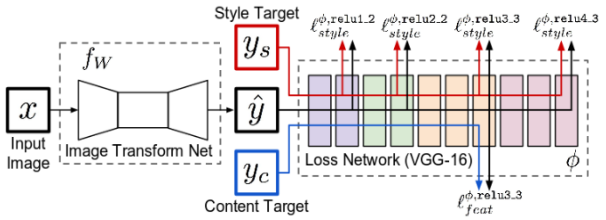

Fig. 12 Model structure for calculating perceptual loss [54]

The image is trained by the Image Transform Net to produce output, where the output is fed to the loss network, which was pre-trained for image classification. The perceptual loss measures perceptual differences in content and style between images. In practice, we can combine different kinds of loss functions, but each loss function mentioned has particular property. There is not a single loss function that works for all kinds of data.

6 Challenges and Trends

Despite of the success of deep learning for SISR tasks, there are opening research questions regarding to SISR model design as discussed below:

1) *Require for light structure model:* Although the deeper is the better, most recent SISR models contain no more than a hundred layers due to the overfitting problem. This is because SISR models work on pixel level, which requires much more parameters than that of image classification. As the model is getting deeper, vanishing gradient is becoming more challenging. This suggests the preference of a light structure model with less parameters and computation.

2) *Adapt well to unknown degradation:* Most algorithms highly depend on predetermined assumption that LR images are simply down-sampling from HR images. They were unsuccessful in recovering SR images with big scale factors due to the lack of learnable features on LR images. If noise is present, the accuracy of reconstruction is deteriorated as the result of the increasing ill-posed problems. A good way to feasibly deal with unknown degradation is to use transfer learning or a huge number of training examples. However, there has been few research on this task hence this needs be further investigated.

3) *Requirement for different assessment criteria:* No methods can achieve low distortion and good perceptual quality at the same time. The traditional measurements such as L1/L2 loss can help to generate images with low distortion, but there are still considerable disagreement with regard to human perception. In contrast, the integration of perceptual assessment produces more realistic images, but it suffers from low PSNR. Therefore, it is necessary to extend more criteria of assessment for particular applications.

4) *Efficiently interpret and exploit prior knowledge to reduce ill-posed problems:* Until recently, the deep architecture appears like a black box and we have limited knowledge of why it works and how it works. We also know a little about image representation for deep networks in term of which space it should be represented. Meanwhile, most SISR algorithms have introduced different structures or connections based on the experiments, neglecting to explain further on why the result is improved. Another important solution for ill-posed problems is to combine different constraints as regulizers for prediction. For example, the combination of different loss functions, or the use of image segmentation information to constraint reconstructed images. That is why semantic categorical prior [55] was introduced, attempting to achieve richer and more realistic textures. The simple ways to use more prior knowledge are that we can use MLE as a proxy to incorporate prior knowledge as conditional probability or feed directly into the network whilst forcing parameters sharing for all kinds of inputs.

Table 2 Quantitative evaluation of the-state-of-the-art SR algorithm. Average PSNR/SSIM for scale factor 2x, 3, 4x. Red text indicates that the best and blue text indicates the second best performance.

	Scale	Set5 PSNR/SSIM	Set14 PSNR/SSIM	B100 PSNR/SSIM	Urban100 PSNR/SSIM
SRCNN	2	36.66/0.9542	32.45/0.9067	-	-
	3	32.75/0.9090	29.30/0.8215	-	-
	4	30.49/0.8628	27.50/0.7513	-	-
VDSR	2	37.53/0.9587	33.03/0.9124	31.90/0.8960	30.76/0.9140
	3	33.66/0.9213	29.77/0.8314	28.82/0.7976	27.14/0.8279
	4	31.35/0.8838	28.01/0.7674	27.29/0.7251	25.18/0.7524
DRCN	2	37.63/0.9588	33.04/0.9118	31.85/0.8942	30.75/0.9133
	3	33.82/0.9226	29.76/0.8311	28.80/0.7963	27.15/0.8276
	4	31.53/0.8854	28.02/0.7670	27.23/0.7233	25.14/0.7510
DRRN	2	37.74/0.9591	33.23/0.9136	32.05/0.8973	31.23/0.9188
	3	34.03/0.9244	29.96/0.8349	28.95/0.8004	27.53/0.8378
	4	31.68/0.888	28.21/0.7720	25.44/07634	25.44/0.7638
RED30	2	37.66/0.9599	32.94/0.9144	-	-
	3	33.82/0.9230	29.61/0.8341	-	-
	4	31.51/0.8869	27.86/0.7718	-	-
MemNet	2	37.78/0.9597	33.28/0.9142	32.08/0.8978	31.31/0.9195
	3	34.09/0.9248	30.00/0.8350	28.96/0.8001	27.56/0.8376
	4	31.74/0.8893	28.26/0.7723	27.40/0.7281	25.50/0.7630
LapSRN	2	37.52/0.959	33.08/0.913	31.80/0.895	30.41/0.910
	4	31.54/0.885	28.19/0.772	27.32/0.728	25.21 / 0.756
	8	26.14/0.738	24.44/0.623	24.54/0.586	21.81/0.581
Zero Shot	2	37.37/0.9570	33.00/0.9108	-	-
	3	33.42/0.9188	29.800.8304	-	-
	4	31.13/0.8796	28.01/0.7651	-	-
EDSR	2	38.20/0.9606	34.02/0.9204	32.37/0.9018	33.10/0.9363
	3	34.77/0.9290	30.66/0.8481	29.32/0.8104	29.02/0.8685
	4	32.62/0.8984	28.94/0.7901	27.79/0.7437	26.86/0.8080
IDN	2	37.83/0.9600	33.30/0.9148	32.08/0.8985	31.27/0.9196
	3	34.11/0.9253	29.99/0.8354	28.95/0.8013	27.42/0.8359
	4	31.82/0.8903	28.25/0.7730	27.41/0.7297	25.41/0.7632
CARN	2	37.76/0.9590	33.52/0.9166	32.09/0.8978	31.92/0.9256
	3	34.29/0.9255	30.29/0.8407	29.06/0.8034	28.06/0.8493
	4	32.13/0.8937	28.60/0.7806	27.58/0.7349	26.07/0.7837
RDN	2	38.30/0.9616	34.10/0.9218	32.40/0.9022	33.09/0.9368
	3	34.78/0.9300	30.67/0.8482	29.33/0.8105	29.00/0.8683
	4	32.61/0.9003	28.92/0.7893	26.82/0.8069	26.82/0.8069
SRCliqueNet+	2	38.28/0.9630	34.03/0.924	32.40/0.906	32.95/0.937
	3	-	-	-	-
	4	32.67/0.903	28.95/0.797	27.81/0.752	26.80/0.810
RCAN+	2	38.27/0.9614	34.23/0.9225	32.46/0.9031	33.54/0.9399
	3	34.85/0.9305	30.76/0.8494	29.39/0.8122	29.31/0.8736
	4	32.73/0.9013	28.98/0.7910	27.85/0.7455	27.10/0.8142

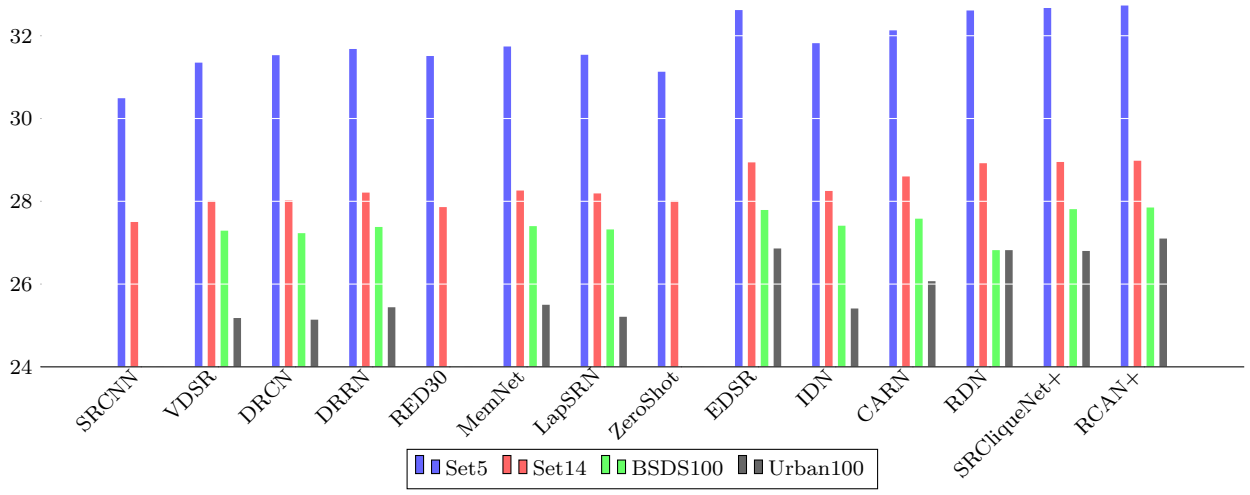

Fig. 11 Comparing the PSNR accuracy of different algorithms on 4 Testing Datasets with factor of 4x.

7 Conclusion

This survey has reviewed most of papers in Single Image Super-Resolution that underlie example-based learning methods. Among them, we noticed that deep learning based methods have recently achieved the state-of-the-art performance. Before going into more detail of each algorithm, the general background in each of the categories was introduced. We have highlighted the important contribution of these algorithms, discussed their pros and cons and suggested future work possible in either within categories or in designated section. Up to present, we cannot define which SISR algorithms is the most state-of-the-art, as this is highly dependent on applications. The algorithm is good for medical imaging or facing processing purposes is not necessarily good for remote sensing images. The different constraints imposed on the problem indicates a need to generate a benchmark database which is specified the concerns of applications in different fields. Finally, there are remaining challenges to bring algorithms into practical applications since they have been applied to standard benchmark datasets and poorly adapt with differently scenarios.

This survey paper has enhanced the understanding of deep learning based algorithms on Single Image Super-Resolution, which can be used as a comprehensive guide for beginner and throw up many questions in need of further investigation.

Acknowledgement

The authors would like to thank the support from the Shanxi Hundred People Plan of China and colleagues from the Image Processing group in Strathclyde University for their valuable suggestions.

References

[1] D. Glasner, S. Bagon, and M. Irani, "Super-resolution from a single image," in *Proceedings of the IEEE Conference on Computer Vision*, pp. 349–356, IEEE, 2009.

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

[3] W. T. Freeman, E. C. Pasztor, and O. T. Carmichael, "Learning low-level vision," *International Journal of Computer Vision*, vol. 40, no. 1, pp. 25–47, 2000.

[4] W. T. Freeman, T. R. Jones, and E. C. Pasztor, "Example-based super-resolution," *IEEE Computer Graphics and Applications*, vol. 22, no. 2, pp. 56–65, 2002.

[5] H. Chang, D.-Y. Yeung, and Y. Xiong, "Super-resolution through neighbor embedding," in *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 1, IEEE, 2004.

[6] C.-Y. Yang and M.-H. Yang, "Fast direct super-resolution by simple functions," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 561–568, 2013.

[7] R. Timofte, V. De Smet, and L. Van Gool, "Anchored neighborhood regression for fast example-based super-resolution," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1920–1927, IEEE, 2013.

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

[9] S. Schulter, C. Leistner, and H. Bischof, "Fast and accurate image upscaling with super-resolution forests," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3791–3799, 2015.

[10] E. Pérez-Pellitero, J. Salvador, J. Ruiz-Hidalgo, and B. Rosenhahn, "Psyco: Manifold span reduction for super resolution," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1837–1845, 2016.

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

[12] J. Yang, Z. Wang, Z. Lin, S. Cohen, and T. Huang, "Coupled dictionary training for image super-resolution," *IEEE transactions on image processing*, vol. 21, no. 8, pp. 3467–3478, 2012.

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

[14] S. Wang, L. Zhang, Y. Liang, and Q. Pan, "Semi-coupled dictionary learning with applications to image super-resolution and photo-sketch synthesis," in *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2216–2223, IEEE, 2012.

[15] L. He, H. Qi, and R. Zaretzki, "Beta process joint dictionary learning for coupled feature spaces with application to single image super-resolution," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 345–352, 2013.

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

[17] C. Dong, C. C. Loy, K. He, and X. Tang, "Image super-resolution using deep convolutional networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 2, pp. 295–307, 2016.

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

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

[20] Y. Tai, J. Yang, and X. Liu, "Image super-resolution via deep recursive residual network," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, vol. 1, p. 5, 2017.

[21] X.-J. Mao, C. Shen, and Y.-B. Yang, "Image restoration using convolutional auto-encoders with symmetric skip connections," *arXiv preprint arXiv:1606.08921*, 2016.

[22] J. Yamanaka, S. Kuwashima, and T. Kurita, "Fast and accurate image super resolution by deep cnn with skip connection and network in network," in *ICONIP*, 2017.

[23] T. Tong, G. Li, X. Liu, and Q. Gao, "Image super-resolution using dense skip connections," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 4809–4817, IEEE, 2017.

[24] Y. Zhang, K. Li, K. Li, L. Wang, B. Zhong, and Y. Fu, "Image super-resolution using very deep residual channel attention networks," in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 286–301, 2018.

[25] Z. Zhong, T. Shen, Y. Yang, Z. Lin, and C. Zhang, "Joint sub-bands learning with clique structures for wavelet domain super-resolution," in *Advances in Neural Information Processing Systems*, pp. 165–175, 2018.

[26] Y. Zhang, Y. Tian, Y. Kong, B. Zhong, and Y. Fu, "Residual dense network for image super-resolution," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2472–2481, 2018.

[27] J. Yu, Y. Fan, J. Yang, N. Xu, Z. Wang, X. Wang, and T. Huang, "Wide activation for efficient and accurate image super-resolution," *arXiv preprint arXiv:1808.08718*, 2018.

[28] N. Ahn, B. Kang, and K.-A. Sohn, "Fast, accurate, and lightweight super-resolution with cascading residual network," in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 252–268, 2018.

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

[30] W.-S. Lai, J.-B. Huang, N. Ahuja, and M.-H. Yang, "Deep laplacian pyramid networks for fast and accurate superresolution," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, vol. 2, p. 5, 2017.

[31] R. S. Asamwar, K. M. Bhurchandi, and A. S. Gandhi, "Interpolation of images using discrete wavelet transform to simulate image resizing as in human vision," *International Journal of Automation and Computing*, vol. 7, no. 1, pp. 9–16, 2010.

[32] B. Lim, S. Son, H. Kim, S. Nah, and K. M. Lee, "Enhanced deep residual networks for single image super-resolution," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, vol. 1, p. 4, 2017.

[33] Y. Wang, F. Perazzi, B. McWilliams, A. Sorkine-Hornung, O. Sorkine-Hornung, and C. Schroers, "A fully progressive approach to single-image super-resolution," *arXiv preprint arXiv:1804.02900*, 2018.

[34] M. Haris, G. Shakhnarovich, and N. Ukita, "Deep back-projection networks for super-resolution," *Computing Research Repository*, vol. abs/1803.02735, 2018.

[35] K. Zhang, W. Zuo, and L. Zhang, "Learning a single convolutional super-resolution network for multiple degradations," *CoRR*, vol. abs/1712.06116, 2017.

[36] A. Shocher, N. Cohen, and M. Irani, "Zero-shot" super-resolution using deep internal learning," *arXiv preprint arXiv:1712.06087*, 2017.

[37] Q. Liao and T. Poggio, "Bridging the gaps between residual learning, recurrent neural networks and visual cortex," *arXiv preprint arXiv:1604.03640*, 2016.

[38] W. Han, S. Chang, D. Liu, M. Yu, M. Witbrock, and T. S. Huang, "Image super-resolution via dual-state recurrent networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[39] Y. Tai, J. Yang, X. Liu, and C. Xu, "Memnet: A persistent memory network for image restoration," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4539–4547, 2017.
[40] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7794–7803, 2018.
[41] D. Liu, B. Wen, Y. Fan, C. C. Loy, and T. S. Huang, "Non-local recurrent network for image restoration," in *Advances in Neural Information Processing Systems*, pp. 1680–1689, 2018.
[42] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in Neural Information Processing Systems*, pp. 2672–2680, 2014.
[43] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. P. Aitken, A. Tejani, J. Totz, Z. Wang, *et al.*, "Photo-realistic single image super-resolution using a generative adversarial network.," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, vol. 2, p. 4, 2017.
[44] M. S. Sajjadi, B. Schölkopf, and M. Hirsch, "Enhancenet: Single image super-resolution through automated texture synthesis," in *Proceedings of the 2017 IEEE International Conference on Computer Vision*, pp. 4501–4510, IEEE, 2017.
[45] J. Johnson, A. Alahi, and F. Li, "Perceptual losses for real-time style transfer and super-resolution," *CoRR*, vol. abs/1603.08155, 2016.
[46] S.-J. Park, H. Son, S. Cho, K.-S. Hong, and S. Lee, "Srfeat: Single image super-resolution with feature discrimination," in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 439–455, 2018.
[47] M. Bevilacqua, A. Roumy, C. Guillemot, and M. L. Alberi-Morel, "Low-complexity single-image super-resolution based on nonnegative neighbor embedding," 2012.
[48] R. Zeyde, M. Elad, and M. Protter, "On single image scale-up using sparse-representations," in *International conference on curves and surfaces*, pp. 711–730, Springer, 2010.
[49] D. Martin, C. Fowlkes, D. Tal, J. Malik, *et al.*, "A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics," Iccv Vancouver:, 2001.
[50] J.-B. Huang, A. Singh, and N. Ahuja, "Single image super-resolution from transformed self-exemplars," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5197–5206, 2015.
[51] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, "Improved techniques for training gans," in *Advances in Neural Information Processing Systems*, pp. 2234–2242, 2016.
[52] M. Arjovsky and L. Bottou, "Towards principled methods for training generative adversarial networks," *arXiv preprint arXiv:1701.04862*, 2017.
[53] L. Metz, B. Poole, D. Pfau, and J. Sohl-Dickstein, "Unrolled generative adversarial networks," *arXiv preprint arXiv:1611.02163*, 2016.
[54] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *Proceedings of the European Conference on Computer Vision*, pp. 694–711, Springer, 2016.
[55] X. Wang, K. Yu, C. Dong, and C. C. Loy, "Recovering realistic texture in image super-resolution by deep spatial feature transform," *arXiv preprint arXiv:1804.02815*, 2018.

Viet Khanh Ha received his B. Eng. degrees in Electrical and Electronics from Le Quy Don University, Viet Nam, 2008, the M. Eng. in Electrical and Electronics at Wollongong University, Australia, 2012. He is currently a Ph. D at the University of Strathclyde, Glasgow, Scotland. His research interests focus mainly on image super resolution using deep learning.

Jinchang Ren received his B. Eng. degrees in computer software, the M. Eng. in image processing, D. Eng. in computer vision all from the Northwestern Polytechnical University (NWPU), Xian, China. Currently, He was also awarded a Ph. D in Electronic Imaging and Media Communication from Bradford University, U.K. Currently he is with CeSIP, University of Strathclyde, Glasgow, U.K.

He has published over 150 peer reviewed journal and conferences papers. His research interests focus mainly on visual computing and multi- media signal processing, especially on semantic content extraction for video analysis and understanding and more recently hyperspectral imaging.

Dr Ren acts as an Associate Editor for two international journals including Multidimensional Systems and Signal Processing (Springer) and Int. J. of Pattern Recognition and Artificial Intelligence.

E-mail: jinchang.ren@strath.ac.uk (Corresponding author)

Xinying Xu is an Professor with the College of Information Engineering, Taiyuan University of Technology.

He has published more than 30 academic papers. His research interests include computational intelligence, data mining, wireless networking, image processing, and fault diagnosis.

Prof Xu is a Member of the Chinese Computer Society, and has been a Visiting Scholar in Department of Computer Science, San Jose State University, CA, USA.

Sophia Zhao is a Research Assistant with the Department of Electronic and Electrical Engineering, University of Strathclyde. Her research interests cover image/signal analysis, machine learning and optimisation.

Gang Xie is a Professor and Vice Principle of Taiyuan University of Science and Technology.

He has published over 80 research papers. His research interests include rough set, intelligent computing, image processing, automation and big data analysis.

Valentin Masero received a B. Eng. degree in Computer Science and Business Administration from University of Extremadura (UEX) in Spain, also a B. Eng. degree in Computer Engineering specialized in Software Development and Artificial Intelligence from University of Granada (UGR). He received the Ph. D. degree in Computer Engineering from UEX. He is an Associate Professor at UEX since 2004.

He focus mainly on Programming, Image Processing, Computer Graphics, Software Development, computer applications in Industrial Engineering, computer applications in Agricultural Engineering and computer applications in Healthcare.

Amir Hussain received his B. Eng. and Ph. D degrees, from the University of Strathclyde, Glasgow, U.K., in 1992 and 1997, respectively. Following postdoctoral and academic positions held at the West of Scotland (1996-98), Dundee (1998-2000) and Stirling Universities (2000-18) respectively, he is currently Professor and founding Head of the Cognitive Big Data and Cybersecurity Research Lab at Edinburgh Napier University, UK.

He has (co)authored more than 350 papers, including over a dozen books and around 150 journal papers.

Prof Hussain is founding Editor-in-Chief of two leading journals: Cognitive Computation (Springer Nature), and BMC Big Data Analytics, and of the Springer Book Series on Socio-Affective Computing, and Cognitive Computation Trends. He is Associate Editor for a number of prestigious journals, includ-

ing: Information Fusion (Elsevier), AI Review (Springer), the IEEE Transactions on Neural Networks and Learning Systems, the IEEE Computational Intelligence Magazine, and the IEEE Transactions on Emerging Topics in Computational Intelligence. Amongst other distinguished roles, he is General Chair for IEEE WCCI 2020 (the world's largest technical event in Computational Intelligence), and Vice-Chair of the Emergent Technologies Technical Committee of the IEEE Computational Intelligence Society.