International Journal of Automation and Computing 16(4), August 2019, 413-426
DOI: 10.1007/s11633-019-1183-x

Deep Learning Based Single Image Super-resolution: A Survey

Viet Khanh Ha[1] Jin-Chang Ren[2,1] Xin-Ying Xu[2] Sophia Zhao[1] Gang Xie[3]
Valentin Masero[4] Amir Hussain[5,6]

[1]Department of Electronic and Electrical Engineering, University of Strathclyde, Glasgow G1 1XW, UK

[2]College of Information Engineering, Taiyuan University of Technology, Taiyuan 030024, China

[3]School of Electronic Information Engineering, Taiyuan University of Science and Technology, Taiyuan 030024, China

[4]Department of Computer Systems and Telematics Engineering, University of Extremadura, Badajoz 06006, Spain

[5]School of Computing, Edinburgh Napier University, Edinburgh EH10 5DT, UK

[6]School of Computer Science and Technology, Anhui University, Anhui 230039, China

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 and blossomed, producing "the state-of-the-art" in many domains. Due to their capability in feature extraction and mapping, it is very helpful to predict high-frequency details lost in low-resolution images. In this paper, we give an overview of recent advances in deep learning-based models and methods that have been applied to single image super-resolution tasks. We also summarize, compare and discuss various models from the past and present for comprehensive understanding and finally provide open problems and 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 multiple solutions exist for a given input LR image, SISR is to solve an ill-posed inverse problem. There are various techniques to solve an 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 provide a comprehensive introduction to the field of SISR.

The remainder of this paper is 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 convolutional neural networks (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.

2 Background

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 an unobserved LR image to recover the most likely HR version. Example-based methods can be classified into two types: internal learning and external learning-based methods.

2.1 Internal learning based methods

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

Review
Manuscript received January 10, 2019; accepted April 19, 2019
Recommended by Associate Editor Bin Luo

To determine the similarity, Glasner et al.[1] made a test by comparing the original image and multiple cascades of images of decreasing resolutions. A scale-space pyramid was constructed to exploit the self-similarity in given LR image, which was then used to impose a set of constraints on the unknown HR image, as shown in Fig. 1[1]. Since the dictionary is limited on the given LR-HR patch pairs, Huang et al.[2] extended the search space to both planar perspectives and affine transforms 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.

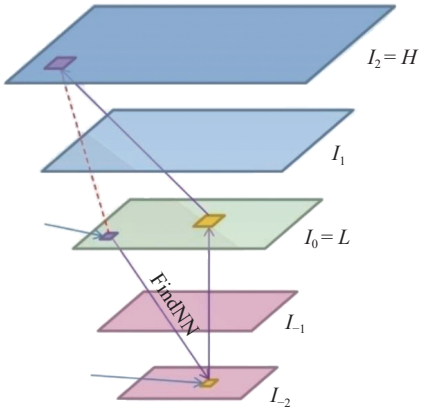


Fig. 1 Pyramid model[1] for SISR. From the bottom, when a similar patch found in a down-scale patch (yellow at level I_{-2}), its parent (yellow at level I_0) is copied to an unknown HR image with an appropriate gap in scale and support of different kernels. Color versions of the figures in this paper are available online.

2.2 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 an 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 determined 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 was 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 local 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 an appropriately over-complete dictionary. Let $D \in \mathbf{R}^{n\times K}$ be an over-complete dictionary $(K \gg n)$, we can build a dictionary for most scenarios of inputs and then any new image (patch) $X \in \mathbf{R}^n$ 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

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

Unlike standard sparse coding, the 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 down-sampled version of the high-resolution X:

$$Y = S \cdot H \cdot X \quad (2)$$

where H represents a blurring filter and S the down-sampling 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

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

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 a high-resolution patch is the same as the sparse representation of the corresponding low-resolution patch. A dictionary D_l was first trained to best fit the LR patches, then the D_h dictionary was trained that worked best with α_l. When these steps were completed, α_l was then used to recover a high-resolution image based on the high-resolution dictionary D_h.

One of the major drawbacks of this method is that the

two dictionaries are 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 the simplifying assumption of $\alpha_h = \alpha_l$. To avoid this invariance assumption, Peleg and Elad[13] connect α_h, α_l 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 a mapping in sparse domain $f(\cdot)$: $\alpha_l \to \alpha_h$ 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 the decomposition of 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 convolutional neural networks (CNNs) have been developed rapidly in the last two decades. The first CNN 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, $i = 1 \cdots N$, 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 consists of the following steps, as shown in Fig. 2[16]:

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

2) Feature extraction: Extract a set of feature maps 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.

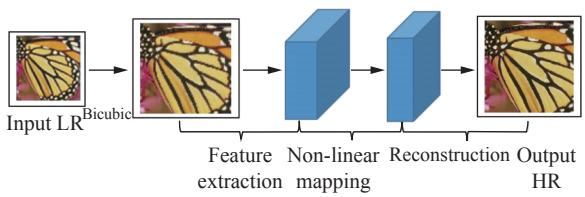


Fig. 2 SRCNN model for SISR

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 the question of whether "the deeper the better" is or is not the case in super resolution (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 (DRCN)[19], which both stack 20 convolutional layers, as shown in Figs. 3 (a) and 3 (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.

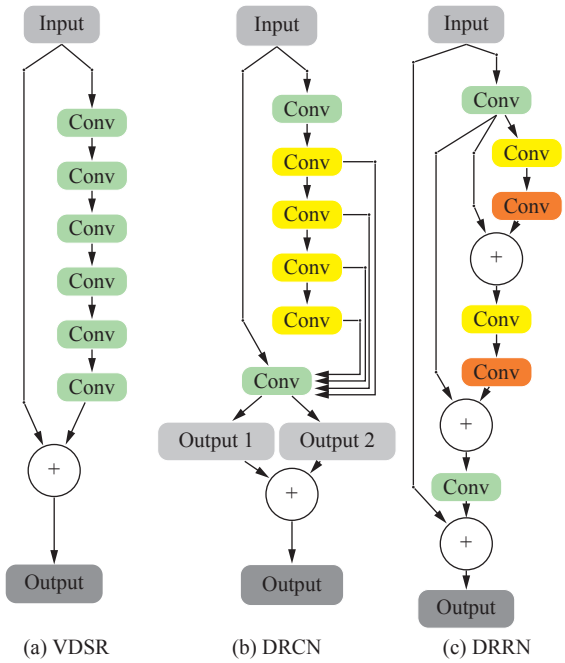


Fig. 3 VDSR, DRCN, DRRN model for SISR. The same color of yellow or orange indicates the sharing parameters.

Instead of predicting the whole image as was done 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 networks. In order to gain more benefits from residual learning, Tai et al.[20] used both global residual connections and local residual connections in deeply recursive residual networks (DRRN). The global residual learning is used in the identity branch and recursive learning in the local residual branch, as illustrated in Fig. 3(c). Mao et al.[21] proposed a 30-layer convolutional auto-encoder network, namely the residual encoder-decoder network (RED30). The convolutional layers work as a feature extractor and encode image content, while the de-convolutional layers decode

and recover image details. Unlike other methods as mentioned above, the encoder reduces the feature map to encode the most important features. By doing it in this way, noise/corruption can be efficiently eliminated. Hence, this model has completed extended tests on several tasks of image restoration such as image de-noising, JPEG de-blocking, non-blind de-blurring and image inpainting[21].

Recent advances in CNN architecture such as DenseNet, Network in Network, and Residual Network have been exploited for SISR applications[22, 23]. Among them, Residual Channel Attention Network (RCAN) and SRCliqueNet have recently been the-state-of-the-art (up to 2018) in terms 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 the interrelationship between channels outside the region. The RCAN[24] has been the deepest model (about 400 layers) for the SISR task. It integrated a channel attention mechanism inside the residual block, as shown in Fig. 4[24]: The input with shape of a $H\times W\times C$ is squeezed into the channel descriptor by averaging through a spatial dimension of $H\times W$ to generate the output shape of $1\times 1\times C$. 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.

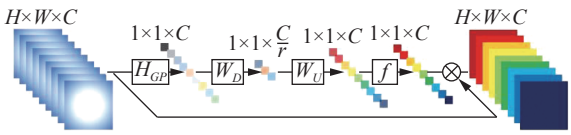


Fig. 4 Channel attention block[24]

Joint sub-band learning with clique structure – SRCliqueNet[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.

The Clique block encourages the features to be refined, which provides more discrimination and leads to a better performance. 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 the model goes deeper, the feature in each layer 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 operations on the DenseNet to build hierarchical features from all layers, as shown in Fig. 6.

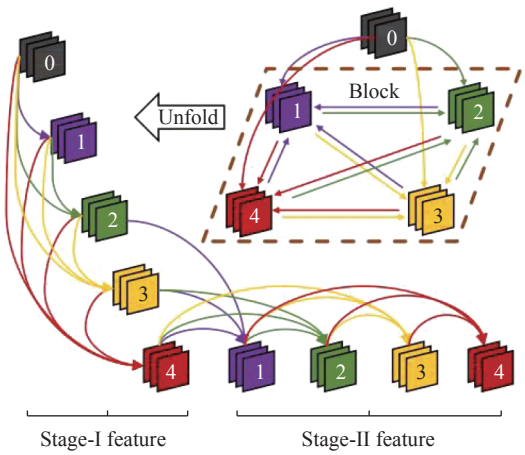


Fig. 5 Clique block with two stages updated. Four layers 1, 2, 3, 4 in blocks are stacked in the order of 1, 2, 3, 4, 1, 2, 3, 4 and bilaterally connected by the residual shortcut. It has more skip connection compared with the Densenet block.

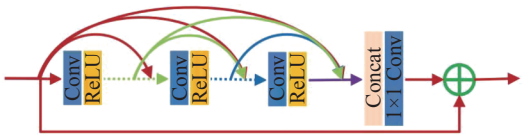


Fig. 6 Residual dense block[26]. All previous feature are concatenated to build hierarchical features.

Wide activation in residual block – Wide-activated deep super-resolution network (WDSR)[27]. The efficiency and higher accuracy image resolution can be achieved with fewer parameters than that of enhanced deep super-resolution network (EDSR) by expanding the number of channels by a factor of $\sqrt{r}$ before rectified linear unit (RELU) activation in residual blocks. As such, the residual identity mapping path slimmed as a factor of $\sqrt{r}$ to maintain constant output channels.

Cascading residuals to incorporate the features from multiple layers – Cascading Residual Network (CARN)[28]. The most interesting finding was that there are similar mechanisms in MemNet (Section 3.2), RDN and CARN models. In addition to the ResNet architecture, they all use 1×1 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 the block before it passes to the next level.

When we use neural networks 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 have 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 networks (Lap-SRN) to reconstruct the 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 the 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 of a certain size. It is proposed by Asamwar 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 factors can benefit from the pre-trained model on lower upscaling factors[32]. In other words, it can be described as 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-sampling and down-sampling stages has been used for reaching such high up-scaling factors. These experiments support recommendations 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 down-sampled 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 6000 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 objects, for example, is better to be found inside the image, other than in any external database of examples. A "Zero Shot" SR (ZSSR)[36] 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 being fed again to the resulting trained network. Because little 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 are 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 a deteriorated result. Another reason is that the down-sampling operation at every stage can lead to information loss at the final reconstruction layer.

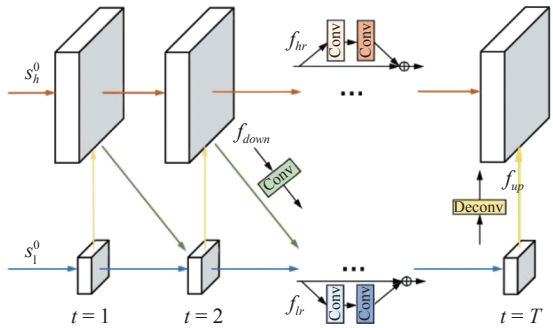


Fig. 7 Dual state model[38]. The top branch operates on the HR space, where the bottom branch works on the LR space. A connection from LR to HR using de-convolution operation; a delayed feedback mechanism is to connect previous predicted HR to LR at the next stage.

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 is 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 to see more prior states and decide whether the information should be kept or discarded, Tai et al.[39] proposed a 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.

Unlike convolutional operations, which capture features by repeatedly processing local neighborhoods of

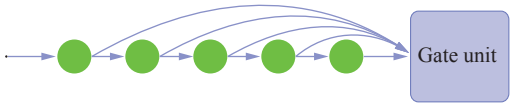


Fig. 8 Memory block in MemNet[39] includes multiple recursive units and a gate unit MemNet model

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 $f(x_i, x_j) = \mathrm{e}^{\Theta(x_i)^{\mathrm{T}} \varphi(x_j)}$ 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].

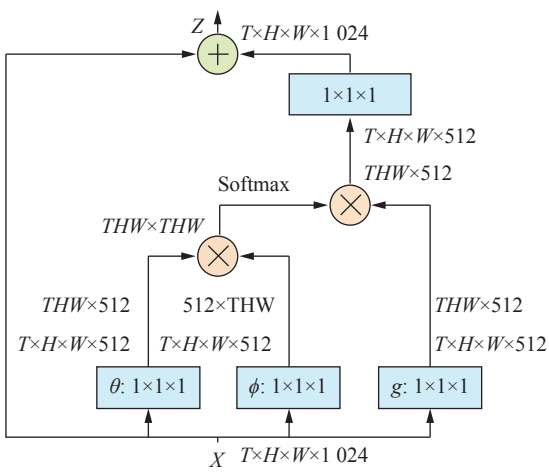


Fig. 9 A non-local block[40]

For SISR tasks, Liu et al.[41] incorporated this model into the RNN network by maintaining two paths: a regular path, that contains convolution operations 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 $z \sim p(z)$ 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 $G(z) \sim p_g(G(z))$ and the ground truth data sample $x \sim p_{\text{data}}(x)$. In other words, 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 they reach a point called Nash Equilibrium that fake images are indistinguishable from real images. The objective function is represented as

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

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. A 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].

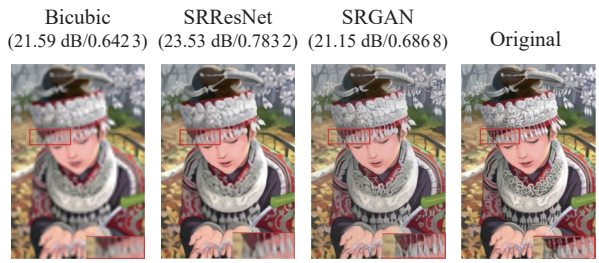


Fig. 10 From left to right, image is reconstructed by bicubic interpolation, deep residual network (SRResNet) measured by MSE, SRGAN optimize more sensitive to human perception, and original image. Corresponding PSNR and SSIM are provided on top. The zoom of red rectangles are shown at right bottom.

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 and employed an additional discriminator

in the 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 the current performance of deep learning-based SISR algorithms, we compare some recent work in Tables 1 and 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. The PSNR can be described as

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

where MSE is mean squared error between two images of I_1 and I_2:

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

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], RCAN[24], SRCliqueNet[25], RDN[26], CARN[28], IDN[29], LapSRN[30], EDSR[32], Zero Shot[36], and MemNet[39]. 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], URBAN100[2] which are popularly used for comparison of SR algorithms. The downsampling 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/SSIM measurement whilst requiring smaller model size. GAN-based models are in favour of perceptual reconstruction, which we do not include in Table 2 and Fig. 11.

5 Discussion on optimization objectives

Generally, when a random variable X has been observed, the aim 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

Table 1 Comparison of different SISR models

Models	Input	Type of network	Number of params	Mult-adds	Reconstructions	Train data	Loss function
SRCNN	LR + Bicubic	Supervised	8 K	52.7 G	Direct	Yang91	L2(MSE)
VDSR	LR + Bicubic	Supervised	666 K	612 G	Direct	G200+Yang91	L2
DRCN	LR + Bicubic	Supervised	1, 775 K	17974 G	Direct	Yang91	L2
DRRN	LR + Bicubic	Supervised	297 K	6796 G	Direct	G200+Yang91	L2
RED30	LR + Bicubic	Supervised	4, 2 M	–	Direct	BSD300	L2
LapSRN	LR	Supervised	812 K	29.9 G	Progressive	G200+Yang91	Charbonnie
MemNet	LR + Bicubic	Supervised	677 K	2662 G	Direct	G200+Yang91	L2
Zero-Shot	LR + Bicubic	Unsupervised	225 K	–	Direct	–	L1(MAE)
Dual State	LR + Bicubic	Supervised	1, 2 M	–	Progressive	Yang91	L2
SRGAN	LR	Supervised	–	–	Direct	ImageNet	L2 + Perceptual loss
EDSR	LR	Supervised	43 M	2890 G	Direct	DIV2K	L1
IDN	LR	Supervised	677 K	–	Direct	G200+Yang91	L1
CARN	LR	Supervised	1, 6 M	222 G	Direct	DIV2K+Yang91+B200	L1
RDN	LR	Supervised	22.6 M	1300 G	Direct	DIV2K	L1
SRCliqueNet+	LR	Supervised	–	–	Direct	DIV2K+Flickr	L1 + L2
RCAN+	LR	Supervised	16 M	–	Direct	DIV2K	L1

Table 2 Quantitative evaluation of the-state-of-the-art SR algorithm. Average PSNR/SSIM for scale factor 2x, 3x, 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	31.68/0.8880	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	37.52/0.9590	33.08/0.9130	31.80/0.8950	30.41/0.9100
	3	–	–	–	–
	4	31.54/0.8850	28.19/0.7720	27.32/0.7280	25.21/0.7560
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	34.03/0.9240	32.40/0.9060	32.95/0.9370
	3	–	–	–	–
	4	32.67/0.9030	28.95/0.7970	27.81/0.7520	26.80/0.8100
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

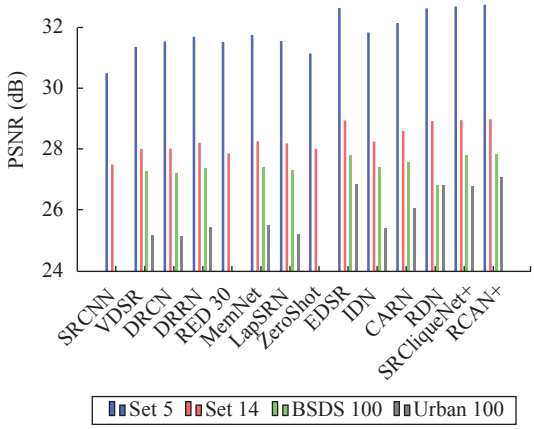


Fig. 11 Comparing the PSNR accuracy of different algorithms on 4 testing datasets with factor of 4x

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

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

then, the optimum W can be determined by using the maximum likelihood estimation (MLE):

$$W_{\mathrm{MLE}} = \arg\max_{W} N(\hat{y}|W^{\mathrm{T}}x, \sigma^2) = \arg\max_{W} \exp\left(-\frac{(\hat{y} - W^{\mathrm{T}}x)^2}{2\sigma^2}\right). \tag{11}$$

Taking the logarithm of the likelihood function, and making use of the standard form ($\sigma = 1$), we obtain the objective function:

$$W_{\mathrm{MLE}} = \arg\min_{W} \frac{1}{2}(\hat{y} - W^{\mathrm{T}}x)^2 \tag{12}$$

which is equal to the 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 to 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. Another problem with content loss is that a minor change in pixels, for example shifting, can lead to a dramatically decreased PSNR. This problem has been mentioned in our previous work[50] with experimental results.

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 (KL) 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 of it 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:

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

The left term is entropy of $P_{x|data}$ which is dependent on the 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 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 distributions to 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 the complicated manifold distributions in natural images. However, training a 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.

2) Vanishing problem[52]: As given in (5), when the discriminator knows 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 of the input. We have demonstrated that L_1 and L_2 loss are special cases of MLE and further KLD is equivalent to 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, e.g., the ImageNet Dataset, as given in Fig. 12).

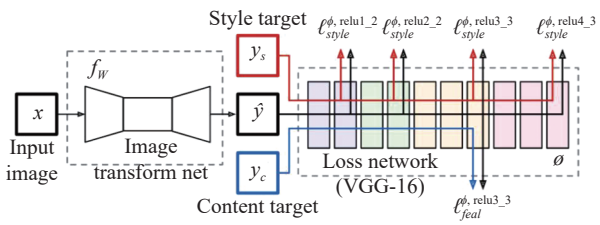


Fig. 12 Model structure for calculating perceptual loss[45]

Given an input image x, Image Transform Net transforms it into the output image $\hat{y}$. Rather than matching the pixels of output image to the pixels of the target image, they were encouraged to have similar feature represents as measured by loss network. The perceptual loss was defined by computing MSE between later set of activations, particularly in applied super-resolution or style transfer. In practice, we can combine different kinds of loss functions, but, each loss function mentioned has a particular property. There is not a single loss function that works for all kinds of data.

6 Challenges and trends

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

1) Need for light structure model: Although deeper is 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 many more parameters than that of image classification. As the model is getting deeper, the vanishing gradient is becoming more challenging. This suggests the preference of a light structure model with fewer parameters and less computation.

2) Adapt well to unknown degradation: Most algorithms highly depend on predetermined assumptions 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 a 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 little 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 is 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. 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 constrain reconstructed images. That is why a semantic categorical prior[54] 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.

7 Conclusions

This survey has reviewed key papers in single image super-resolution that underly example-based learning methods. Among these, we noticed that deep learning based methods have recently achieved 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 important contributions of these algorithms, discussed their pros and cons and suggested future work possible either within categories or in designated sections. Up to now, we cannot define which SISR algorithms are the most state-of-the-art, as this is highly dependent on applications. For instance, an algorithm which is good for medical imaging or facing processing purposes is not necessarily effective for remote sensing images. The different constraints imposed in a problem indicates a need to generate a benchmark database that specifies the concerns of applications in different fields. Finally, there are outstanding challenges to exploit algorithms in practical applications since they have been mainly applied to standard benchmark datasets and poorly adapted to different scenarios. This survey paper has enhanced the understanding of deep learning based algorithms applied to single image super-resolution, which can be used as a comprehensive guide for the beginner and throws up many questions in need of further investigation.

Acknowledgements

The authors would like acknowledge the support from the Shanxi Hundred People Plan of China and colleagues from the Image Processing Group in Strathclyde University (UK) , Anhui University (China) and Taibah Valley (Taibah University, Saudi Arabia) respectively, for their valuable suggestions.

References

[1] D. Glasner, S. Bagon, M. Irani. Super-resolution from a single image. In *Proceedings of the 12th International Conference on Computer Vision*, IEEE, Kyoto, Japan, pp. 349–356, 2009. DOI: 10.1109/ICCV.2009.5459271.

[2] J. B. Huang, A. Singh, N. Ahuja. Single image super-resolution from transformed self-exemplars. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, Boston, USA, pp. 5197–5206, 2015. DOI: 10.1109/CVPR.2015.7299156.

[3] W. T. Freeman, E. C. Pasztor, O. T. Carmichael. Learning low-level vision. *International Journal of Computer Vision*, vol. 40, no. 1, pp. 25–47, 2000. DOI: 10.1023/A:1026501619075.

[4] W. T. Freeman, T. R. Jones, E. C. Pasztor. Example-based super-resolution. *IEEE Computer Graphics and Applications*, vol. 22, no. 2, pp. 56–65, 2002. DOI: 10.1109/38.988747.

[5] H. Chang, D. Y. Yeung, Y. M. Xiong. Super-resolution through neighbor embedding. In *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Washington, USA, 2004. DOI: 10.1109/CVPR.2004.1315043.

[6] C. Y. Yang, M. H. Yang. Fast direct super-resolution by simple functions. In *Proceedings of IEEE International Conference on Computer Vision*, Sydney, Australia, pp. 561–568, 2013. DOI: 10.1109/ICCV.2013.75.

[7] R. Timofte, V. De Smet, L. Van Gool. Anchored neighborhood regression for fast example-based super-resolution. In *Proceedings of IEEE International Conference on Computer Vision*, Sydney, Australia, pp. 1920–1927, 2013. DOI: 10.1109/ICCV.2013.241.

[8] R. Timofte, V. De Smet, L. Van Gool. A+: Adjusted anchored neighborhood regression for fast super-resolution. In *Proceedings of the 12th Asian Conference on Computer Vision*, Springer, Singapore, pp. 111–126, 2015. DOI: 10.1007/978-3-319-16817-3_8.

[9] S. Schulter, C. Leistner, H. Bischof. Fast and accurate image upscaling with super-resolution forests. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, Boston, USA, pp. 3791–3799, 2015. DOI: 10.1109/CVPR.2015.7299003.

[10] E. Pérez-Pellitero, J. Salvador, J. Ruiz-Hidalgo, B. Rosenhahn. PSyCo: Manifold span reduction for super resolution. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, USA, pp. 1837–1845, 2016. DOI: 10.1109/CVPR.2016.203.

[11] J. C. Yang, J. Wright, T. S. Huang, Y. Ma. Image super-resolution via sparse representation. *IEEE Transactions on Image Processing*, vol. 19, no. 11, pp. 2861–2873, 2010. DOI: 10.1109/TIP.2010.2050625.

[12] J. C. Yang, Z. W. Wang, Z. Lin, S. Cohen, T. Huang. Coupled dictionary training for image super-resolution. *IEEE Transactions on Image Processing*, vol. 21, no. 8, pp. 3467–3478, 2012. DOI: 10.1109/TIP.2012.2192127.

[13] T. Peleg, 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. DOI: 10.1109/TIP.2014.2305844.

[14] S. L. Wang, L. Zhang, Y. Liang, Q. Pan. Semi-coupled dictionary learning with applications to image super-resolution and photo-sketch synthesis. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, Providence, USA, pp. 2216–2223, 2012. DOI: 10.1109/CVPR.2012.6247930.

[15] L. He, H. R. Qi, R. Zaretzki. Beta process joint dictionary learning for coupled feature spaces with application to single image super-resolution. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, Portland, USA, pp. 345–352, 2013. DOI: 10.1109/CVPR.2013.51.

[16] C. Dong, C. C. Loy, K. M. He, X. O. Tang. Learning a deep convolutional network for image super-resolution. In *Proceedings of the 13th European Conference on Computer Vision*, Springer, Zurich, Switzerland, pp. 184–199, 2014. DOI: 10.1007/978-3-319-10593-2_13.

[17] C. Dong, C. C. Loy, K. M. He, X. O. Tang. Image super-resolution using deep convolutional networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 2, pp. 295–307, 2016. DOI: 10.1109/TPAMI.2015.2439281.

[18] J. Kim, J. Kwon Lee, K. Mu Lee. Accurate image super-resolution using very deep convolutional networks. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, USA, pp. 1646–1654,

2016. DOI: 10.1109/CVPR.2016.182.

[19] J. Kim, J. Kwon Lee, K. Mu Lee. Deeply-recursive convolutional network for image super-resolution. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, USA, pp. 1637–1645, 2016. DOI: 10.1109/CVPR.2016.181.

[20] Y. Tai, J. Yang, X. M. Liu. Image super-resolution via deep recursive residual network. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, USA, vol. 1, pp. 2790–2798, 2017. DOI: 10.1109/CVPR.2017.298.

[21] X. J. Mao, C. H. Shen, Y. B. Yang. Image Restoration Using Convolutional Auto-encoders with Symmetric Skip Connections, [Online], Available: https://arxiv.org/abs/1606.08921, May, 2018.

[22] J. Yamanaka, S. Kuwashima, T. Kurita. Fast and accurate image super resolution by deep CNN with skip connection and network in network. In *Proceedings of the 24th International Conference on Neural Information Processing*, Springer, Guangzhou, China, 2017. DOI: 10.1007/978-3-319-70096-0_23.

[23] T. Tong, G. Li, X. J. Liu, Q. Q. Gao. Image super-resolution using dense skip connections. In *Proceedings of IEEE International Conference on Computer Vision*, IEEE, Venice, Italy, pp. 4809–4817, 2017. DOI: 10.1109/ICCV.2017.514.

[24] Y. L. Zhang, K. P. Li, K. Li, L. C. Wang, B. N. Zhong, Y. Fu. Image super-resolution using very deep residual channel attention networks. In *Proceedings of the 15th European Conference on Computer Vision*, Springer, Munich, Germany, pp. 286–301, 2018. DOI: 10.1007/978-3-030-01234-2_18.

[25] Z. S. Zhong, T. C. Shen, Y. B. Yang, Z. C. Lin, C. Zhang. Joint sub-bands learning with clique structures for wavelet domain super-resolution. In *Proceedings of the 32nd Conference on Neural Information Processing Systems*, Curran Associates, Inc., Montréal, Canada, pp. 165–175, 2018.

[26] Y. L. Zhang, Y. P. Tian, Y. Kong, B. N. Zhong, Y. Fu. Residual dense network for image super-resolution. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, IEEE, Salt Lake City, USA, pp. 2472–2481, 2018. DOI: 10.1109/CVPR.2018.00262.

[27] J. H. Yu, Y. C. Fan, J. C. Yang, N. Xu, Z. W. Wang, X. C. Wang, T. Huang. Wide Activation for Efficient and Accurate Image Super-resolution, [Online], Available: https://arxiv.org/abs/1808.08718v1, April 8, 2019.

[28] N. Ahn, B. Kang, K. A. Sohn. Fast, accurate, and lightweight super-resolution with cascading residual network. In *Proceedings of the 15th European Conference on Computer Vision*, Springer, Munich, Germany, pp. 252–268, 2018. DOI: 10.1007/978-3-030-01249-6_16.

[29] Z. Hui, X. M. Wang, X. B. Gao. Fast and accurate single image super-resolution via information distillation network. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, IEEE, Salt Lake City, USA, pp. 723–731, 2018. DOI: 10.1109/CVPR.2018.00082.

[30] W. S. Lai, J. B. Huang, N. Ahuja, M. H. Yang. Deep laplacian pyramid networks for fast and accurate super-resolution. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, Honolulu, USA, vol. 2, pp. 5835-5843, 2017. DOI: 10.1109/CVPR.2017.618.

[31] R. S. Asamwar, K. M. Bhurchandi, 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. DOI: 10.1007/s11633-010-0009-7.

[32] B. Lim, S. Son, H. Kim, S. Nah, K. M. Lee. Enhanced deep residual networks for single image super-resolution. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition Workshops*, IEEE, Honolulu, USA, vol. 1, pp. 1132–1140, 2017. DOI: 10.1109/CVPRW.2017.151.

[33] Y. F. Wang, F. Perazzi, B. McWilliams, A. Sorkine-Hornung, O. Sorkine-Hornung, C. Schroers. A fully progressive approach to single-image super-resolution. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, IEEE, Salt Lake City, USA, 2018. DOI: 10.1109/CVPRW.2018.00131.

[34] M. Haris, G. Shakhnarovich, N. Ukita. Deep back-projection networks for super-resolution. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, IEEE, Salt Lake City, USA, pp. 1664–1673, 2018. DOI: 10.1109/CVPR.2018.00179.

[35] K. Zhang, W. M. Zuo, L. Zhang. Learning a single convolutional super-resolution network for multiple degradations. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, IEEE, Salt Lake City, USA, pp. 3262–3271, 2018. DOI: 10.1109/CVPR.2018.00344.

[36] A. Shocher, N. Cohen, M. Irani. Zero-shot super-resolution using deep internal learning. In *Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, IEEE, Salt Lake City, USA, pp. 3118-3126, 2018. DOI: 10.1109/CVPR.2018.00329.

[37] Q. L. Liao, T. Poggio. Bridging the Gaps Between Residual Learning, Recurrent Neural Networks and Visual Cortex, [Online], Available: https://arxiv.org/abs/1604.03640, July 10, 2018.

[38] W. Han, S. Y. Chang, D. Liu, M. Yu, M. Witbrock, T. S. Huang. Image super-resolution via dual-state recurrent networks. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, IEEE, Salt Lake City, USA, pp. 1654–1663, 2018. DOI: 10.1109/CVPR.2018.00178.

[39] Y. Tai, J. Yang, X. M. Liu, C. Y. Xu. MemNet: A persistent memory network for image restoration. In *Proceedings of IEEE International Conference on Computer Vision*, IEEE, Venice, Italy, pp. 4539–4547, 2017. DOI: 10.1109/ICCV.2017.486.

[40] X. L. Wang, R. Girshick, A. Gupta, K. M. He. Non-local neural networks. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, IEEE, Salt Lake City, USA, pp. 7794–7803, 2018. DOI: 10.1109/CVPR.2018.00813.

[41] D. Liu, B. H. Wen, Y. C. Fan, C. C. Loy, T. S. Huang. Non-local recurrent network for image restoration. In *Proceedings of the 32nd Conference on Neural Information Processing Systems*, Curran Associates, Inc., Montréal, Canada, pp. 1680–1689, 2018.

[42] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio. Generative adversarial nets. In *Proceedings of the 27th International Conference on Neural Information Processing Systems*, MIT Press, Montreal, Canada, 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. H. Wang, W. Z. Shi. Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of IEEE Conference on Computer Vision and Pattern*

Recognition, IEEE, Honolulu, USA, vol. 2, pp. 105–114, 2017. DOI: 10.1109/CVPR.2017.19.

[44] M. S. Sajjadi, B. Schölkopf, M. Hirsch. EnhanceNet: Single image super-resolution through automated texture synthesis. In *Proceedings of IEEE International Conference on Computer Vision*, IEEE, Venice, Italy, pp. 4501–4510, 2017. DOI: 10.1109/ICCV.2017.481.

[45] J. Johnson, A. Alahi, F. F. Li. Perceptual losses for real-time style transfer and super-resolution. In *Proceedings of the 14th European Conference on Computer Vision*, Springer, Amsterdam, The Netherlands, pp. 694–711, 2016. DOI: 10.1007/978-3-319-46475-6_43.

[46] S. J. Park, H. Son, S. Cho, K. S. Hong, S. Lee. SRFeat: Single image super-resolution with feature discrimination. In *Proceedings of the 15th European Conference on Computer Vision*, Springer, Munich, Germany, pp. 439–455, 2018. DOI: 10.1007/978-3-030-01270-0_27.

[47] M. Bevilacqua, A. Roumy, C. Guillemot, M. L. Alberi-Morel. Low-complexity single-image super-resolution based on nonnegative neighbor embedding. In *Proceedings of British Machine Vision Conference*, BMVA Press, Surrey, UK, 2012.

[48] R. Zeyde, M. Elad, M. Protter. On single image scale-up using sparse-representations. In *Proceedings of the 7th International Conference on Curves and Surfaces*, Springer, Avignon, France, pp. 711–730, 2010. DOI: 10.1007/978-3-642-27413-8_47.

[49] D. Martin, C. Fowlkes, D. Tal, J. Malik. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In *Proceedings the 8th IEEE International Conference on Computer Vision*, IEEE, Vancouver, Canada, 2001. DOI: 10.1109/ICCV.2001.937655.

[50] V. K. Ha, J. C. Ren, X. Y. Xu, S. Zhao, G. Xie, V. M. Vargas. Deep learning based single image super-resolution: A survey. In *Proceedings of the 9th International Conference on Brain Inspired Cognitive Systems*, Springer, Xi′an, China, pp. 106–119, 2018. DOI: 10.1007/978-3-030-00563-4_11.

[51] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, X. Chen. Improved techniques for training GANs. In *Proceedings of the 30th Conference on Neural Information Processing Systems*, Curran Associates, Inc., Barcelona, Spain, pp. 2234–2242, 2016.

[52] M. Arjovsky, L. Bottou. Towards Principled Methods for Training Generative Adversarial Networks, [Online], Available: https://arxiv.org/abs/1701.04862, April 8, 2018.

[53] L. Metz, B. Poole, D. Pfau, J. Sohl-Dickstein. Unrolled Generative Adversarial Networks, [Online], Available: https://arxiv.org/abs/1611.02163, June 10–20, 2018.

[54] X. T. Wang, K. Yu, C. Dong, C. C. Loy. Recovering Realistic Texture in Image Super-resolution by Deep Spatial Feature Transform, [Online], Available: https://arxiv.org/abs/1804.02815, October, 2018.

Viet Khanh Ha received the B. Eng. degrees in electrical and electronics from Le Quy Don University, Viet Nam in 2008, the M. Eng. degree in electrical and electronics from Wollongong University, Australia in 2012. He is currently a Ph. D. degree candidate at the University of Strathclyde, UK.

His research interests include image super resolution using deep learning.

E-mail: ha-viet-khanh@strath.ac.uk

ORCID iD: 0000-0002-6965-4024

Jin-Chang Ren received the B. Eng. degree in computer software in 1992, the M. Eng. degree in image processing in 1997, the Ph. D. degree in computer vision in 2000, all from the North-western Polytechnical University (NWPU), China. He was also awarded a Ph. D. in electronic imaging and media communication from Bradford University, UK in 2009. Currently, he is with Centre for Signal and Image Processing (CeSIP), University of Strathclyde, UK. He has published over 150 peer reviewed journals and conferences papers. He acts as an associate editor for two international journals including *Multidimensional Systems and Signal Processing* and *International Journal of Pattern Recognition* and *Artificial Intelligence*.

His research interests focus mainly on visual computing and multi-media signal processing, especially on semantic content extraction for video analysis and understanding more recently hyperspectral imaging.

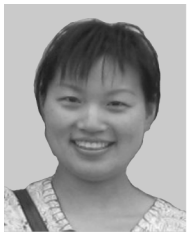
E-mail: jinchang.ren@strath.ac.uk (Corresponding author)

ORCID iD: 0000-0001-6116-3194

Xin-Ying Xu received the B. Sc. and Ph. D. degrees from the Taiyuan University of Technology, China, in 2002 and 2009, respectively. He is currently a professor with the College of Information Engineering, Taiyuan University of Technology, China. He has published more than 30 academic papers. He is a member of the Chinese Computer Society, and has been a visiting scholar in Department of Computer Science, San Jose State University, USA.

His research interests include computational intelligence, data mining, wireless networking, image processing, and fault diagnosis.

E-mail: xuxinying@tyut.edu.cn



Sophia Zhao received the B. Sc. degree in education from Henan University, China in 1999, and several qualifications from Shipley College, UK during 2003–2005. Currently, she is a research assistant with the Department of Electronic and Electrical Engineering, University of Strathclyde, UK.

Her research interests include image/signal analysis, machine learning and optimisation.

E-mail: sophia.zhao@strath.ac.uk

Gang Xie received the B. S. degree in control theory and the Ph. D. degree in circuits and systems from the Taiyuan University of Technology, China, in 1994 and 2006, respectively. He has been a professor and vice principle of Taiyuan University of Science and Technology, China. He has published over 80 research papers.

His research interests include rough sets, intelligent computing, image processing, automation and big data analysis.

E-mail: xiegang@tyut.edu.cn

Valentin Masero received the B. Eng. degree in computer science and business administration from University of Extremadura (UEX), Spain, and another B. Eng. degree in computer engineering specialized in software development and artificial intelligence from University of Granada, Spain. He received the Ph. D. degree in computer engineering from UEX, Spain. Now he is an associate professor at UEX.

His research interests include image processing, machine learning, artificial intelligence, computer graphics, computer programming, software development, computer applications in industrial engineering, computer applications in agricultural engineering and computer applications in healthcare.

E-mail: vmasero@unex.es

Amir Hussain received the B. Eng. and Ph. D. degrees from the University of Strathclyde in Glasgow, UK, in 1992 and 1997, respectively. Following postdoctoral and academic positions at the Universities of West of Scotland (1996–1998), Dundee (1998–2000) and Stirling (2000–2018), respectively, he joined Edinburgh Napier University (UK) in 2018, as founding director of the Cognitive Big Data and Cybersecurity (CogBiD) Research Laboratory, managing over 25 academic and research staffs. He has been appointed to invited visiting professorships at several Universities and Research and Innovation Centres, including at Anhui University (China) and Taibah Valley (Taibah University, Saudi Arabia). He has (co)authored three international patents, around 400 publications, including over a dozen books and 150 journal papers. He has led major multi-disciplinary research projects, funded by national and European research councils, local and international charities and industry, and supervised more than 35 Ph. D. students. He is founding Editor-in-Chief of (Springer Nature′s) *Cognitive Computation* journal and *BMC Big Data Analytics* journal. He has been appointed Associate Editor of several other world-leading journals including, *IEEE Transactions on Neural Networks and Learning Systems*, (Elsevier′s) *Information Fusion* journal, *IEEE Transactions on Emerging Topics in Computational Intelligence*, and *IEEE Computational Intelligence Magazine*. Amongst other distinguished roles, he is General Chair for IEEE WCCI 2020 (the world′s largest and top IEEE technical event in computational intelligence, comprising IJCNN, FUZZ-IEEE and IEEE CEC), Vice-Chair of Emergent Technologies Technical Committee of the IEEE Computational Intelligence Society, and chapter Chair of the IEEE UK & Ireland, Industry Applications Society Chapter.

His research interests include developing cognitive data science and AI technologies, to engineer the smart and secure systems of tomorrow.

E-mail: A.Hussain@napier.ac.uk