IET Image Processing

*Review Article*

IET Journals
The Institution of Engineering and Technology

# Survey of single image super-resolution reconstruction

ISSN 1751-9659
Received on 16th November 2019
Revised 28th February 2020
Accepted on 24th March 2020
E-First on 23rd July 2020
doi: 10.1049/iet-ipr.2019.1438
www.ietdl.org

*Kai Li[1,2], Shenghao Yang[1,2], Runting Dong[1], Xiaoying Wang[1], Jianqiang Huang[1,2]* ✉

[1]*State Key Laboratory of Plateau Ecology and Agriculture, Department of Computer Technology and Application, Qinghai University, Xining, Qinghai 810016, People's Republic of China*

[2]*Department of Computer Science and Technology, Tsinghua University, Beijing 100084, People's Republic of China*

✉ *E-mail: hjq16@mails.tsinghua.edu.cn*

**Abstract:** Image super-resolution reconstruction refers to a technique of recovering a high-resolution (HR) image (or multiple images) from a low-resolution (LR) degraded image (or multiple images). Due to the breakthrough progress in deep learning in other computer vision tasks, people try to introduce deep neural network and solve the problem of image super-resolution reconstruction by constructing a deep-level network for end-to-end training. The currently used deep learning models can divide the SISR model into four types: interpolation-based preprocessing-based model, original image processing based model, hierarchical feature-based model, and high-frequency detail-based model, or shared the network model. The current challenges for super-resolution reconstruction are mainly reflected in the actual application process, such as encountering an unknown scaling factor, losing paired LR–HR images, and so on.

## 1 Introduction

Before the advent of deep learning-based methods, image super-resolution reconstruction mostly used methods based on interpolation and regularisation. Image interpolation is an imaging method that increases the number of image pixels. It is a process of estimating the pixel value of a particular position between image pixels [1]. Image interpolation generates high-resolution (HR) images by up-sampling low-resolution (LR) images [2]. On the other hand, since image super-resolution reconstruction is an ill-posed problem, regularisation is widely used as a method to solve the ill-posed problem [3]. In recent years, the application of image super-resolution reconstruction in various fields has received increasing attention, and many models based on deep neural network have achieved excellent results [4]. Since super-resolution convolutional neural network (SRCNN) was put forward, many convolutional neural network (CNN)-based network models have been born. At the same time, due to the proposed residual network, the deep model can be effectively prevented from disappearing [5], so that the deeper CNN model can improve the performance of super-resolution better. Dual-state recurrent network (DSRN) [6] is an improvement based on recursive neural network (RNN)). DSRN can perform super-resolution reconstruction at different spatial resolutions, with higher performance than DenseNet and ResNet. Seif and Androutsos[7] proposed to use the one-dimensional separable filter and Atrous conv to achieve the performance of conventional conv for parameters that have been parameterised in deep network, achieving the same performance using fewer conv layers. EDSR [8] removes the batch normalisation layer in the residual block based on SRResNet [9], saving memory resources to stack more network layers under the same computing resources or extract more features in each layer, so that the quality of the image is improved. Wide deep super resolution (WDSR) [10] removes most of the redundant convolutional layers based on EDSR, which reduces memory and increases computational speed. Zhang *et al*. [11] proposed a new residual-dense network (RDN), which uses residual-dense blocks (RDBs) to connect to my richer local features. Cascading residual network (CARN) [12] is a cascaded residual network that achieves better performance with fewer parameters and operands. RefSR [13] proposed an end-to-end CNN based on the idea of 'warping + synthesis', using MDSR as a sub-module for LR image feature extraction and RefSR synthesis. Zhang *et al*. [14] proposed a very deep residual channel attention network (RCAN) to solve the difficulty of training deep network, which can adaptively re-adjust the characteristics of the channel mode by considering the interdependencies between channels.

For the first time, super-resolution generative adversarial network (SRGAN) [9] applied generative adversarial network (GAN) to the field of super-resolution reconstruction and proposed a perceptual loss function instead of the mean square error (MSE) loss function. The $4\times$ down-sampled super-resolution images reconstructed has been achieved, and the obtained images have more detailed details and texture. Enhanced SRGAN (ESRGAN) removed the BN layer based on SRGAN and eliminated artefacts in the original model results. Wang *et al*. [15] proposed a method that is gradual in both network structure and training. Its network structure is an asymmetric pyramid structure that not only reduces memory consumption, but also enhances the receiving field concerning the original image. Therefore, it can maintain high efficiency while achieving a higher up-sampling rate.

At present, there are many technical and application challenges using deep learning to solve super-resolution reconstruction. For instance, images in practical applications often encounter unknown ambiguities such as camera noise [16–18], human factors [19], and motion blur. Therefore, the models trained on artificially constructed data sets tend to perform poorly in real-world data sets. Many researchers now have proposed several ways to solve this problem (e.g. Camera-SR [20], Dual CNN [21], NatSR [22], KMSR [23]). However, these methods have some congenital defects, such as difficulty in training and better effect on artificial data sets. In addition, super-resolution can be applied not only directly to specific areas of data and scenes, but also to other visual tasks. Therefore, it is also a challenge to apply SR to more specific areas, such as video surveillance [24–26], face recognition [27–29], target tracking [30–32], medical imaging [33, 34] and so on. At the same time, since most SR models currently perform SR with a fixed amplification factor while we often use SR with arbitrary scale factor in practical applications, the development of a single model with multi-scale super-resolution is also a potential development direction, such as Meta-SR [35], but it is not easy to achieve the quality of a single fixed-factor SR model.

Here are some classic models and methods. This paper is organised as follows: Section 2 provides an overview of HR and LR images. Section 3 introduces several evaluation indicators for

image super-resolution work. Section 4 introduces traditional methods in the field of image processing. In Sections 5 anf 6 outline the work of CNN networks in the field of image super-resolution and describes the work of GAN in the area of image super-resolution. In Section 7, some models are selected for comparison, and Section 8 is summarised the paper.

## 2 Super-resolution image overview

At present, the process of image acquisition and processing is often affected by many factors, such as image blurring, image downsampling, etc., resulting in the acquired image not meeting subsequent processing to achieve the desired index. The low-quality image obtained here is called a low-resolution image (LR image).

For single-image super-resolution (SISR) tasks, we often need a large number of LR images to learn how to map to super-resolution images. Most researchers usually downsample to obtain a LR image of the original image. This method mainly reduces the spatial resolution of the image by sampling the original image. The process of image downsampling belongs to the process of irreversible information loss. Therefore, if the downsampling ratio is too large, it will cause severe distortion of the image. Recently, there have been new methods for acquiring LR images. Chang Chen *et al*. proposed a LR image of a camera lens. A conventional camera captured a HR image of a set of LR images. Training on this dataset can better handle the recovery of the original LR images. Fig. 1 is a flowchart of obtaining a LR image. Equation (1) for obtaining an LR image is defined as, where $D()$ stands for the downsampling operation, and $n$ represents the image noise.

$$\mathrm{LR} = D(\mathrm{HR}, \mathrm{scale}) + n \tag{1}$$

The LR image obtained from the original image becomes an image down-sampling. In contrast, an image generated by passing a LR image through a super-resolution model is called a super-resolution image (SR image). Because the image down-sampling process is a morbid process, LR images and original images often cannot correspond one to one. Therefore, it is necessary to use the deep learning model to mine the prior knowledge of the image and use the prior knowledge of the image to build a mathematical model for image restoration, to recover the texture information and edge information of the LR image.

## 3 Evaluation index

In recent years, image processing technology has become popular increasingly, since image quality has had a lot of influence in many research fields, such as medical images and satellite imaging. Consequently, the evaluation method of image processing technology has also become an emerging research direction [36]. According to whether human is involved, we can divide image quality evaluation methods into two categories: subjective evaluation and objective evaluation. The evaluator of subjective evaluation is human, which can truly reflect human visual perception. The objective evaluation method uses a mathematical model to reflect the visual perception of the human eye and gives numerical results of the evaluation. At present, the mainstream objective evaluation methods mainly include peak-signal-to-noise ratio (PSNR), SSIM, Perceptual index (PI) and root MSE(RMSE), which will be introduced in turn.

### *3.1 PSNR*

The PSNR [37] is a mathematical method based on image pixel statistics. It uses statistical methods to measure the quality of the resulting image by calculating the difference between the greyscale values of the pixels of the resulting image corresponding to the original image.

Equation (2) is the calculation formula of PSNR, $F$ refers to the resulting image, $R$ refers to the original image, and their sizes are both $M \times N$. Where the $M$ stands for the image height, $N$ refers to the image width.

$$\mathrm{PSNR} = 10\log_{10}\frac{255^2}{\frac{1}{MN}\sum_{i=1}^{M}\sum_{i=1}^{N}|R(i,j)-F(i,j)|^2} \tag{2}$$

The larger the PSNR value, the smaller the difference between the resulting image and the original image, which means the better the image quality. This method is relatively simple and easy to implement and has full applications in the fields of image denoising and image super-resolution. However, since the PSNR is based on the global statistics of image pixel values, the local visual factors of the human eye are not considered. As for human eyes, the sensitivity to different regions is different, and the perception result of a specific area is also affected by the surrounding neighbouring areas, so the evaluation results of PSRN may have deviated from the perception of the human eye.

### *3.2 Structural similarity (SSIM)*

Considering the high structure of natural images, there is a strong correlation between their pixels, which often carries essential information about the structure of the object. While the human visual system mainly acquires structural information from the visible region, so it is feasible to perceive the approximate knowledge information of the image distortion by detecting the deterioration of the structural information.

The measurement system of SSIM consists of three measurement modules: brightness, contrast, and structure [38].

First, we use the average grey level to estimate the brightness, as shown in (3), where the brightness contrast function $l(x, y)$ is a function of $\mu_x$, $\mu_y$. $N$ are the pixels of the image.

$$\mu_x = \frac{1}{N}\sum_{i=1}^{Nx_i} \tag{3}$$

Then, the standard deviation is used to estimate the contrast. As shown in (4), the contrast function $c(x, y)$ is a function of $\sigma_x$, $\sigma_y$.

$$\sigma_x = \left(\frac{1}{N-1}\sum_{i=1}^{N}(x_i-\mu_x)^2\right)^{\frac{1}{2}} \tag{4}$$

Next, the structure comparison function $S(x,y)$ is defined as a function of $((x-\mu_x))/\sigma_x$, $((x-\mu_y))/\sigma_y$.

Finally, the complete SSIM function is shown in (5). Among them, $l(x, y)$ compares brightness, $c(x, y)$ compares contrast, and $s(x, y)$ compares structure.

$$S(x,y) = f(l(x,y), c(x,y), s(x,y)) \tag{5}$$

The feature statistics of an image are usually unevenly distributed in the pixel space while the distortion of the image varies in space. Therefore, in the calculation process of image quality, the local solution of SSIM is more accurate than the global. Considering people often only focus on a region of the image in reality, so local processing is more in line with the characteristics of the human visual system. Also, the local quality detection of the image can obtain the mapping matrix of the change of the picture quality more accurately, so the result can be applied to other aspects.

### *3.3 PI*

PI is an evaluation criterion for the ECCV2018 workshop PIRM2018's Perceptual SR Image Reconstruction Challenge [39]. According to the definition of the event organisers

$$\mathrm{PI} = \frac{1}{2}((10 - Ma) + \mathrm{NIQE}) \tag{6}$$

Among them, Ma is a non-reference quality indicator applied in the field of image super-resolution reconstruction, which does not refer to real images. It designs the types of low-level statistical

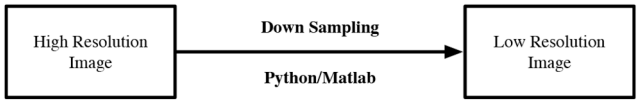


**Fig. 1** *Flowchart of LR image acquisition*

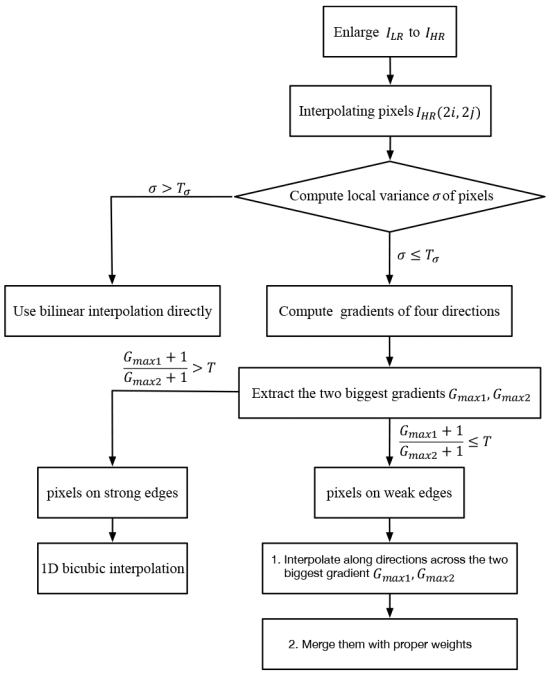


**Fig. 2** *Flowchart of directional BI method*

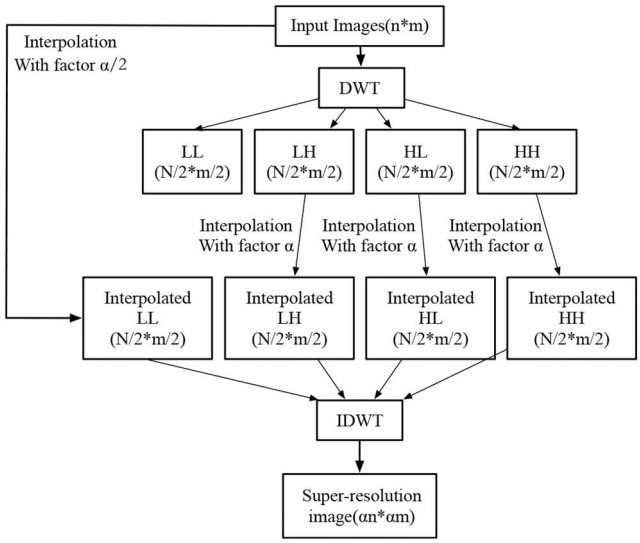


**Fig. 3** *Flowchart of DWT and BI method*

features in the spatial and frequency domains to quantify super-resolution artefacts, and learning a two-stage regression model to predict the quality score of SR images.

Natural image quality evaluator (NIQE) [40] is an image quality evaluation algorithm. It does not need to use the human-rated distorted images for training. After calculating the localised MSCN normalised image, part of the image blocks is selected as the training data according to the local activity. The model parameters are obtained by fitting the generalised Gaussian model as features, and the multivariate Gaussian model describes these features. The image quality is then determined based on the distance between the image feature model parameters to be evaluated and the pre-established model parameters in the evaluation process.

The human opinion study verified that the perceptual coefficient is highly correlated with the rating of human observers and a lower PI indicates a better perceived quality.

### *3.4 RMSE*

RMSE [41] is a frequently used measure of the differences between values (sample or population values) predicted by a model or an estimator and the values observed. The RMSE is very sensitive to very large or very small errors in a set of measurements, so the RMSE is a good reflection of the precision of the measurement. While in actual measurements, the number of observations $n$ is always finite, and the true value can only be replaced with the most reliable (best) value.

In the ECCV2018 workshop PIRM2018's Perceptual SR Image Reconstruction Challenge, the RMSE is defined as the square root of the MSE of all pixels in all images. Equation (7) is expressed as follows:

$$\mathrm{RMSE} = \sqrt{\frac{1}{M}\sum_{i=1}^{M}\frac{1}{N_i}\parallel X_i^{\mathrm{HR}} - X_i^{\mathrm{EST}} \parallel^2} \tag{7}$$

where $X_i^{\mathrm{HR}}$, $X_i^{\mathrm{EST}}$ are the $i$th real image and evaluation image, respectively. $N_i$ is the number of pixels in $X_i^{HR}$, and $M$ is the number of images in the test.

## 4 Image super-resolution methods before deep learning

### *4.1 Methods based on interpolation*

*4.1.1 Method based on directional bicubic interpolation (BI):* Among the image super-resolution reconstruction methods based on interpolation, BI has become a standard method because of its low complexity and relatively good results. However, it only interpolates the image edges horizontally and vertically so that the edges are vulnerable to artefacts. Liu *et al*. [42] proposed a directional BI method, which used different methods to interpolate lost pixels based on local intensity and direction to better preserve sharp edges and details. The flowchart of this method is shown in Fig. 2.

As shown in Fig. 2, the authors first estimated edge strength and direction from local image gradients and then interpolated in different ways for strong or weak edges and expectations on textures. $T_\sigma$ and $T$ are both thresholds.

*4.1.2 Method based on DWT and BI:* Kumar and Singh. [43] proposed a super-resolution technique based on the interpolation of high-frequency subband images obtained by the discrete wavelet transform (DWT) and input images. As shown in Fig. 3, the authors first used DWT to decompose an image into different subband images, namely low-low (LL), low-high (LH), high-low (HL), and high-high (HH). Then interpolate high-frequency subband images and the LR input image. LL is not used because LL is the LR of the original image and contains less information. It is worth noting that the input image is interpolated with half the interpolation factor. Finally, use inverse DWT (IDWT) to combine all these images to generate a new SR image.

### *4.2 Methods based on regularisation*

*4.2.1 Method based on regularisation with stationary gradient fidelity:* Yu *et al*. [3] proposed a SISR method based on regularisation with stationary gradient fidelity. The authors use a stationary fidelity gradient method based on error interpolation to estimate the smooth gradient of the fidelity term.

As shown in Fig. 4, first define the regularisation energy $E(I_h)$ used to estimate the best reconstructed HR image, as follows:

$$E(I_h) = F(I_h) + \lambda \cdot R(I_h) \tag{8}$$

where $F(I_h)$ is fidelity term, which brings the reconstruction result close to the true value, and $R(I_h)$ is regularisation term used to overcome the ill-posed problem. In order to find the minimum value of $E(I_h)$, use the gradient descent method and use (9) to continuously update the HR image.

$$I_h = I_h - \tau \frac{\partial E}{\partial I_h} \tag{9}$$

where $\tau$ is the time step. The authors believe that when defining the fidelity term, not only the down-sampled pixels should be considered, but also the pixels lost during the down-sampling process, i.e. the HR should be more concerned than the LR, so the blurred HR image $I_h$ is used to define $F(\cdot)$, the equation is as follows:

$$F(I_h) = \| I_h \otimes g_s - \tilde{I}_h \|^2 \tag{10}$$

where $g_s$ is the Gaussian kernel, $\otimes$ is the convolution operator. When calculating the gradient of the fidelity term, the equation is as follows:

$$\frac{\partial F}{\partial I_h} = (I_h \otimes g_s - \tilde{I}_h) \otimes g_s \tag{11}$$

It can be seen from (12) that the gradient of the fidelity term mainly depends on the error $(I_h \otimes g_s - \widetilde{I_h})$, So the author proposed to use interpolation to estimate this error, as shown below:

$$(I_h \otimes g_s - I \sim_h) \simeq U_s^b(D_s(I_h \otimes g_s) - I_l) \tag{12}$$

where $U_s^b(\cdot)$ is a BI, $D_s$ is a down-sampling operator. Where $\partial F/\partial I_h$, use the method of (9) to reconstruct the HR image. Therefore, (13) can be estimated as follows:

$$\frac{\partial F}{\partial I_h} \simeq U_s^b(D_s(I_h \otimes g_s) - I_l) \otimes g_s \tag{13}$$

The author named this method as error interpolation fidelity gradient (EIFG), which uses the difference in pixels after down-sampling to interpolate the differences in pixels lost during downs-sampling. The fidelity gradient obtained using this method is more stationary.

*4.2.2 Method based on joint regularisation:* Chang *et al*. [44] proposed a reconstruction-based single image super-resolution method by using joint regularisation, which combined group-residual-based regularisation (GRR) and a ridge-regression-based regularisation (3R). The flowchart of this method is shown in Fig. 5.

As shown in Fig. 5, the authors construct the minimisation problem shown in the following equation to realise the reconstruction from LR image to HR image.

$$\hat{X} = \arg\min_x \frac{1}{2} \| Y - S\boldsymbol{H}X \|_2^2 + \alpha \Psi_{\mathrm{GRR}}(X) + \beta \Psi_{3\mathrm{R}}(X) \tag{14}$$

where $\alpha$ and $\beta$ are trade-off parameters, $Y$ and $X$ represent the LR image and its HR version, $\boldsymbol{H}$ represents the blurring matrix, $S$ represents the decimation operator. The first term in the above formula is the fidelity term, and the last two terms are regularisation terms. Among them, $\Psi_{\mathrm{GRR}}(X)$ uses the structural information in the image, and $\Psi_{3\mathrm{R}}(X)$ introduces HR information from the external data set.

$\Psi_{\mathrm{GRR}}(X)$ is defined as follows:

$$\Psi_{\mathrm{GRR}}(X) = \sum_i \| W_i \circ (\boldsymbol{F}_i DX - E_i) \|_1 \tag{15}$$

where $\boldsymbol{F}_i$ represents the matrix which extracts the $i$th group of similar patches in the gradient domain, $E_i$ represents the estimation of the $i$th group of similar patches in the gradient domain, then $(\boldsymbol{F}_i - DX - E_i)$ represents the residual of a group of similar patches in the gradient domain, $\boldsymbol{W}_i$ is a weight matrix used to compensate the unreliable estimation of the $i$th group, $\circ$ is the Hadamard product. $\Psi_{3\mathrm{R}}(X)$ is defined as follows:

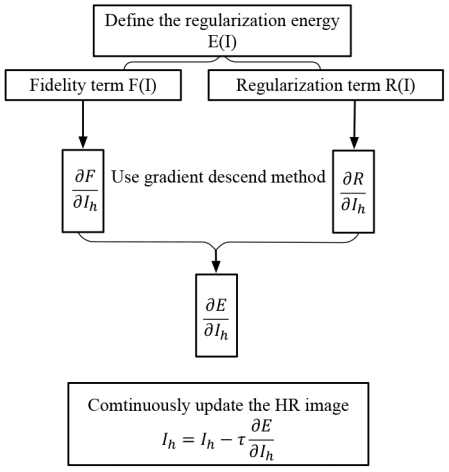


**Fig. 4** *Flowchart of regularisation with stationary gradient fidelity method*

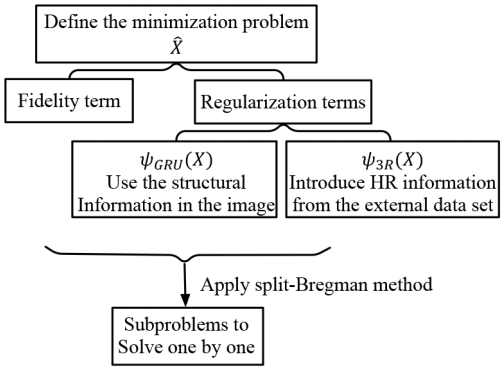


**Fig. 5** *Flowchart of joint regularisation method*

$$\Psi_{3\mathrm{R}}(X) = \sum_l \| \boldsymbol{R}_l \boldsymbol{B} X - \boldsymbol{P}_j R_l \boldsymbol{B} X \|_1 \tag{16}$$

where $\boldsymbol{R}_l$ represents the matrix which extracts the feature of the $l$th image patch, $\boldsymbol{B}$ represents the feature transformation matrix, $\boldsymbol{P}_j$ is the projection matrix.

In order to solve the minimisation problem shown in (14), the authors used split-Bregman method [45] to split it into multiple sub-problems and solve them one-by-one.

## 5 Image SRCNN

### *5.1 SRCNN*

SRCNN [46] is the pioneering work of deep learning applications in the field of super-resolution reconstruction. SRCNN's structure is very simple, using only three convolution layers. The network structure of SRCNN is shown in Fig. 6.

For a given LR image, it uses the bicubic algorithm to zoom to the target size firstly and then set the processed image to $Y$, after that through the three convolution layers, the following functions are implemented.

*5.1.1 Feature extraction and representation:* The patches are extracted from the LR image and each patch is represented as a high-dimensional vector. These vectors include a set of characteristic surfaces equal in number to the dimension of the vector. The first convolutional layer is expressed by the following euation:

$$F_1(Y) = \max(0, W_1 \times Y + \boldsymbol{B}_1) \tag{17}$$

where $W_1$ and $\boldsymbol{B}_1$ represent filters and offsets, respectively. The size of $W_1$ is $c \times f_1 \times f_1 \times n_1$, where $c$ is the number of channels in the input image, $f_1$ is the channel size of the filter, $n_1$ is the number of

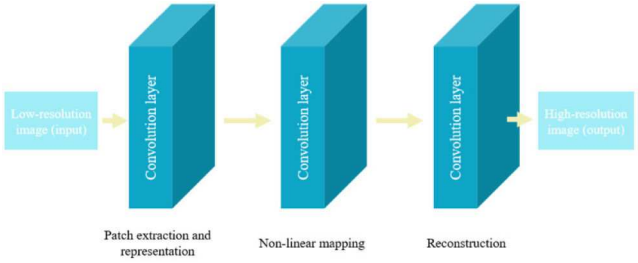


**Fig. 6** *SRCNN structure*

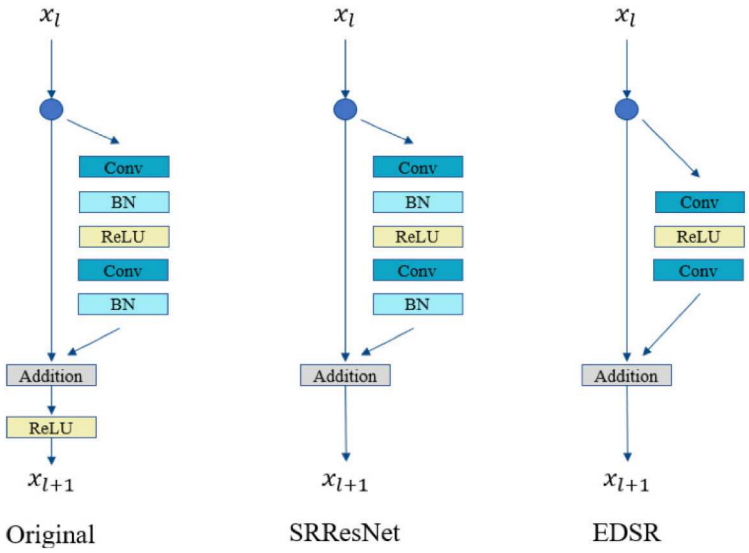


**Fig. 7** *Comparison of the original residual network, SRResNet and EDSR residual blocks*

filters, and $\boldsymbol{B}_1$ is a vector of $n_1$ dimensions, each of which is associated with the filters.

*5.1.2 Non-linear mapping:* This operation maps each high-dimensional vector onto another high-dimensional vector non-linearly. The vector of each map is conceptually a representation of a HR patch. These vectors include another set of characteristic surfaces. The first layer extracts an $n_1$ dimension feature from each patch. In the second operation, we map each of these $n_1$-dimensional vectors to an $n_2$-dimensional vector. The second operation can be expressed by the following euation:

$$F_2(Y) = \max(0, W_2 \times F_1(Y) + \boldsymbol{B}_2) \tag{18}$$

Here, the size of $W_2$ is $n_1 \times 1 \times 1 \times n_2$, and $\boldsymbol{B}_2$ is a $n_2$-dimensional vector. The $n_2$-dimensional vector of each output is conceptually a representation of the HR patch that will be used for reconstruction.

*5.1.3 Reconstruction:* This operation aggregates the HR patch representation described above to generate a final HR image that is as similar as possible to the original HR image. Expressed by the following equation:

$$F(Y) = W_3 \times F_2(Y) + \boldsymbol{B}_3 \tag{19}$$

where the size of $W_3$ is $n_2 \times f_3 \times f_3 \times c$, and $\boldsymbol{B}_3$ is a $c$-dimensional vector.

SRCNN uses MSE as a loss function, which is beneficial to obtain a higher PSNR. Let the original HR image be $X$, so the loss function formula will be shown in (20), where the function $F(\cdot)$ is the method of SRCNN, $Y_i$ represents the original image, and $X_i$ represents the generated image.

$$L_{\text{MSE}} = \frac{1}{n}\sum_{i=1}^{n} \| F(Y_i - X_i) \|^2 \tag{20}$$

## 5.2 EDSR

EDSR [8] is a model of the winner in the NTIRE2017 Super-Resolution Challenge [47]. The network structure of EDSR is based on the improvement of SRResNet [9]. On the basis of SRResNet, the batch normalisation (BN) [48] layer in the residual block is removed, and the ReLU active layer is not set outside the residual block. The model also has no residual scaling layers because it uses only 64-dimensional feature for each convolutional layer. The comparison of the residual block structure of EDSR with the original residual network and SRResNet is shown in Fig. 7 The network structure of EDSR is shown in Fig. 7.

The BN layer in SRResNet comes from the most primitive ResNet [49]. And in the case that the original ResNet was first proposed to solve high-level computer vision problems, such as classification and detection, applying the structure of ResNet directly to low-level computer vision problems like super-resolution does not perform well. At the same time, since the BN layer consumes the same amount of memory as the convolutional layer in front of it, it can save memory resources after being removed, which means that EDSR can stack more network layers or extract more features for each layer for better performance with the same computing resources.

EDSR uses the $L_1$ loss function to optimise the network model (see Fig. 8). The loss function is shown below:

$$L_1(P) = \frac{1}{N}\sum_{p \in P} |x(p) - y(p)| \tag{21}$$

where $p$ is the pixel number, $P$ is the patch, $x(p)$ and $y(p)$ are the pixel values, respectively, for processing the patch and the original image.

During training, a low-multiple up-sampling model is trained first, and then a high-multiplied up-sampling model is initialised by training the parameters obtained by the low-multiplied up-sampling model. This can reduce the training time of the high-magnification up-sampling model, and the training results are better.

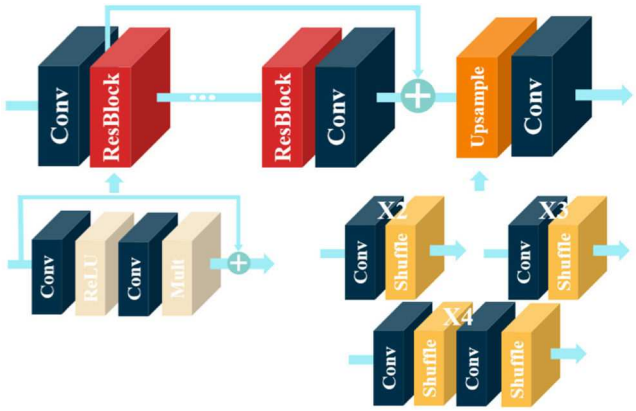


**Fig. 8** *Network structure of EDSR*

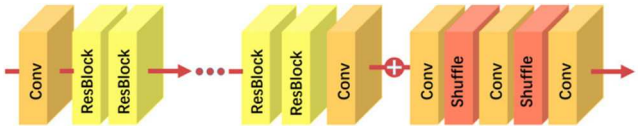


**Fig. 9** *Overall structure of WDSR and EDSR*

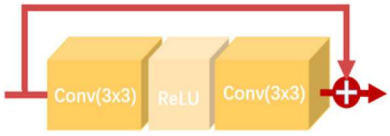


**Fig. 10** *ResBlock of EDSR*

### 5.3 WDSR

WDSR [10] is a super-resolution framework proposed by JiaHui Yu *et al*. in 2018. At the same time, the WDSR-based image super-resolution method also obtained the first name of single image super-resolution in all three real tracks in the NTIRE 2018 challenge [50]. WDSR is an improved algorithm based on the CNN optimisation model, and the CNN-based SR algorithm can be optimised in the following four directions.

*5.3.1 Up-sampling algorithm:* The current CNN method is mainly to learn HR features that are converted from LR up-sampling to HR images. The traditional up-sampling methods are deconvolution and bilinear interpolation, etc., which are not suitable for restoring the details and texture of LR images, so it tends to produce images that are too smooth and also introduces too many artefacts. At the 2016 CVPR conference, a new convolution algorithm for pixel shuffle [51] completely tailored for image super-resolution was proposed. Through the way of inserting LR features into LR images periodically at specific locations, the risk of loss of detail caused by artefacts can be significantly reduced.

*5.3.2 Deep neural network:* The depth of the neural network is one of the key factors affecting the performance of the SR algorithm. At the same time, using the method of the cyclic neural network can increase the reusability of weights.

*5.3.3 Skip-connecting:* Based on the Resnet algorithm, the front layer output is connected to the deep layer output. First, the gradient dispersion of backpropagation can be effectively solved, and second, shallow feature information can be effectively utilised. The current well-performing SR algorithms will contain ResBlock almost.

*5.3.4 Batch normalisation:* Various image super-resolution algorithms seem to be inseparable from BN. However, BN is not the only normalisation method. Currently, the popular normalisation methods of the SR algorithms are BN and weight normalisation.

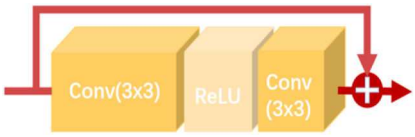


**Fig. 11** *ResBlock of WDSR-A*

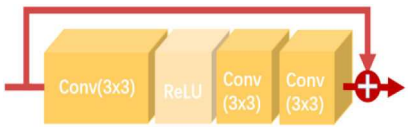


**Fig. 12** *ResBlock of WDSR-B*

WDSR is improved based on the EDSR algorithm. EDSR and WDSR are consistent across the overall framework, as shown in Fig. 9, consisting of the convolutional layers, the residual blocks, and the up-sampled layers. WDSR has partially improved the details of EDSR. The first is to remove most of the redundant convolutional layers, which reduces memory and increases computational speed.

On the other hand, WDSR changed the ResBlock structure of EDSR. The removed redundant layers are absorbed into the ResBlock. Through a lot of experiments, the results of the image are not degraded, so it can be indirectly proved that removing the redundant convolutional layer outside the ResBlock can reduce the computational overhead.

Figs. 10–12 show the ResBlock of EDSR, WDSR-A, and WDSR-B. For the ResBlock of EDSR, its activation function, shown as ReLU, is operated between two convolutional layers, and the number of filters per convolutional layer is small. In contrast, ResBlock of WDSR-A increases the width of the feature map by increasing the number of active layer pre-convolution layer convolution kernel filters without adding computational overhead. It allows the network to learn more detailed features of the picture. In addition, the main frontier of WDSR is to increase the number of channels of the shallow feature map of the activation function to restore the details and texture parts of the image. WDSR-B further liberates the computational overhead relative to WDSR-A, splitting the large convolutional layer into two small convolutional layers after the activation function. This allows a wider range of feature maps to be obtained with the same computational overhead.

Secondary, WDSR has another innovation compared to EDSR is replacing BN [48] with weight normalisation [52]. Why use weight normalisation instead of BN? First, the weight normalisation accelerates the convergence of deep neural network parameters by rewriting deep learning weights. In addition, since the introduction of mini-batch, it is also applied to RNN-based deep learning network. At the same time, BN calculates the mean and variance of the data set based on mini-batch instead of the entire data set, which is equivalent to performing the gradient calculation and introducing noise. Therefore, BN is not used to reinforce learning and generating models. In contrast, weight normalisation rewrites the weight $W$ by the scalar g and the vector $\boldsymbol{v}$, and the rewriting vector $\boldsymbol{v}$ is fixed. Thus we can think that being based on weight normalisation can introduce less noise than BN. Using weight normalisation does not require additional storage space to preserve the mean and variance of the mini batch, and the additional computational caused by the forward signal propagation and inverse gradient calculations in deep learning network is also small. Apparently, the normalisation with weight normalisation needs less memory while leads faster calculations.

### 5.4 RDN

The RDN [11] proposes a RDB to mine rich local features through densely connected convolutional layers. The overall structure of the model is shown in Fig. 13.

As shown in the above, let $I_{LR}$ and $I_{HR}$ be the input and output of the model, respectively. The model consists of four main modules.

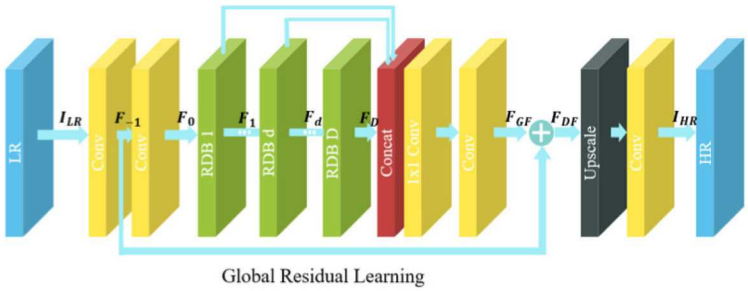


**Fig. 13** *Network structure of RDN*

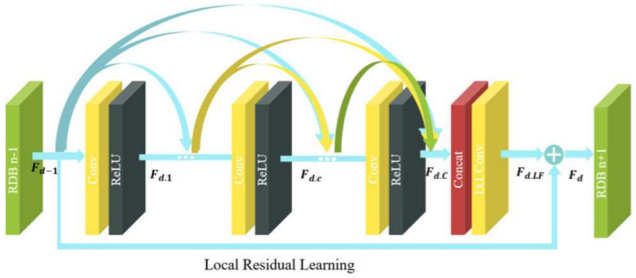


**Fig. 14** *Structure of the RDBs module*

*5.4.1 Shallow feature extraction net (SFENet):* This module includes two convolutional layers for extracting shallow features. The first convolutional layer extracts the feature $F_{-1}$ from the LR input as follows:

$$F_{-1} = H_{\mathrm{SFE1}}(I_{\mathrm{LR}}) \tag{22}$$

where $F_{\mathrm{SFE1}}(\cdot)$ represents the convolution operation of the first shallow feature extraction layer, and $F_{-1}$ is used for further shallow feature extraction and global residual learning. So we can go further as follows:

$$F_{-0} = H_{\mathrm{SFE2}}(F_{-1}) \tag{23}$$

where $F_{\mathrm{SFE2}}(\cdot)$ represents the convolution operation of the second shallow feature extraction layer, and $F_{-0}$ is the input of the RDBs.

*5.4.2 RDBs:* This module consists of multiple RDBs. The structure of each RDB is shown in Fig. 14.

Assuming the total number of RDBs is $D$, so the output of the $d$th RDB can be expressed as

$$\begin{aligned} F_d &= H_{\mathrm{RDB}.d}(F_{d-1}) \\ &= H_{\mathrm{RDB}.d}(H_{\mathrm{RDB}.d-1}(\cdots(H_{\mathrm{RDB}.1}(F_0))\cdots)) \end{aligned} \tag{24}$$

where $H_{\mathrm{RDB}.d}$ represents the operation of the $d$th RDB. This operation is a composite function operation, such as a convolution operation and a ReLU activation function. $F_d$ is the output of the $d$th RDB and $F_{d-1}$ refers to the $(d-1)$th RDB.

The RDB integrates the modules of the residual blocks and the dense blocks. The difference among these three modules is shown in Fig. 15.

The RDB mainly consists of three parts: the contiguous memory module passes the state of the previous RDB to each layer of the current RDB. The local feature fusion module combines the state of the previous RDB with the state of each Conv layer in the current RDB. The local residual learning module combines the input of the RDB with the features of the output after the 1*1 convolution operation to help improve the expressiveness of the model.

*5.4.3 Dense feature fusion (DFF):* The DFF module consists of global feature fusion and global residual learning.

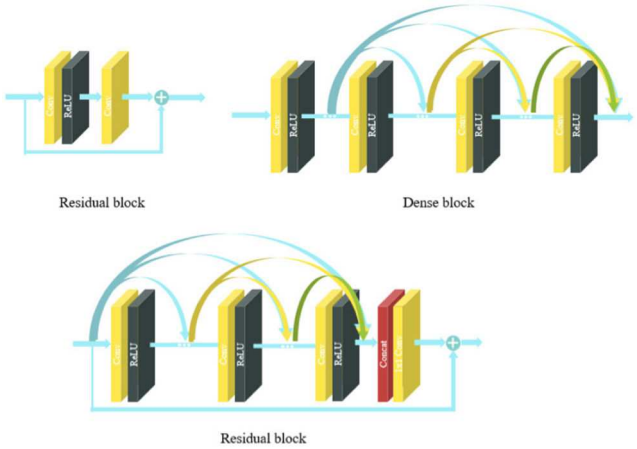


**Fig. 15** *Comparison of RDB with residual blocks and dense blocks*

Global feature fusion combines the features of each RDB output $[F_1, \ldots, F_D]$ to extract a global feature $F_{\mathrm{GF}}$ which can be expressed as

$$F_{\mathrm{GB}} = H_{\mathrm{GFF}}([F_1, \ldots, F_D]) \tag{25}$$

where $H_{\mathrm{GFF}}$ is a composite function, including a 1*1 convolution layer for adaptively merging different levels of features, and a 3*3 convolution layer for further extracting the features of global residual learning.

Global residual learning combines the original shallow feature $F_{-1}$ with the $F_{\mathrm{GF}}$ obtained in global feature fusion, which can be expressed as

$$F_{\mathrm{DF}} = F_{-1} + F_{\mathrm{GF}} \tag{26}$$

*5.4.4 Up-sampling net (UPNet):* This module represents the last up-sampling and convolution operation of the network, which is able to enlarge the input picture. The output of the entire model can be expressed as

$$I_{\mathrm{SR}} = H_{\mathrm{RDN}}(I_{\mathrm{LR}}) \tag{27}$$

where $H_{\mathrm{RDN}}$ represents all operations of the entire RDN model, $I_{\mathrm{SR}}$ is the super-resolution image and $I_{\mathrm{LR}}$ refers to the low-resolution image.

### 5.5 DSRN

Many super-resolution models can convert their networks into an expanded limited single-state RNN. DSRN [6] is proposed as a two-state design based on the finite expansion of single-state RNN. Compared to the models that are using the fixed-resolution signal, DSRN utilises both LR and HR signals and performs two-way fusion in it. Excellent qualitative and quantitative results are obtained in the baseline data set, and DSRN has better performance than the most advanced methods in terms of memory consumption and image reconstruction quality.

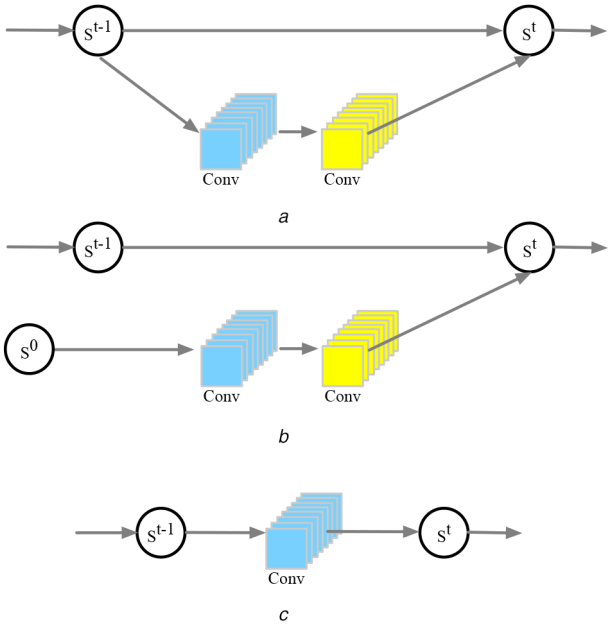


**Fig. 16** *Network structure of DSRN*
***(a)*** ResNet, ***(b)*** DRRN, ***(c)*** DRCN

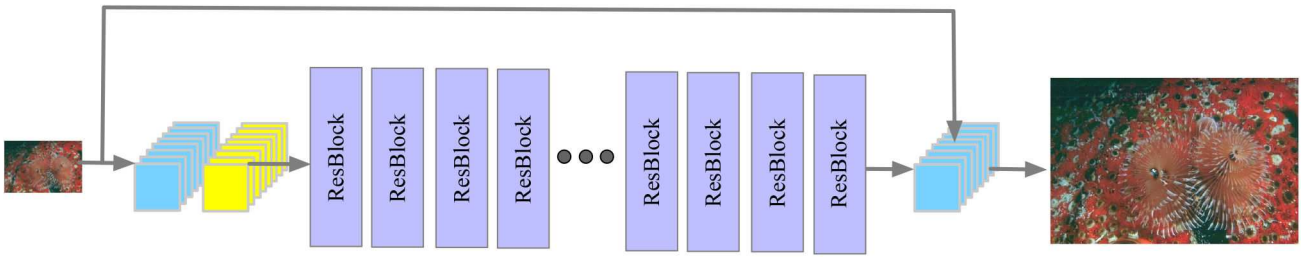


**Fig. 17** *Network structure of LRFNet*

Also, in the deeper CNN model, the performance of super-resolution can be better improved, because the mapping relationship between LR and SR can be better learned in the deeper model with more parameters, and the deep model can be effectively prevented from disappearing due to the proposed residual network.

After referring to the methods of DRN and DRCN, the method is found to be implemented in a single state structure RNN. In contrast, the mapping functions of the two networks are not the same. Our dual-state recursive network can perform super-resolution reconstruction at different spatial resolutions. Resnet, DRCN [53], and DRRN [54] are modified for the RNN. ResNet branches in the upper layer of the network, which is divided into direct connections and jump connections through two convolution blocks. By contrast, the DRCN change method is more straightforward, which is directly connected by adding a convolution layer between the upper and lower layers. The DRRN is connected in a single-state RNN by connecting the initial state weight to the hop connection. The network structure is shown in Fig. 16.

### *5.6 LRFNet*

CNN is the basis for deep learning of SISR. However, since the introduction of ResNet, the existing network is so deep that it has many network parameters, which leads the time, and the amount of the occupation of the GPUs and CPUs is too much during the test and training process. LRFNet uses an Atrous conv and a 1D separable kernel to reduce parameter requirements. Because using a 1D detachable filter and Atrous conv can achieve the performance of a conventional conv with fewer parameters. Furthermore, the input of LRFNet-B [7] is two three-times magnified LR images connected using the global and local plus, and each residual block uses an EDSR structure. Among them, LRFNet-B uses 12 residual blocks, Conv core 3*3, 64 filters, and ReLU.

Using a 1D Conv layer needs fewer parameters while it has the same performance as a multi-dimensional Conv. This is because the separable filter is a combination of the matrix products of two low-dimensional convolution filters, and the ordinary 2D one is composed of two 1D filters. LRFNet-A uses Atrous Conv as the convolutional layer of this method to extend the conventional convolutional layer, and for the 1D core, each residual block has a conv of 1*$k$ and $K$*1. Since it only adds Atrous Conv extended convolution to some blocks, the effect is not as good as RDN and ESRGAN. The network structure of the model is shown in Fig. 17.

### *5.7 RCAN*

This approach addresses the problem of difficult training in deep SISR network. It believes that LR inputs contain a lot of low-frequency information. Many CNN-based methods treat each feature in the channel equally, including this low-frequency information, which hinders the ability for the deep network to express. To this end, a RCAN [14] has been proposed to obtain a very deep network.

The paper believes that a simple stacking residual block is difficult to be improved by deepening the network. Therefore, in the network structure, a residual structure (RIR) is proposed to help form a deep network which can reach the maximum depth currently known and is capable of providing very large receiving areas. The structure consists of a plurality of residual groups (RGs) with long skip connections, and each RG contains a plurality of simplified residual blocks with short skip connections. RIR allows the main network to focus on high-frequency information by bypassing the connection to bypass rich low-frequency information. In addition, a channel attention mechanism is proposed to adaptively re-adjust the characteristics of the channel

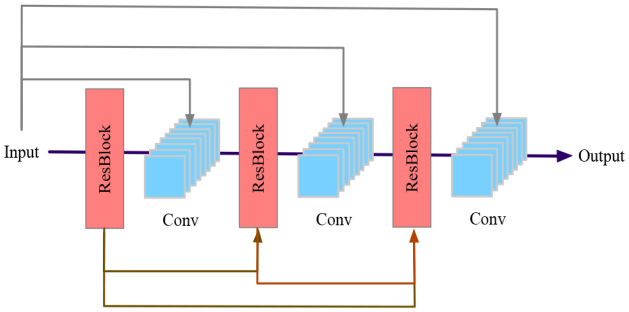


**Fig. 18** *Local cascading block structure of CARN*

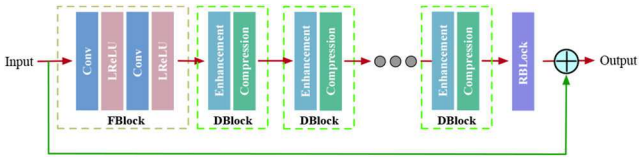


**Fig. 19** *Structure diagram of the IDN*

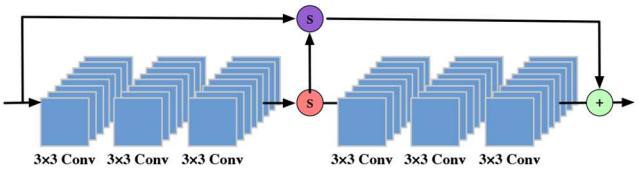


**Fig. 20** *Structure diagram of the enhancement unit*

mode by considering the interdependencies between the channels. Each RG includes a residual channel attention block (RCAB) with a short hop connection, and the global average information of the channel mode is incorporated into the channel descriptor using the global average pool.

RCAN is divided into four parts in the network structure.

Extract the shallow feature $F_0$ from the $I_{LR}$ using a convolutional layer

$$F_0 = H_{SF}(I_{LR}) \tag{28}$$

where the $H_{SF}(\cdot)$ represents a convolution operation. $F_0$ is then used for depth feature extraction.

$$F_{DF} = H_{RIR}(F_0) \tag{29}$$

Think of the RIR output as a deep feature and zoom by using the amplification module. where the $F_{UP}(\cdot)$ and $H_{UP}(\cdot)$ stand for the upscale module and upscaled feature, respectively,

$$F_{UP} = H_{UP}(F_{DF}) \tag{30}$$

The magnified features are finally reconstructed by a convolutional layer. $H_{REC}(\cdot)$ and $H_{RCAN}(\cdot)$ represent the reconstruction layer and RCAN method function of SISR method, respectively.

$$F_{SR} = H_{REC}(F_{UP}) = H_{RCAN}(I_{LR}) \tag{31}$$

### 5.8 CARN

This paper is dedicated to providing an accurate and lightweight depth network for image super-resolution to easily apply it. A lightweight deep learning model is designed to reduce parameters and reduce operands. A neural network CARN [12] based on the cascade module is proposed. By combining effective residual block and recursive network scheme, a variant model CARN-M is proposed to improve efficiency further.

The middle part of the model is based on ResNet, and the main difference is the existence of local and global cascading modules. Using a layered mechanism at the local and global level to merge multiple layers of functionality, the output of the middle layer is cascaded to a higher layer and finally converged on a $1 \times 1$ convolutional layer. Local cascading and global cascading except for the unit block is almost the same as other aspects of the ordinary residual block (as shown in Fig. 18), and to improve efficiency, a residual-E block is also proposed.

### 5.9 IDN

As the depth and width of the network increase, the CNN-based SISR method faces the problem of significant computation and memory overhead. Therefore, Zheng *et al*. proposed a deep and concise convolutional network (IDN) to implement the super-resolution function [55]. The IDN includes three parts: a feature extraction module, a stacked information distillation module, and a reconstruction module. The structure of the IDN is shown in Fig. 19.

The feature extraction module is the FBlock in the network structure diagram, which is mainly composed of two convolution layers. Then the focus of the network is the information distillation block (DBlock) shown in the structure diagram, which includes an enhancement unit and a compression unit. The enhancement unit is shown in Fig. 20. The enhancement unit consists of three $3 \times 3$ convolutions, each of which is followed by an LRelu activation function. The compression unit is composed of a $1 \times 1$ convolution layer, which is mainly responsible for fusing the information distilled from the enhancement unit.

### 5.10 Characteristics of CNN models

Below we summarise the characteristics of the CNN-based models introduced in this paper. SRCNN is a pioneering work, and its network structure is straightforward. EDSR improved the residual block structure of SRResNet and removed the BN layer to save memory resources. WDSR removed most of the redundant convolutional layer based on EDSR and changed BN to weight normalisation, which requires less memory and is faster to calculate. RDN proposed the residual dense network by integrating the modules of the residual block and the dense block to mine rich local features through densely connected convolutional layers. DSRN used both LR and HR signals and performed a two-way fusion in the model. LRFNet used atrous conv and a 1D separable kernel to reduce parameter requirements. RCAN proposed a residual structure (RIR), which allows the main network to focus on high-frequency information by passing the connection to avoid rich low-frequency details and offered a channel attention mechanism to adaptively readjust the characteristics of the channel mode by considering the interdependencies between the channels. CARN proposed an accurate and lightweight deep network with fewer parameters and operands. IDN is aimed at the problem of increased computation and memory overhead caused by the increase of network depth and width and proposes the use of the deep and concise convolutional network.

In conclusion, most CNN-based methods are pursuing deeper networks to improve performance. This is because theoretically, the depth of the network is critical to the performance of the model. When the number of network layers is increased, the network can extract more complex feature patterns, so theoretically, better

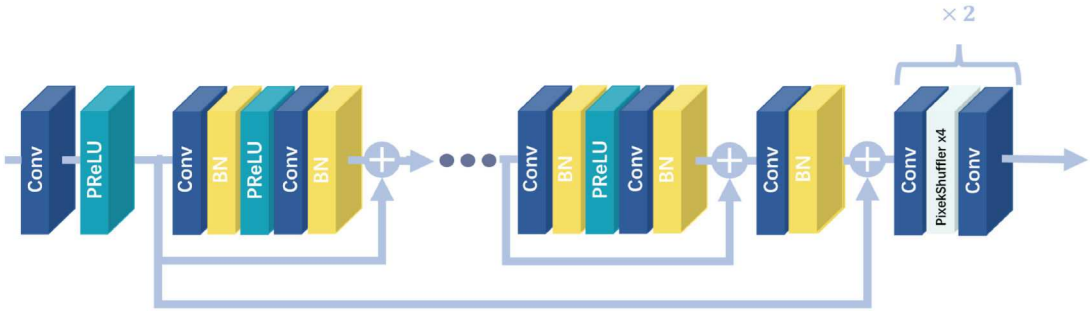


**Fig. 21** *Generation network structure of SRGAN*

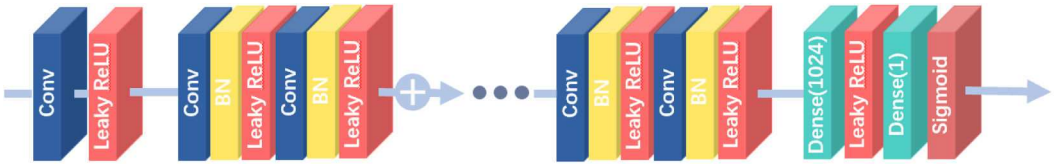


**Fig. 22** *Discriminant network structure of SRGAN*

results can be obtained when the model is deeper. However, deep network often has the problem of gradient disappearance or explosion. ResNet's proposal solves this critical problem. Its appearance has increased the number of layers of the network several times, which is a milestone. At the same time, Researchers have also noticed the considerable consumption of memory and time by deep network, so some models that focus on reducing memory consumption and the number of parameters have appeared, and have achieved some good results.

## 6 Image SRGAN

### 6.1 SRGAN

SRGAN [9] was proposed by Christian *et al*. in 2017 to use GAN for super-resolution reconstruction. The network is the first framework to recover 4 × down-sampled images. At the same time, the structure also modified the loss function from the mean squared loss function commonly used in the CNN method is replaced by a new perceptual loss function consisting of resistance loss and content loss. The loss function of SRGAN is shown in (32), where $l_X^{\mathrm{SR}}$ is the content loss, and $10^{-3} l_{\mathrm{Gen}}^{\mathrm{SR}}$ is the adversarial loss. A higher PSNR can be obtained with a MSE as a loss function, but at a higher recovery multiple, the reconstructed image is too smooth and lacks some detail realism. The adversarial loss function is based on the probability of the discriminator output, as shown in (33), where $D_{\theta_G}(\cdot)$ is the probability that an image belongs to a real HR image $G_{\theta_G}(I^{\mathrm{LR}})$, which is a reconstructed HR image. The content loss is a pixel-by-pixel loss of the feature map of an individual layer, including the minimum MSE (MSE loss) of the pixel space, as shown in (34). The minimum MSE of the feature space is a high-level feature that uses the VGG network [56] to extract images. By comparing the features of the generated image through the CNN and the features of the target image after convolving the neural network $\phi_{i,j}$, the generated picture and the target picture are more similar in semantics and style, as shown in (35). In (33)–(35), $N$ represents the number of pixels of the image, and $W$ and $H$ represent the width and height of the image, respectively.

$$l^{\mathrm{SR}} = l_X^{\mathrm{SR}} + 10^{-3} l_{\mathrm{Gen}}^{\mathrm{SR}} \tag{32}$$

$$l_{\mathrm{Gen}}^{\mathrm{SR}} = \sum_{n=1}^{N} -\log D_{\theta_G}\left(G_{\theta_G}\left(I^{\mathrm{LR}}\right)\right) \tag{33}$$

$$l_{\mathrm{MSE}}^{\mathrm{SR}} = \frac{1}{r^2 WH} \sum_{x=1}^{rW} \sum_{y=1}^{rH} \left(I_{x,y}^{\mathrm{HR}} - G_{\theta_G}\left(I^{\mathrm{LR}}\right)_{x,y}\right)^2 \tag{34}$$

$$l_{\mathrm{VGG}/i,j}^{\mathrm{SR}} = \frac{1}{W_{i,j} H_{i,j}} \sum_{x=1}^{W_{i,j}} \sum_{y=1}^{H_{i,j}} \left(\phi_{i,j}\left(I^{\mathrm{HR}}\right)_{x,y} - \phi_{i,j}\left(G_{\theta_G}\left(I^{\mathrm{HR}}\right)\right)_{x,y}\right)^2 \tag{35}$$

So we can derive the formula for generating the network loss function, as shown in (36), where $I_{\mathrm{G}}$ stands for the generator loss, $I_{\text{content loss}}$ refers the content loss, $I_{\text{VGG loss}}$ refers the VGG network loss, and $I_{\text{adversarial}}$ is the adversarial loss. The judgment network only has an adversarial loss function, as shown in (37). $E$ represents the operation of averaging all the fake data in the mini-batch. where $\sum_I[\cdot]$ represents the operation of averaging all the fake data and true data in the mini-batch.

$$l_G = l_{\text{content loss}} + l_{\text{VGG loss}} + l_{\text{adversarial}} \tag{36}$$

$$\begin{aligned} l_{\text{adversarial}} = \min_{\theta_G} \max_{\theta_D} & \sum_{I^{\mathrm{HR}} \sim p_{\text{train}}(I^{\mathrm{HR}})} \left[\log D_{\theta_G}\left(I^{\mathrm{HR}}\right)\right] \\ & + \sum_{I^{\mathrm{LR}} \sim p_G(I^{\mathrm{LR}})} \left[\log\left(1 - D_{\theta_G}\left(G_{\theta_G}\left(I^{\mathrm{HR}}\right)\right)\right)\right] \end{aligned} \tag{37}$$

The model's generation network and judgment network are shown in Figs. 21 and 22. The Residual Network Part (SRResNet) section contains multiple residual blocks, each of which contains two 3 × 3 convolutional layers, and the convolutional layer is followed by BN and PReLU as activation functions. Two 2 × sub-pixel convolution layers are used to increase the feature size. In the discriminating network part, there are eight convolutional layers. As the network layers deepens, the number of features increases continuously, and the feature size decreases. The activation function is selected as LeakyReLU. The probability of predicting a natural image is finally obtained by two fully connected layers and the final sigmoid activation function.

The main contribution of the model is presented in SRResnet proposed to obtain detailed textures and details in the image results. At the same time, the perceptual loss is applied to the anti-neural network, and the 4 × down-sampled image super-resolution is realised. But the model also has some problems, such as using the BN layer as part of the activation function. Although it can speed up the convergence, it will produce artefacts when the network is deeper and more complex.

### 6.2 ESRGAN

ESRGAN [57] is a model proposed based on the improvement of SRGAN. As mentioned above, the enlarged details of the images produced by SRGAN are often accompanied by unpleasant artefacts. In order to further improve the visual quality, ESRGAN mainly aims to improve the three key parts of SRGAN: network structure, adversarial loss, and perceived loss.

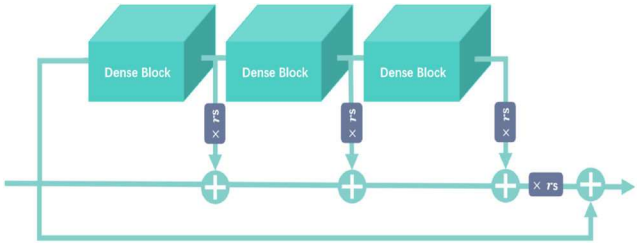


**Fig. 23** *Schematic diagram of the RRDB model*

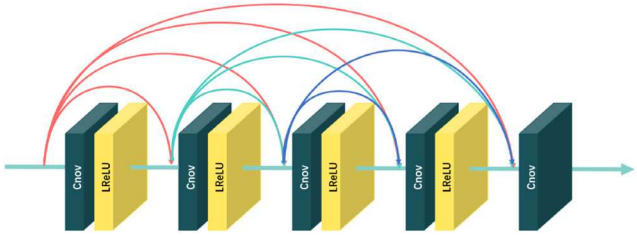


**Fig. 24** *Schematic diagram of the dense block model*

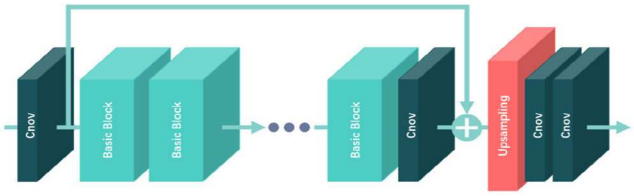


**Fig. 25** *Basic structure of SRResNet*

On the network structure, ESRGAN removed all BN layers. The original basic block is replaced by a residual-in-RDB (RRDB) [14], which combines a multi-level residual network with dense connections. The BN layer is removed because it tends to produce unpleasant artefacts and limit generalisation when training and testing data sets. Removing the BN layer can help improve generalisation and performance, and reduce computational complexity. RRDB uses deeper, more complex structures than the original residual blocks in SRGAN, since more layers and connections always means higher performance. Fig. 23 shows a schematic diagram of the RRDB model, and Fig. 24 shows a schematic diagram of the dense block.

In addition, ESRGAN improved the discriminator based on relativistic GAN [58]. The discriminator in SRGAN is used to estimate the probability that the image input to the discriminator is a real and natural image, while the relativistic discriminator attempts to estimate the probability that the actual image is more realistic than the fake image. This modification for discriminator helps to learn sharper edges and more excellent textures. In the generation of the network part, SRResNet is used as the basic network architecture, and in order to transfer most of the calculations to the feature space of the LR image, the amount of calculation is reduced. Fig. 25 shows the basic architecture of SSResNet. Each residual block contains two $3 \times 3$ convolutional layers. PReLU acts as an activation function after the convolutional layers, and two $2\times$ sub-pixel convolution layers are used to increase the feature size. Parametric leaky ReLU (PReLU) works better on more larger data sets and has overfitting risk on smaller data sets. It also prevents the problem of gradient disappearing.

In addition to improvements in network architecture, ESRGAN also uses several techniques to facilitate very deep network training: one is residual scaling, which is to reduce the residual by multiplying a constant between 0 and 1, and then add them to the main path to improve stability. The second is to use a smaller initialisation, making the initial parameter variance smaller and more comfortable to train.

In terms of the perceptual domain loss function, ESRGAN proposes a more efficient perceptual domain loss, using pre-activation features (VGG16 network). This will overcome two shortcomings. First, the activated features are very sparse, especially in a deep network. This sparse activation provides a weak monitoring effect and causes poor performance. Second, the use of activated features may cause the reconstructed image to be inconsistent with the brightness of the GT. This loss is based on a VGG16 network (MINCNet) for material identification, which focuses more on texture than on objects.

In addition, ESRGAN also implements the interpolation function of the network in order to be able to produce all possible results without introducing artefacts. At the same time, the perceived quality and fidelity of the image can be balanced, and the training time can be reduced without retraining the deep neural network. Equation (38) is the calculation method, and $\alpha$ is the interpolation parameter, where the $\theta_G^{\mathrm{PI}}$ stands for the parameters of interpolation, the $\theta_G^{\mathrm{PSNR}}$ represents the parameters of a PSNR-oriented network, and the $\theta_G^{\mathrm{GAN}}$ is the parameters of a GAN-based network.

$$\theta_G^{\mathrm{PI}} = (1-\alpha)\theta_G^{\mathrm{PSNR}} + \alpha\theta_G^{\mathrm{GAN}} \tag{38}$$

### 6.3 ProSR

This model [15] is mainly aimed at obtaining high-quality results in the case of large up-sampling factors and proposes a method that is gradual in both structure and training. In terms of structure, an asymmetric pyramid structure is proposed. Each level consists of a cascading dense compression unit (DCU) followed by a sub-pixel convolutional layer. The DCU is composed of dense blocks whose structure is CONV (1,1)-RELU-CONV (3,3). The function of each level of the pyramid is to perfect the feature and perform $2\times$ up-sampling on its input, allocate more DCUs at the lower level, reduce memory consumption, increase the receiving field relative to the original image, and achieve a higher up-sampling rate with the efficiency maintained. The network structure is shown in Fig. 26.

In order to achieve a more realistic effect, the GAN framework was adopted and a discriminator was designed. It matches the progressive nature of the generator network by calculating the fertility output for each scale. In the part of loss function, the more stable least squares loss is used instead of the original cross entropy loss.

Using the method of curriculum learning to improve training by gradually increasing the difficulty of learning tasks can improve training time and generalisation performance. Model training begins with the $2\times$ portion of the network, and the new level of the pyramid merges as it enters the new phase of the model, which reduces its impact on previously trained layers. This progressive training strategy that feeds different scales of training simultaneously into the network reduces the total training time

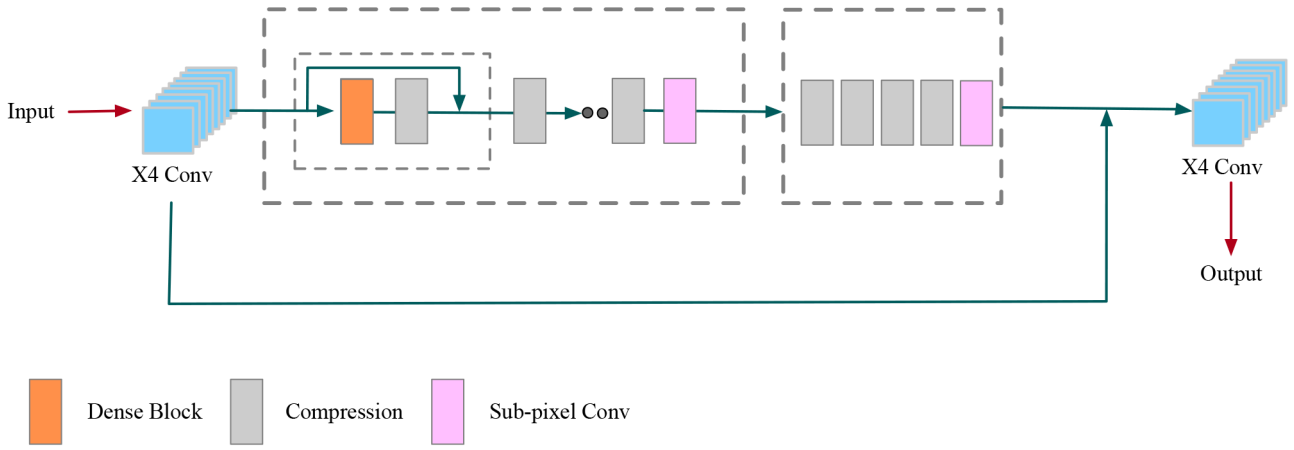


**Fig. 26** *Structure diagram of the ProSR network*

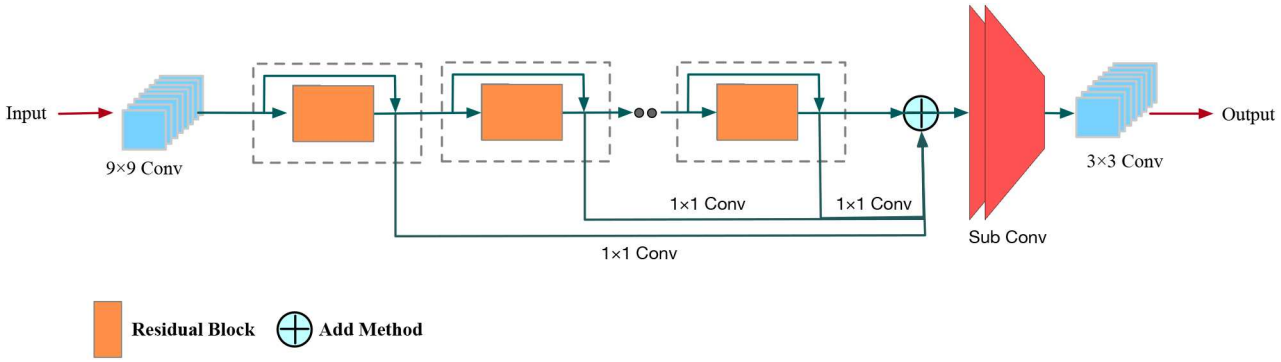


**Fig. 27** *Structure diagram of the SRFeat network*

greatly, and it has further performance improvements at all included scales and mitigates the instability in GAN training.

Compared to the other similar methods of accuracy reconstruction, the model's asymmetric pyramid structure contributes to faster runtimes and has five times faster runtime than the top team in the NTIRE 2018 challenge [50].

### 6.4 CinCycle

For LR inputs, the quality is further reduced due to noise and blurring, and for this complex case, supervised learning and accurate fuzzy kernel estimation will not be possible. This method [59] inputs a LR image with fuzzy kernel noise and maps it to a LR space with no noise blur and then up-sample image using a pre-trained super-resolution model to obtain a super-resolution image. And we adjust these two methods to get better super-resolution image output.

The method consists of two CycleGANs [60]. The first one will Cycleangle map the LR image to the noise free fuzzy LR space so that this module ensures proper denoising/deblurring of the LR input. Another CycleGAN is a pre-trained super-resolution model to up-sample intermediate results to the desired size. Finally, a training approach to learning is used to fine tune the network. The generation model of this network is similar to the architecture of ResNet.

Its effect is similar to that of CNN-based monitoring algorithms. The only advantage is that it can be used in the absence of data pairs, but it is not very good if it is for the game.

### 6.5 SRFeat

In 2018, Park *et al.* proposed a new SISR framework called SRFeat [61]. This framework mainly solves the problems of current GAN-based super-resolution methods used to generate real texture information. Among them, GAN methods tend to produce less meaningful high-frequency noise that has nothing to do with the input image. Therefore, the SRFeat model adds a discriminative network acting on the feature domain, so that the generation network can generate high-frequency features related to the image structure. The main innovations of the SRFeat model include the following two aspects:

1. Two types of discriminators are proposed, which are image domain and feature domain discriminators, which are mainly used to discriminate that the model generates high-frequency information instead of noise.
2. A ResNet-like skip connection was used in the generation network.

SRFeat's generation network is similar to SRResNet. Generate the network structure diagram, as shown in Fig. 27. However, compared to SRResNet, it uses more long-range hop connections. Park *et al*. believed that the SSResNet model is equivalent to each layer feature. The model uses $1 \times 1 - \text{Conv}$ as a bottleneck to connect the features of different layers, and then dynamically adjusts their weights. The layer adds all the features. It is similar to the attention mechanism in the RNN model. By doing so, the gradient can be more easily updated during the backward propagation process, and the features of the middle layer can be fully utilised to improve the final aggregation features.

Then use the sub-pixel convolutional layer to complete the scale-up operation. The structure of the discrimination network is similar to SRResNet. The structure diagram is shown in Fig. 25.

### 6.6 Characteristics of GAN models

Below we summarise the characteristics of the GAN-based models introduced in this paper. SRGAN is the first application of GAN to the field of super-resolution reconstruction. Its proposed generator network, SSResnet, can generate result images with detailed textures and details. At the same time, it applied perceptual loss to GAN, and it is the first framework capable of recovering $4\times$ down-sampled images. ESRGAN removed all BN layers based on SRGAN and replaces the original basic block with the RRDB. Besides, based on relativistic GAN, it improves the discriminator and improves the perceptual loss function. ProSR proposes a

method that is progressive in both structure and training to achieve high-quality results with large up-sampling factors. CinCycle maps LR image input with blur kernel noise to LR space without blur noise, which solves the problem of low-quality input quality degradation caused by noise and blur. SRFeat is directed at the problem that GAN-based methods tend to generate high-frequency noise that is unrelated to the input image and adds a discriminative network that acts on the feature domain so that the generation network can create high-frequency features related to the image structure.

In conclusion, most GAN-based methods are mostly based on the original SRGAN framework and continue to propose optimisation methods for some details of the inventive method's network structure and loss function. The most significant difference between GAN-based methods and CNN-based methods is that they are not committed to obtaining high PSNR values. In the work of SRGAN, it is pointed out that the pictures with high PSNR values are too smooth and lack some realism in detail. Therefore, a perceptual loss is proposed, and detailed textures and details are obtained. Although SRGAN itself has many problems, as a pioneer, its SRResNet generation network and perceptual loss function set the tone for GAN-based methods. Later, researchers have carried out a series of optimisations based on it and have achieved good results.

## 7 Compare results

### *7.1 Characteristics comparison*

Now, we analyse and compare the characteristics of the models mentioned in the paper. The detailed model features are compared in Table 1. Among them, the '91 images' is a data set proposed by Yang *et al.* [65], the '291 images' is composed of '91 images' and '200 images', and '200 images' is an image data set introduced by Martin *et al.* [66]. First, we observe Table 1 and find that before the application of artificial intelligence methods in the field of SISR, the mainstream algorithms are divided into interpolation and regularisation. The interpolation method is mainly based on the bicubic method. Because the bicubic method only interpolates the image edges in the horizontal and vertical directions, the traditional bicubic method is prone to generate image artefacts. Therefore, most of the optimisation methods of interpolation perform interpolation processing on the image based on the change of interpolation direction and filter selection. Because the difference method only refers to the original image, its method cannot learn deeper image features to restore sharper image textures. The regularisation method is to obtain a clearer image by adding a penalty to the loss function. Most of the current regularisation methods are regularisation based on ridge regression and Lasso regression. Regularisation can be combined with deep learning to restore image details, so the regularisation method is highly flexible. Second, according to Table 1, it is found that most of the methods of the current SISR model belong to one of CNN or GAN. And through experimental observation, it is found that the image texture generated by using the CNN model as the basic framework of the SISR method is often too smooth and lacks details. The texture produced by the SISR process of the GAN model is vibrant, but many artefacts will appear to affect the image quality. With the development of the SISR method, we can see from Table 1 that the SISR model is developing towards a lightweight model. This is also in line with our current needs for mobile devices and the Internet of Things (IoT). Besides, through the analysis of the network structure of the SISR model, we found that the SISR model was quickly applied to the SISR field with the introduction of the network structure, such as the application of network structures such as VGGNet [56], ResNet [49], and DenseNet [67]. At the same time, with the development of SISR, substantial changes have also taken place in the training data set. Looking at Table 1, we can see that the training data set is moving towards the real data set, without using downsampling to generate the corresponding data set. This shows that we pay more attention to the low-scoring images in the real world, which can help the implementation of the training model in real life. Besides, we can find that with the development of SISR, the training data set is gradually becoming richer. Most of the datasets used in the early days of SISR were low-quality image datasets, such as ImageNet subset, '91 images' and '200 images'. The current SISR training data set has been changed to use very rich DIV2K [68] and Flickr2K [69]. This is because the rapid development of machine hardware has increased computing power, and more complex images are required for learning due to more complex scene requirements.

**Table 1** Comparison of features of characteristics models

| Model | Structure | Parameters | Train data | Years |
|---|---|---|---|---|
| bicubic | — | — | — | — |
| DBI | bicubic | — | — | 2013 |
| DWT-BI | wavelet transform | — | — | 2013 |
| SRCNN | CNN | 8.032 K | ImageNet subset | 2015 |
| SCN [62] | CNN&Linear | 31 K | 91 images | 2015 |
| FSRCNN [63] | CNN | 3.937 K | General-100 | 2016 |
| VDSR [62] | VGG-Net | 665 K | 291 images | 2016 |
| DRCN | RNN & ResNet | 1.77M | 91 images | 2016 |
| SelfExSR [63] | — | — | Urban 100 | 2017 |
| LapSRN [62] | ResNet | 812 K | 291 images | 2017 |
| DRRN | RNN & ResNet | 297 K | 291 images | 2017 |
| EDSR | ResNet | 43M | DIV2K | 2017 |
| SRGAN | GAN | 1.5M | DIV2K | 2017 |
| EIFG | CNN | — | — | 2017 |
| JRSR | — | — | 91 images | 2018 |
| DSRN | RNN & ResNet | 1.25M | 291 images | 2018 |
| LRFNet-S [7] | ResNet | 1.086M | DIV2K | 2018 |
| ProSR | DenseNet | 1.89M | DIV2K | 2018 |
| MSRN [64] | ResNet | 6.3M | DIV2K | 2018 |
| RDN | ResNet | 22.6M | DIV2K | 2018 |
| RCAN | ResNet | 15.6M | DIV2K | 2018 |
| CARN | ResNet | 1.592M | DIV2K | 2018 |
| ESRGAN | DenseNet | — | DIV2K & Flickr2K | 2018 |

**Table 2** Results of the tested models on multiple test sets

| Datasets | Set5 [70] | | Set14 [71] | | BSDS100 [72] | | Urban100 [73] | | Manga109 [74] | |
|---|---|---|---|---|---|---|---|---|---|---|
| methods | PSNR | SSIM | PSNR | SSIM | PSNR | SSIM | PSNR | SSIM | PSNR | SSIM |
| Bicubic | 28.42 | 0.8104 | 26 | 0.7027 | 25.69 | 0.6675 | 23.14 | 0.6577 | 24.89 | 0.7866 |
| DBI | 28.97 | 0.8213 | 26.75 | 0.7088 | 25.83 | 0.6728 | 23.55 | 0.6609 | 24.92 | 0.7877 |
| DWT-BI | 29.92 | 0.8562 | 26.97 | 0.7502 | 26.1 | 0.6895 | 23.90 | 0.6972 | 26.18 | 0.8022 |
| SRCNN | 30.48 | 0.8628 | 27.5 | 0.7513 | 26.9 | 0.7101 | 24.53 | 0.7221 | 27.58 | 0.8555 |
| FSRCNN [63] | 30.72 | 0.8666 | 27.61 | 0.7555 | 26.98 | 0.715 | 24.62 | 0.728 | 27.9 | 0.861 |
| SelfExSR [73] | 30.33 | 0.8611 | 27.54 | 0.7563 | 26.84 | 0.716 | 24.82 | 0.7401 | 28.1 | 0.8632 |
| SCN [62] | 30.39 | 0.8623 | 27.48 | 0.7512 | 26.87 | 0.712 | 24.52 | 0.7552 | 27.91 | 0.8602 |
| VDSR [75] | 31.35 | 0.882 | 28.02 | 0.7681 | 27.29 | 0.0711 | 25.18 | 0.751 | 28.83 | 0.887 |
| JRSR | 31.32 | 0.8809 | 27.99 | 0.7702 | 27.11 | 0.7128 | 25.01 | 0.7434 | 28.08 | 0.8845 |
| DRCN | 31.53 | 0.8841 | 28.04 | 0.7704 | 27.24 | 0.7243 | 25.14 | 0.7518 | 28.99 | 0.8891 |
| LapSRN [76] | 31.54 | 0.8855 | 28.19 | 0.7721 | 27.32 | 0.7286 | 25.21 | 0.7562 | 29.02 | 0.89 |
| DRRN | 31.68 | 0.8891 | 28.19 | 0.7721 | 27.38 | 0.7281 | 25.44 | 0.7643 | 29.33 | 0.8201 |
| DSRN | 31.4 | 0.8332 | 28.07 | 0.7702 | 27.25 | 0.7241 | 25.08 | 0.7472 | 29 | 0.8897 |
| LRFNet-S [7] | 31.91 | 0.8901 | 28.44 | 0.7789 | 27.47 | 0.7334 | 25.7 | 0.7736 | 30.5 | 0.8991 |
| EDSR | 32.46 | 0.8968 | 28.8 | 0.7876 | 27.71 | 0.742 | 26.64 | 0.8033 | 31.02 | 0.9148 |
| EIFG | 32.49 | 0.8982 | 28.76 | 0.7721 | 26.82 | 0.7065 | 26.78 | 0.8145 | 30.22 | 0.8976 |
| ProSR | 32.61 | 0.8966 | 28.94 | 0.7881 | 27.79 | 0.7511 | 26.89 | 0.8211 | 31.22 | 0.9221 |
| MSRN [64] | 32.07 | 0.8903 | 28.6 | 0.7751 | 27.52 | 0.7273 | 26.04 | 0.7896 | 30.17 | 0.9034 |
| RDN | 32.47 | 0.899 | 28.81 | 0.7871 | 27.72 | 0.7419 | 26.61 | 0.8028 | 31 | 0.9151 |
| RCAN+ | 32.73 | 0.9013 | 28.98 | 0.791 | 27.85 | 0.7455 | 27.1 | 0.8142 | 31.65 | 0.9208 |
| CARN | 33.01 | 0.9211 | 29.81 | 0.8011 | 27.98 | 0.7522 | 27.51 | 0.8343 | 31.98 | 0.9402 |
| SRGAN | 29.4 | 0.8472 | 26.02 | 0.7397 | 25.16 | 0.6688 | — | — | — | — |
| ESRGAN | 32.73 | 0.9011 | 28.99 | 0.7917 | 27.85 | 0.7455 | 27.03 | 0.8153 | 31.66 | 0.9196 |
| Meta-SR | — | — | 28.84 | 0.7872 | 27.75 | 0.7423 | — | — | 31.03 | 0.9154 |

### *7.2 Quantitative comparison*

We selected some models to be tested on multiple test sets. The CPU of the test server environment is Intel Core i9, and the GPU is NVIDIA Tesla P100. The results are shown in Table 2.

The models in the above table are generally arranged in chronological order, and it can be seen that some new models tend to have better test results. Bicubic is an interpolation-based non-deep learning method with a PSNR value that ranks lowest in all test sets. DBI also belongs to a type of interpolation method, which is to improve the edge texture and sharpness of interpolation according to the intensity and direction of image energy by the bicubic method. Through the results of PSNR and SSIM, it is found that there is a certain improvement over bicubic. Because its principle is similar to Bicubic, the improvement effect is not apparent. DWT-BI is a traditional wavelet transform filter to generate a sub-band image and then combines it with a LR image to restore the image texture. We can find that the interpolation-based method cannot learn the deeper features of the image, so the results cannot achieve the effect of the deep model. The regularisation method is mostly used as an objective function to optimise the edge texture and details of the image, so this method can often be combined with advanced techniques to get better results. SRCNN is the first model of deep learning applications in the SR field, which has a significant improvement over bicubic. FSRCNN is an improvement based on the SRCNN, whose PSNR value has been slightly improved and the training speed has been greatly improved, which is more importantly. SelfExSR is a non-deep learning method based on self-similarity, and its result is better than bicubic but not as good as other deep learning methods. The advantage is that there is no need to use external training samples. SCN uses a sparse coding model based on SRCNN, which reduces the complexity of the algorithm. VDSR uses the method of residual learning to deepen the network structure, which greatly improves the test results, but the operation efficiency is lower because the network is too deep. The DRCN uses a RRN and also uses the idea of residual learning to achieve some improvement in the test results. Both DRRN and DSRN are the idea of combining the residual network and the recursive network, and the results are similar to other models in the same period. LRFNet-S uses a large receiving field network to achieve large-scale image super-resolution, with a certain improvement in PSNR values. SRGAN is the first SR method to use the GAN model. It is not optimised for PSRN values, but focuses on the detailed texture of the image, which is good in perception. EDSR uses the generation network SRResNet in SRGAN and removes the BN layer on the basis of it, so that the memory resources are saving and the result has been greatly improved. LapSRN, MSRN, and ProSr are dedicated to multi-scale super-resolution reconstruction. Among them, LapSRN uses a step-by-step upsampling method with a faster running speed, MSRN uses a residual network, ProSR uses the GAN model, and they are all perform well on PSRN. RDN proposed RDB to mine richer features, and its test results ranked higher. RCAN proposed RIR to help train the deep model, and the test results performed very well. CARN is a lightweight, deep learning model that performs best in our small sample tests. ESRGAN is based on the improvement of SRGAN, which eliminates the artefacts of the original model, achieves better visual effects, and its performance on PSNR is also good. While Meta-SR may not be as effective as other advanced SR methods, this method is a multi-scale SR model that can be applied to any SR task with magnification.

### *7.3 Qualitative comparison*

We selected two images for testing. The scenes of the two images are simpler and the other one is more complicated. The result is shown in Fig. 28. First focus on the test results of simple scene pictures. In the interpolation-based method, compared to the traditional BI method, the directional BI (DBI) has a significant improvement in image sharpness, and the method based on DWT and BI (DWT–BI) performs better in the detail position of the window railing, but the details appear slightly distorted, and two methods based on regularisation have achieved good details but also appeared some redundant details, the method based on joint regularisation (JRSR) can even see obvious warped textures. In the CNN-based method, except for the original SRCNN and FSRCNN, which are relatively fuzzy, the result pictures of most CNN-based models are relatively clear. CARN as a lightweight model has achieved comparable performance with most CNN-based models,

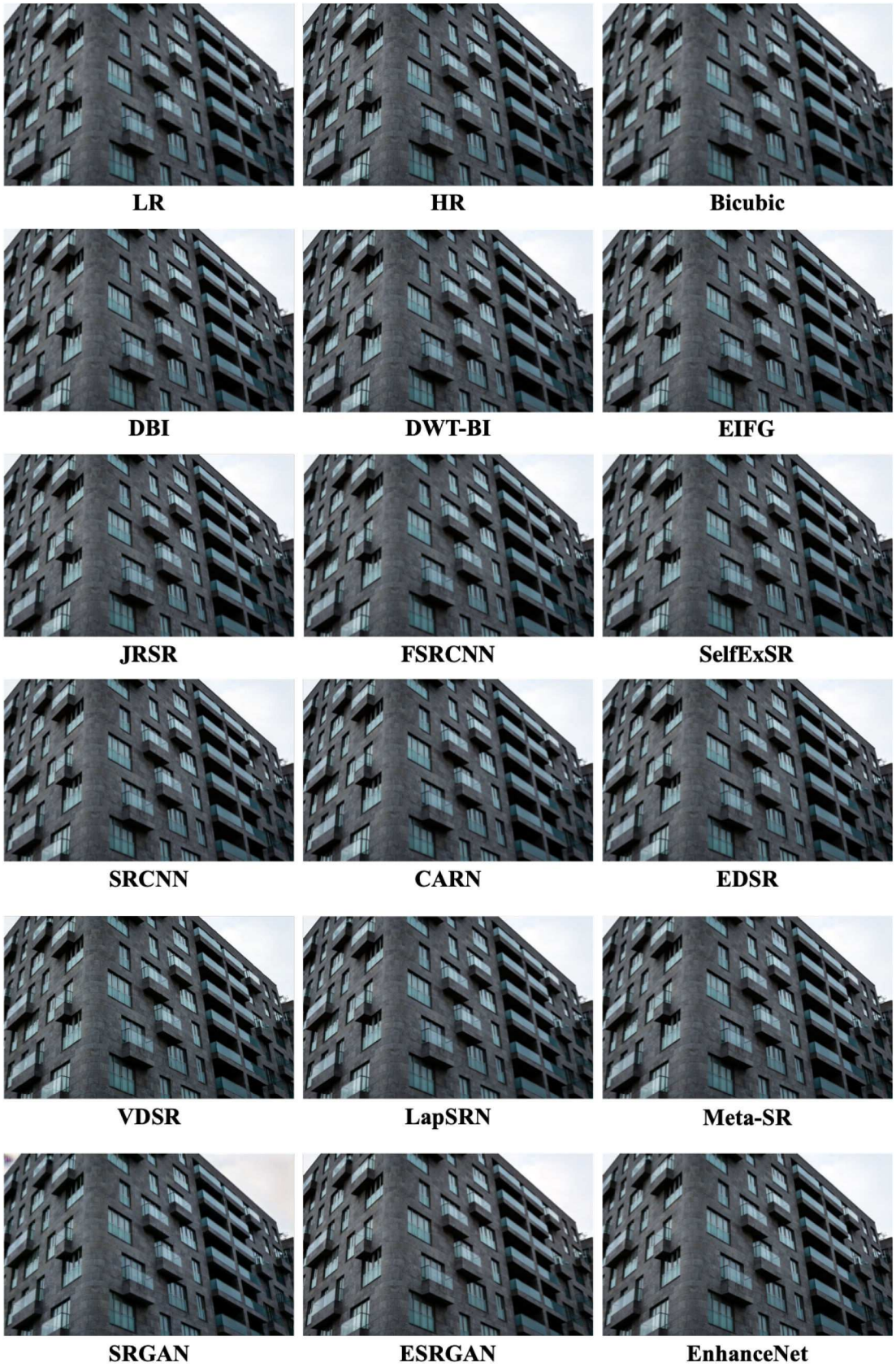


**Fig. 28** *Test results for a simple texture image*

but like other CNN-based models, it has the problem of being too smooth and lacking details. The result of VDSR is very sharp, and the overall image is very clear. Even the rightmost branch of the image can be clearly restored, but misalignment will occur in the details. The GAN-based model is dedicated to obtaining images that match the look and feel of the human eye. Therefore, the obtained images have good details and textures, but some locations will be distorted. As the first method to apply GAN to the SR field, SRGAN did not successfully resolve artefacts appearing at the edges of the image. In the resulting image of ESRGAN, not only artifacts are eliminated, but distortion is also reduced, but some locations will have unwanted textures. In the test results of complex scene images, similarly, DBI and DWT–BI have obtained clearer results images than the original BI, but they and two regularisation-based methods have limited ability to restore the details of complex scenes, which are close to SRCNN and


**Fig. 29** *Test results for complex image textures*

FSRCNN levels. Most of the CNN-based methods are smeared (see Fig. 29). The CARN image is distorted in the details at the top of the remote building. The restoration of the building's window railings is also not good. The VDSR has rich details and can clearly restore the window railings. However, because the picture is more complicated overall, the effect of excessive sharpening is a bit messy, especially at the branches. The result image based on the GAN method is rich in details, but the distortion is more obvious in complex scenes, and SRGAN still has strange artefacts. ESRGAN's performance is still good, and even the details of the windows of the remote building are restored. The disadvantage is still that some locations will have unwanted textures.

### 7.4 Prospects and opinions

In view of the future development in the field of SISR reconstruction, we put forward a few prospects here:

1. *Design smaller models to maintain performance while reducing run time*. With the introduction of ResNet, the depth of the network has been deepened again and again, while the performance has been improved, it also has some problems, such as excessive training time and excessive memory consumption. Considering the future development in practical applications, it is necessary to design a lightweight model that balances performance and consumption.
2. *Find the loss function that is most suitable for image super-resolution reconstruction.* Most of the loss functions used in current image super-resolution reconstruction work are inherited from previous work in other image fields. In practical applications, the appropriate loss function is often selected through experience. Therefore, finding a loss function dedicated to image super-resolution reconstruction is of great significance for improving model performance.
3. *The BN layer is gradually deprecated in the SR field, and new standardised technologies need to be proposed.* BN has been widely used in the image field as a standardised technology, but its performance on SR is not satisfactory, bringing a lot of memory consumption. In EDSR, the authors removed the BN layer to improve model performance. Since then, the BN layer has been gradually deprecated in the SR field. Therefore, new standardisation technologies are needed.
4. *Find more accurate image evaluation methods.* This paper introduces four evaluation indicators. Most CNN-based methods use PSNR and SSIM as evaluation indicators. In the work of SRGAN, it is pointed out that the images with high PSNR are often too smooth and do not meet the real perception of people. Therefore, the perceptual loss is proposed. So, there is still no unified standard in actual research. Finding a more accurate evaluation method to unify the evaluation standard is very important for future work development.
5. *Extend the application of SR in practical scenarios.* Most of the current SR-related work is based on the training of a large number of paired data sets, which does not meet the application requirements in actual scenarios. Therefore, in recent years, many researchers have begun to propose some unsupervised SR methods, which should also be a future research focus.
6. *Transfer from image super-resolution to video super-resolution reconstruction.* At present, a series of results have been achieved in the field of image super-resolution. Researchers have gradually turned their attention to the area of video super-resolution. Compared with image super-resolution, video super-resolution technology is more complicated. It not only needs to generate frame-by-frame images with vibrant details but also to maintain coherence between images. But at the same time, it also has considerable application value and a wide range of application scenarios.

## 8 Conclusion

This paper studies the related work in the field of image super-resolution reconstruction. Firstly, several evaluation indexes, PSNR, SSIM, PI and RMSE of image super-resolution reconstruction work are introduced. Then, according to the model-based categories, the interpolation-based, regularisation-based, CNN-based and GAN-based models are introduced, respectively. Based on the SISR model of the interpolation method, we introduce two more classic models from the interpolation method and filter, respectively. The method of regularisation is mainly introduced by introducing two models that add a regularisation loss function, and then experimentally observe the texture and details generated by the method. In the CNN-related models, since the birth of the SRCNN, the depth of the CNN has become larger and larger with the introduction of the ResNet network. In recent years, researchers have pursued better performance while taking into account the complexity of structures and algorithms, and a lightweight network model presented. The GAN-based models do not pursue a better PSNR than the CNN-based ones with considering that the reconstructed image will be too smooth and lacks detail. Therefore, it is proposed to apply the perceived loss to the network to make the constructed image have excellent texture and detail. However, the original model is prone to distortion and accompanied by artefacts. ESRGAN eliminates artefacts by removing the BN layer, which improves the perception overall. Finally, we looked forward to the development prospects of SISR reconstruction from aspects of model design, evaluation methods, performance optimisation, and application development, and put forward some opinions and suggestions.

## 9 Acknowledgments

The authors were grateful to the reviewers for valuable comments that have greatly improved the paper. This paper was partially supported by the National Natural Science Foundation of China (grant nos. 61762074, 61962051), National Natural Science Foundation of Qinghai Province (grant nos. 2019-ZJ-7034, 2020-ZJ-943Q). 'Qinghai Province High-end Innovative Thousand Talents Program - Leading Talents'. Project Support, National College Students Innovation and Entrepreneurship Training Program (grant no. 201910743002), The Open Project of State Key Laboratory of Plateau Ecology and Agriculture, Qinghai University (grant no. 2020-ZZ-03).

## 10 References

[1] Patil, V.H., Bormane, D.S.: 'Interpolation for super resolution imaging'. Innovations and Advanced Techniques in Computer and Information Sciences and Engineering, Bridgeport, CT, USA, 2007, pp. 483–489
[2] Siu, W.C., Hung, K.W.: 'Review of image interpolation and super-resolution'. Proc. of The 2012 Asia Pacific Signal and Information Processing Association Annual Summit and Conf., Hollywood, CA, USA, 2012, pp. 1–10
[3] Yu, L., Cao, S., He, J., *et al.*: 'Single-image super-resolution based on regularization with stationary gradient fidelity'. 2017 10th Int. Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI), Shanghai, People's Republic of China, 2017, pp. 1–5
[4] Milanfar, P.: '*Super-resolution imaging*' (CRC Press, USA, 2010)
[5] Simonyan, K., Zisserman, A.: 'Very deep convolutional networks for large-scale image recognition'. arxiv [cs. cv]. 2014, 2018
[6] Han, W., Chang, S., Liu, D., *et al.*: 'Image super-resolution via dual-state recurrent networks'. Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 2018, pp. 1654–1663
[7] Seif, G., Androutsos, D.: 'Large receptive field networks for high-scale image super-resolution'. Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition Workshops, Salt Lake City, UT, USA, 2018, pp. 763–772
[8] Lim, B., Son, S., Kim, H., *et al.*: 'Enhanced deep residual networks for single image super-resolution'. Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition Workshops, Honolulu, HI, USA, 2017, pp. 136–144
[9] Ledig, C., Theis, L., Huszár, F., *et al.*: 'Photo-realistic single image super-resolution using a generative adversarial network'. Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 2017, pp. 4681–4690
[10] Fan, Y., Yu, J., Huang, T.S.: 'Wide-activated deep residual networks based restoration for bpg-compressed images'. The IEEE Conf. on Computer Vision and Pattern Recognition (CVPR) Workshops, Salt Lake City, UT, USA, 2018, pp. 2621–2624
[11] Zhang, Y., Tian, Y., Kong, Y., *et al.*: 'Residual dense network for image super-resolution'. Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 2018, pp. 2472–2481
[12] Ahn, N., Kang, B., Sohn, K.A.: 'Fast, accurate, and lightweight super-resolution with cascading residual network'. Proc. of the European Conf. on Computer Vision (ECCV), Munich, Germany, 2018, pp. 252–268
[13] Zhang, Z., Wang, Z., Lin, Z., *et al.*: 'Image super-resolution by neural texture transfer'. Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 2019, pp. 7982–7991
[14] Zhang, Y., Li, K., Li, K., *et al.*: 'Image super-resolution using very deep residual channel attention networks'. Proc. of the European Conf. on Computer Vision (ECCV), Munich, Germany, 2018, pp. 286–301
[15] Wang, Y., Perazzi, F., McWilliams, B., *et al.*: 'A fully progressive approach to single-image super-resolution'. Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition Workshops, Salt Lake City, UT, USA, 2018, pp. 864–873
[16] Anaya, J., Barbu, A.: 'Renoir–a dataset for real low-light image noise reduction', *J. Vis. Commun. Image Represent.*, 2018, **51**, pp. 144–154
[17] Sumiya, T., Fukuhara, M., Sato, M.: 'Tomographic image capturing apparatus and method with noise reduction technique'. Google Patents, uS Patent App. 10/126,1122018

[18] Tang, Z., Lin, Y.S., Lee, K.H., *et al.*: 'Esther: joint camera self-calibration and automatic radial distortion correction from tracking of walking humans', *IEEE Access*, 2019, **7**, pp. 10754–10766

[19] Chang, Y.: 'Research on de-motion blur image processing based on deep learning', *J. Vis. Commun. Image Represent.*, 2019, **60**, pp. 371–379

[20] Chen, C., Xiong, Z., Tian, X., *et al.*: 'Camera lens super-resolution'. Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 2019, pp. 1652–1660

[21] Xu, X., Ma, Y., Sun, W.: 'Towards real scene super-resolution with raw images'. Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 2019, pp. 1723–1731

[22] Soh, J.W., Park, G.Y., Jo, J., *et al.*: 'Natural and realistic single image super-resolution with explicit natural manifold discrimination'. Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 2019, pp. 8122–8131

[23] Zhou, R., Süsstrunk, S.: 'Kernel modeling super-resolution on real low-resolution images'. 2019 Int. Conf. on Computer Vision Conf., Seoul, Republic of Korea, 2019

[24] Zhang, L., Zhang, H., Shen, H., *et al.*: 'A super-resolution reconstruction algorithm for surveillance images', *Signal Process.*, 2010, **90**, (3), pp. 848–859

[25] Molina-Cabello, M.A., Elizondo, D.A., Luque-Baena, R.M., *et al.*: 'Foreground object detection enhancement by adaptive super resolution for video surveillance', 2019

[26] Shamsolmoali, P., Zareapoor, M., Jain, D.K., *et al.*: 'Deep convolution network for surveillance records super-resolution', *Multimedia Tools Appl.*, 2019, **78**, (17), pp. 23815–23829

[27] Chen, Y., Tai, Y., Liu, X., *et al.*: 'Fsrnet: end-to-end learning face super-resolution with facial priors'. Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 2018, pp. 2492–2501

[28] Lu, Z., Jiang, X., Kot, A.: 'Deep coupled resnet for low-resolution face recognition', *IEEE Signal Process. Lett.*, 2018, **25**, (4), pp. 526–530

[29] Peng, Y., Spreeuwers, L.J., Veldhuis, R.N.: 'Low-resolution face recognition and the importance of proper alignment', *IET Biometrics*, 2019, **8**, (4), pp. 267–276

[30] Truong, N.Q., Nguyen, P.H., Nam, S.H., *et al.*: 'Deep learning-based super-resolution reconstruction and marker detection for drone landing', *IEEE Access*, 2019, **7**, pp. 61639–61655

[31] Durova, M.L., Dimou, A., Litos, G., *et al.*: 'Too many eyes: Super-recogniser directed identification of target individuals on CCTV'. 8th Int. Conf. on Imaging for Crime Detection and Prevention (ICDP 2017), Madrid, Spain, 2017, pp. 43–48

[32] Ma, J., Tao, H., Huang, P.: 'Subspace-based super-resolution algorithm for ground moving target imaging and motion parameter estimation', *IET Radar Sonar Navig.*, 2016, **10**, (3), pp. 488–499

[33] Köhler, T.: 'Multi-frame super-resolution reconstruction with applications to medical imaging', arXiv preprint arXiv:181209375, 2018

[34] Ren, S., Jain, D.K., Guo, K., *et al.*: 'Towards efficient medical lesion image super-resolution based on deep residual networks', *Signal Process. Image Commun.*, 2019, **75**, pp. 1–10

[35] Hu, X., Mu, H., Zhang, X., *et al.*: 'Meta-sr: a magnification-arbitrary network for super-resolution'. Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 2019, pp. 1575–1584

[36] Wang, W., He, Q.: 'A survey on emotional semantic image retrieval'. 2008 15th IEEE Int. Conf. on Image Processing, San Diego, CA, USA, 2008, pp. 117–120

[37] Johnson, D.H.: 'Signal-to-noise ratio', *Scholarpedia*, 2006, **1**, (12), p. 2088

[38] Lore, K.G., Akintayo, A., Sarkar, S.: 'Llnet: a deep autoencoder approach to natural low-light image enhancement', *Pattern Recognit.*, 2017, **61**, pp. 650–662

[39] Blau, Y., Mechrez, R., Timofte, R., *et al.*: 'The 2018 PIRM challenge on perceptual image super-resolution'. Proc. of the European Conf. on Computer Vision (ECCV), Munich, Germany, 2018

[40] Mittal, A., Soundararajan, R., Bovik, A.C.: 'Making a 'completely blind' image quality analyzer', *IEEE Signal Process. Lett.*, 2012, **20**, (3), pp. 209–212

[41] Willmott, C.J., Matsuura, K.: 'Advantages of the mean absolute error (mae) over the root mean square error (rmse) in assessing average model performance', *Climate Res.*, 2005, **30**, (1), pp. 79–82

[42] Liu, J., Gan, Z., Zhu, X.: 'Directional bicubic interpolation – a new method of image super-resolution'. 3rd Int. Conf. on Multimedia Technology (ICMT-13), Guangzhou, People's Republic of China, 2013

[43] Kumar, G., Singh, K.: 'Image super resolution on the basis of dwt and bicubic interpolation', *Int. J. Comput. Appl.*, 2013, **65**, (15), pp. 12–17

[44] Chang, K., Ding, P.L.K., Li, B.: 'Single image super resolution using joint regularization', *IEEE Signal Process. Lett.*, 2018, **25**, (4), pp. 596–600

[45] Goldstein, T., Osher, S.: 'The split bregman method for L1-regularized problems', *SIAM J. Imag. Sci.*, 2009, **2**, (2), pp. 323–343

[46] Dong, C., Loy, C.C., He, K., *et al.*: 'Learning a deep convolutional network for image super-resolution'. European Conf. on Computer Vision, Zurich, Switzerland, 2014, pp. 184–199

[47] Timofte, R., Agustsson, E., Van-Gool, L., *et al.*: 'NTIRE 2017 challenge on single image super-resolution: methods and results'. Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition Workshops, Honolulu, HI, USA, 2017, pp. 114–125

[48] Ioffe, S., Szegedy, C.: 'Batch normalization: Accelerating deep network training by reducing internal covariate shift', arXiv preprint arXiv:150203167, 2015

[49] He, K., Zhang, X., Ren, S., *et al.*: 'Deep residual learning for image recognition'. Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 2016, pp. 770–778

[50] Agustsson, E., Timofte, R.: 'NTIRE 2017 challenge on single image super-resolution: dataset and study'. 2017 IEEE Conf. on Computer Vision and Pattern Recognition Workshops (CVPRW), Honolulu, HI, USA, July 2017

[51] Shi, W., Caballero, J., Huszár, F., *et al.*: 'Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network'. Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 2016, pp. 1874–1883

[52] Salimans, T., Kingma, D.P.: 'Weight normalization: a simple reparameterization to accelerate training of deep neural networks'. Advances in Neural Information Processing Systems, Barcelona, Spain, 2016, pp. 901–909

[53] Kim, J., Kwon-Lee, J., Mu-Lee, K.: 'Deeply-recursive convolutional network for image super-resolution'. Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 2016, pp. 1637–1645

[54] Tai, Y., Yang, J., Liu, X.: 'Image super-resolution via deep recursive residual network'. Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 2017, pp. 3147–3155

[55] Hui, Z., Wang, X., Gao, X.: 'Fast and accurate single image super-resolution via information distillation network'. Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 2018, pp. 723–731

[56] Simonyan, K., Zisserman, A.: 'Very deep convolutional networks for large-scale image recognition', arXiv preprint arXiv:14091556, 2014

[57] Wang, X., Yu, K., Wu, S., *et al.*: 'Esrgan: enhanced super-resolution generative adversarial networks'. Proc. of the European Conf. on Computer Vision (ECCV), Munich, Germany, 2018

[58] Jolicoeur-Martineau, A.: 'The relativistic discriminator: a key element missing from standard gan', arXiv preprint arXiv:180700734, 2018

[59] Yuan, Y., Liu, S., Zhang, J., *et al.*: 'Unsupervised image super-resolution using cycle-in-cycle generative adversarial networks'. Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition Workshops, Salt Lake City, UT, USA, 2018, pp. 701–710

[60] Zhu, J.Y., Park, T., Isola, P., *et al.*: 'Unpaired image-to-image translation using cycle-consistent adversarial networks'. Proc. of the IEEE Int. Conf. on Computer Vision, Honolulu, HI, USA, 2017, pp. 2223–2232

[61] Park, S.J., Son, H., Cho, S., *et al.*: 'SRFeat: single image super-resolution with feature discrimination'. Proc. of the European Conf. on Computer Vision (ECCV), Munich, Germany, 2018, pp. 439–455

[62] Wang, Z., Liu, D., Yang, J., *et al.*: 'Deep networks for image super-resolution with sparse prior'. Proc. of the IEEE Int. Conf. on Computer Vision, Boston, MA, USA, 2015, pp. 370–378

[63] Dong, C., Loy, C.C., Tang, X.: 'Accelerating the super-resolution convolutional neural network'. European Conf. on Computer Vision, Las Vegas, NV, USA, 2016, pp. 391–407

[64] Li, J., Fang, F., Mei, K., *et al.*: 'Multi-scale residual network for image super-resolution'. Proc. of the European Conf. on Computer Vision (ECCV), Munich, Germany, 2018, pp. 517–532

[65] Yang, J., Wright, J., Huang, T.S., *et al.*: 'Image super-resolution via sparse representation', *IEEE Trans. Image Process.*, 2010, **19**, (11), pp. 2861–2873

[66] Martin, D., Fowlkes, C., Tal, D., *et al.*: 'A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics'. IEEE Int. Conf. on Computer Vision (ICCV), Vancouver, 2001

[67] Huang, G., Liu, Z., Van Der Maaten, L., *et al.*: 'Densely connected convolutional networks'. Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 2017, pp. 4700–4708

[68] Agustsson, E., Timofte, R.: 'NTIRE 2017 challenge on single image super-resolution: dataset and study'. The IEEE Conf. on Computer Vision and Pattern Recognition (CVPR) Workshops, Honolulu, HI, USA, 2017

[69] Lim, B., Son, S., Kim, H., *et al.*: 'Enhanced deep residual networks for single image super-resolution'. The IEEE Conf. on Computer Vision and Pattern Recognition (CVPR) Workshops, Honolulu, HI, USA, 2017

[70] Bevilacqua, M., Roumy, A., Guillemot, C., *et al.*: 'Low-complexity single-image super-resolution based on nonnegative neighbor embedding'. Proc. of the British Machine Vision Conf., BMVA Press, 2012, pp. 135.1–135.10

[71] Zeyde, R., Elad, M., Protter, M.: 'On single image scale-up using sparse-representations'. Int. Conf. on Curves and Surfaces, Arcachon, France, 2010, pp. 711–730

[72] Arbelaez, P., Maire, M., Fowlkes, C., *et al.*: 'Contour detection and hierarchical image segmentation', *IEEE Trans. Pattern Anal. Mach. Intell.*, 2010, **33**, (5), pp. 898–916

[73] Huang, J.B., Singh, A., Ahuja, N.: 'Single image super-resolution from transformed self-exemplars'. Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition, Boston, MA, USA, 2015, pp. 5197–5206

[74] Matsui, Y., Ito, K., Aramaki, Y., *et al.*: 'Sketch-based manga retrieval using manga109 dataset', *Multimedia Tools Appl.*, 2017, **76**, (20), pp. 21811–21838

[75] Kim, J., Kwon Lee, J., Mu Lee, K.: 'Accurate image super-resolution using very deep convolutional networks'. The IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 2016

[76] Lai, W.S., Huang, J.B., Ahuja, N., *et al.*: 'Fast and accurate image super-resolution with deep laplacian pyramid networks', *IEEE Trans. Pattern Anal. Mach. Intell.*, 2019, **41**, pp. 2599–2613