Research on Image Super-resolution Reconstruction Based on Deep Learning

Jingyu Jiang
School of Computer Science and Engineering
Xi'an Technological University
Xi'an, China
E-mail: jjy1030@126.com;

Yan Jiao
School of Computer Science and Engineering
Xi'an Technological University
Xi'an, China
E-mail: jiaoyan@st.xatu.edu.cn;

Li Zhao
School of Computer Science and Engineering
Xi'an Technological University
Xi'an, China
E-mail: 332099732@qq.com;

***Abstract*—Image super-resolution reconstruction (SR) aims to use a specific algorithm to restore a low-resolution blurred image in the same scene into a high-resolution clear image. Due to its wide application value and theoretical value, image super-resolution reconstruction technology has become a research hotspot in the field of computer vision and image processing, and has attracted widespread attention from researchers. Compared with traditional methods, deep learning methods have shown better reconstruction effects in the field of image super-resolution reconstruction, and have gradually developed into the mainstream technology. Therefore, this paper classifies the image super-resolution reconstruction problem systematically according to the structure of the network model, and divides it into two categories: the super-division method based on the convolutional neural network model and the super-division method based on the generative confrontation network model. The main image super-resolution reconstruction methods are sorted out, several more important deep learning super-resolution reconstruction models are described, the advantages and disadvantages of different algorithms and the applicable application scenarios are analyzed and compared, and the different types of super-resolution algorithms are discussed. The method of mutual fusion and image and video quality evaluation, and a brief introduction to commonly used data sets. Finally, the potential problems faced by the current image super-resolution reconstruction technology are discussed, and a new outlook for the future development direction is made.**

Keywords-Image Super-Resolution Reconstruction; Deep Learning; Convolutional Neural Network; Generative Confrontation Network

I. INTRODUCTION

Image super-resolution reconstruction (SR) refers to the use of image processing and machine learning methods to reconstruct one or more low-resolution (LR) images in the same scene

DOI: 10.2478/ijanmc-2022-0001

with rich image details and High-Resolution (HR) image process with clear texture [1]. It has important application value in the fields of video, remote sensing, medicine and security monitoring. With the rapid development of machine learning in the field of computer vision, deep learning technology has been widely used to solve SR problems and achieved good reconstruction results, and has gradually become the mainstream.

Existing super-resolution reconstruction algorithms are usually divided into three categories: interpolation-based methods, which are simple but provide too smooth reconstructed images, lose some details and produce ringing effects; methods based on modeling, Compared with the interpolation method, this type of algorithm has a better reconstruction effect, but when faced with a large amount of calculation, the calculation process takes a long time, is difficult to solve and is greatly affected by the amplification factor; based on the learning method, this type of algorithm solves the problem of The scale factor is sensitive to the problem and the reconstruction effect is the best, which is the mainstream direction of current research [3].

Convolutional neural network (CNN) and generative adversarial network (GAN) are the current mainstream network models. When the scaling factor is large, using these two network models can restore the height of the image very well. Frequency information to make the output image closer to the original real image [3].

II. NETWORK MODEL OF SUPER-RESOLUTION RECONSTRUCTION METHOD

According to the different network model structure, the image super-resolution reconstruction method based on deep learning can be divided into the following two categories: ① Super-segmentation method based on Convolutional Neural Network (CNN) model; ② Based on generative adversarial network (Generative Adversarial Networks, GAN) model super-division method. In response to various requirements for image super-resolution, super-resolution network models with various characteristics have been produced (Figure 1) [4].

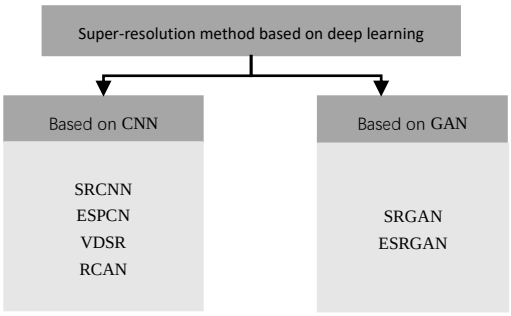

Figure 1. Super-resolution method based on deep learning

A. Super-division method based on CNN network model

1) SRCNN

SRCNN [5] uses the relationship between deep learning and traditional sparse coding as a basis, and divides the 3-layer network into feature extraction, nonlinear mapping, and final reconstruction. For a low-resolution image, as shown in Figure 2, the method first uses bicubic interpolation to enlarge it to the same size as the target, and then extracts and represents the image block, and then uses a three-layer convolutional network to make nonlinear The result of mapping and reconstruction is output as a high-resolution image. SRCNN introduces convolutional neural networks to SR tasks for the first time. Unlike the step-by-step processing of traditional SR algorithms, SRCNN integrates various stages into a deep learning model, which greatly simplifies the SR workflow and can be regarded as a super deep learning based Milestones in resolution methods [6].

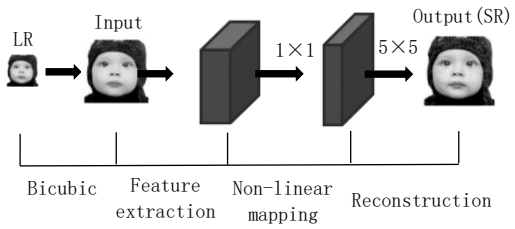

Figure 2. SRCNN network model

Although the network structure of SRCNN is simple in design and is superior to traditional super-resolution algorithms in terms of image reconstruction quality and speed, it has the following problems: ①Does not use any prior knowledge; ②It is only suitable for SR tasks with a single magnification factor. For different magnification factors, the network model needs to be trained again; ③Because the input image needs to be interpolated and magnified to the same size as the target, the entire image reconstruction process is performed in the HR space, which takes up a lot of memory space and increases the amount of calculation. , The error produced by the interpolation process will also have an impact on the reconstruction effect, the model convergence speed is slower, and the training time is longer; ④The number of network layers is less, and the receptive field of the convolution kernel is also small (13 × 13). Good application of image context-related information, resulting in unclear texture of the final reconstructed HR image and limited algorithm adaptability [7].

Initially, the smaller datasets Set5 and Set14 were used to train the SRCNN algorithm. After training, the knowledge learned is relatively small, and the image reconstruction performance is constrained. When the relatively large dataset BSD200 is used, the reconstruction performance is significantly improved. The reconstruction performance of the image is also greatly affected by the size of the data set [8].

After that, although experts have proposed various network models, SRCNN is still used as a benchmark experiment for evaluating the performance of other network models.

2) (2) ESPCN

In 2016, Shi et al. proposed an Efficient Sub-Pixel Convolutional Neural Network (ESPCN) network model based on pixel rearrangement [9], The core concept of ESPCN is a sub-pixel convolutional layer, which performs a convolution operation on the LR image to obtain LR image features, and then expands the features in the LR space to the HR space through the sub-pixel convolutional layer. Reorganize the HR feature map obtained after convolution to obtain an HR image [10].

The ESPCN network mainly improves the reconstruction layer of SRCNN. The LR image is used as the network input. The sub-pixel convolutional layer is used in the reconstruction layer to double the network training speed. The simple network structure and extremely high reconstruction speed make it very It is suitable for high-speed real-time systems that require relatively low reconstruction performance. In the ESPCN network, the interpolation function used for image size enlargement is implicitly included in the previous convolutional layer, which can be automatically learned. Since the convolution operation is performed on the size of the low-resolution image, the model efficiency is higher. The sub-pixel convolutional layer proposed by the ESPCN model has been widely used later. Compared with the deconvolutional layer proposed in the FSRCNN model, the learned nonlinear effect of upsampling from low-resolution images to high-resolution images is better. It is worth noting that the model also modified the activation function, replacing the

ReLU function with the tanh function, and the loss function is the mean square error.

3) (3) VDSR

VDSR (Very Deep CNN for SR) [11] is the first deep model that proposes to use the global residual learning idea to solve the image SR problem. It is an improved network based on SRCNN. It uses a multi-layer convolution kernel for deep convolution, which not only reduces the amount of parameters, but also makes the following the network layer has a larger receptive field, can better utilize the image context information of a larger area, and obtain a better reconstruction effect than SRCNN. The biggest feature of VDSR is the deep network layers, good image reconstruction effect, and faster training speed. Since the author found that the input LR image is very similar to the output HR image, that is, the low frequency information carried by the LR image is very similar to the low frequency information of the HR image [12], So only need to learn the high-frequency residual part between the HR image and the LR image. The VDSR network structure is shown in Figure 3. It will interpolate to obtain an LR image of the same size as the target and input it into the network, and then add this image and the residual error learned by the network to obtain the final HR reconstructed image. The adaptive gradient pruning strategy is to train the network with a higher learning rate. Although the architecture is huge, it can still speed up the convergence. Therefore, on the basis of increasing the network depth, combining residual network and adaptive gradient cropping to accelerate model training can improve network performance and achieve better reconstruction effects. At the same time, through mixed training of images of different scales, the VDSR network can achieve a single Multi-scale SR reconstruction of the model.

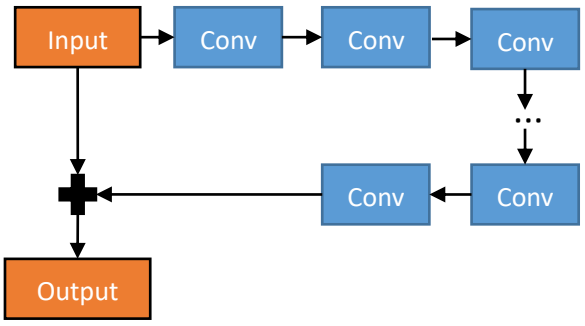

Figure 3. VDSR network structure

4) RCAN

In 2018, Zhang et al. [13] It is believed that the input image contains a large amount of low-frequency information, and the existing SR network treats all channels of this information equally, and lacks the ability to distinguish and learn these channels, which hinders the network's characterization ability. Therefore, the deep residual channel attention network RCAN is proposed [13]. RCAN is the first network to apply the attention mechanism to SR problems. The algorithm obtains a weight value by learning the importance of different channels, which is equivalent to modeling the relationship between channel features, adaptively adjusting each channel feature, thereby effectively strengthening useful feature channels while suppressing useless feature channels, to make fuller use of computing resources. The model uses a locally nested residual structure (residual in residual), which is composed of a residual group (RG) and a long jump connection (LSC). A deeper network is built by simply stacking residual blocks and passes between feature channels the dependence relationship of the selection contains more key information feature channels, and enhances the identification and learning ability of the entire network.

B. Super-division method based on GAN network model

1) SRGAN

The Generative Adversarial Network (GAN) was proposed by Goodfellow et al. It is inspired by the two-person zero-sum game in game theory. The two players in the GAN model are respectively composed of a generative model and a discriminant model (discriminative model) as [14]. SRGAN first applied adversarial training to the problem of image super-resolution reconstruction. The results show that the introduction of adversarial training can enable the network to generate finer texture details. GAN can complete many incredible generation problems, in the fields of image generation, speech conversion, and text generation. Occupy a very important position. As shown in Figure 4, SRGAN inputs the LR image to the generator G for image reconstruction. The discriminator D will train the generated image against the HR image, and finally output the image generated by the training [15]. The collaborative training of the generator and the discriminator enables the network to not only judge the similarity between the generated image and the actual high-resolution image in the pixel domain, but also pay more attention to its distribution similarity in the pixel space. Compared with the previous algorithm, although SRGAN is relatively low in objective evaluation indicators (such as PSNR), it has a better reconstruction effect in visual effects, image details and other intuitive aspects. This is related to its unique network structure and the loss function that combines perceptual loss and adversarial loss. Perceptual loss is a feature extracted by using convolutional neural networks. The generated image is compared with the target image. The feature difference after the convolutional neural network makes the generated picture and the target picture more similar in semantics and style. The confrontation loss is provided by GAN, and the network is trained according to whether the image can be successfully deceived. SRGAN is a milestone in the pursuit of the development of visual experience. SRGAN has significantly improved the overall visual quality of PSNR-based reconstruction.

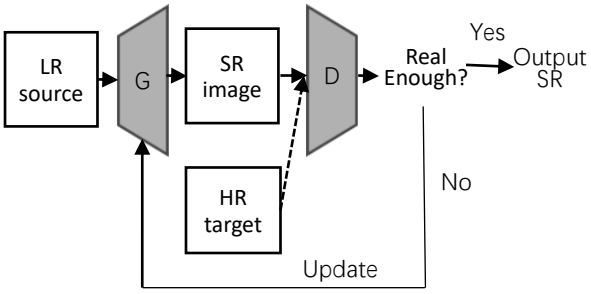

Figure 4. Schematic diagram of the basic structure of SRGAN

2) ESRGAN

In order to improve the reconstruction accuracy of the SRGAN model, on this basis, ESRGAN [15] model has been improved in three aspects: the generator architecture is changed, the batch normalization (BN) layer is removed to reduce the artifacts generated in the reconstructed image, and a denser with higher reconstruction accuracy is introduced. Residual block (RDDB) to improve the structure of the model, make it more capacity and easier to train, help to improve generalization ability, reduce computational complexity and memory usage; improve the perceptual domain loss function, use the VGG before activation Features. This improvement will provide clear edges and more visually consistent results, which can better maintain image brightness consistency and restore better detail texture; enhance the discriminator's ability to discriminate, and use relative discriminators to make them then there is the absolute value of true and false but the relative distance between the predicted generated image and the real image.

Compared with the SRGAN model, on most standard test images, the ESRGAN model can reduce the noise in the reconstructed image while maintaining better details. A large number of experiments have shown that enhanced SRGAN, called ESRGAN, is always better than the most advanced methods in sharpness and detail.

C. Summary

In recent years, deep learning technology has developed vigorously. The super-resolution reconstruction method based on deep learning has gradually become the mainstream of image super-resolution, and a series of network models based on CNN and GAN have been developed, which are introduced in sections 2.1-2.2. The super-division method based on CNN network model and GAN network model is presented. On the whole, the reconstruction performance of the algorithm is continuously improving, and the reconstructed image is getting closer and closer to the original image. The image texture details obtained by the algorithm reconstruction based on the GAN network model are better, the realism of the image is improved, and the human eye perception effect is better. In contrast, the algorithm based on the CNN network model has lower network complexity, lower training difficulty, and higher reconstruction result accuracy, but it will produce artifacts and the reconstruction speed is slower.

III. LOSS FUNCTION CONSTRUCTION

In the field of image super-resolution, the loss function is used to measure the difference between the generated image and the actual high-resolution image. At present, in the field of super-resolution, loss functions play an important role. Commonly used loss functions include pixel-based loss functions and perception-based loss functions.

A. Pixel loss function

Pixel loss mainly measures the pixel difference between the predicted image and the target image, mainly including L1 (mean absolute error) loss and L2 (mean square error) loss.

L1 loss, also known as the minimum absolute deviation (LAD) and the minimum absolute error (LAE), calculates the sum of the absolute difference between the actual value and the target value. In the supervised image super-resolution task, the goal is to make the generated image (SR) as close to the real high-resolution image (HR) as possible, and the L1 loss is used to calculate the value between the corresponding pixel positions of SR and HR The error. The basic expression of mean absolute value error (MAE) is shown in formula (1):

$$L_{MAE}(\theta) = \frac{1}{n}\sum_{i=1}^{n}\left|Y_i - F(X_i, \theta)\right| \qquad (1)$$

Among them, L(θ) represents the loss function that the network needs to optimize, n represents the number of training samples, θ represents the parameters of the deep neural network, F(Xi, θ) is the image reconstructed by the network, and Yi represents the corresponding HR image. L2 loss, also called mean square error loss (MSE), calculates the sum of the squares of the absolute difference between the actual value and the target value, which greatly improves the performance of the image super-resolution model based on deep learning, is the most commonly used loss function. However, MSE loss can punish large losses, but can do nothing for small losses, so it will produce blurred images. Its basic expression is shown in formula (2):

$$L_{MSE}(\theta) = \frac{1}{n}\sum_{i=1}^{n}\left\|Y_i - F(X_i, \theta)\right\|^2 \qquad (2)$$

The meaning of its basic symbols is the same as in formula (1). The application of the minimum mean square error effectively solves the problem of measuring the difference between the SR reconstructed image and the target HR image, making the image SR model based on deep learning a greater improvement than the traditional learning-based SR reconstruction model.

Compared with MSE, MAE has the advantage that it is more robust to outliers and more tolerant. It can be seen from equation (1) that MAE calculates the absolute value of error Yi-F(Xi, θ), so whether it is Yi-F(Xi, θ)>1 or Yi-F(Xi, θ)<1, there is no The effect of the square term, the punishment is the same, and the weight is the same.

Since these loss functions evaluate the class prediction of each pixel vector separately and then average all pixels, they assert that each pixel in the image has the same learning ability.

Compared with the L1 loss function, the L2 loss function will amplify the gap between the maximum error and the minimum error (such as 2*2 and 0.1*0.1), and the L2 loss function is also more sensitive to outliers.

Since the definition of peak signal-to-noise ratio (PSNR) (see 4.1) has a high correlation with pixel difference, and minimizing pixel loss is equivalent to maximizing PSNR, the pixel loss function has gradually become a commonly used loss function. But because it does not explore image quality issues (e.g., perceptual quality [16], Texture detail [17]), Therefore, the results often lack high-frequency details, resulting in too smooth texture details.

B. Perceptual loss function

For tasks such as image stylization and image super-resolution reconstruction, the L2 loss in the image pixel space was used in the early days, but the L2 loss does not match the image quality of the human eye perception loss, and the restored image details often perform poorly. In today's research, L2 loss is gradually replaced by human eye perception loss. The perceptual loss of the human eye is also called perceptual loss. The difference from the MSE loss that uses image pixels for difference is that the calculated space is no longer the image space. The loss function based on perception can recover more high-frequency information and make the reconstruction performance better. At present, the loss function based on perception mainly includes content loss and counter loss.

The content loss function is divided into feature reconstruction loss function and style reconstruction loss function. Bruna et al. [18] proposed the feature reconstruction loss function for the first time. The feature maps corresponding to the reconstructed SR image and the HR image in the feature space were extracted and compared through the pre-trained VGG19 network. The expressions are as follows:

$$L_{VGG/j} = \frac{1}{W_j H_j} \sum_{h=1}^{H_j} \sum_{w=1}^{W_j} (\phi_j(Y)_{x,y} - \phi_j(G_{\theta_G}(X))_{x,y})^2 \quad (3)$$

Among them, W_j and H_j represent the width and height of the jth feature map, ϕ_j represents the feature map obtained by the jth convolution in the VGG19 network, X represents the original LR image, Y is the reconstructed HR image, G_{θ_G} epresents SR image generated by the network. Gatys et al [19] .proposed a style reconstruction loss function based on the feature reconstruction loss function. This function defines a Gram matrix. The calculation formula is as follows:

$$G_j^{\phi}(x)_{c,c'} = \frac{1}{C_j W_j H_j} \sum_{h=1}^{H_j} \sum_{w=1}^{W_j} \phi_j(x)_{h,w,c} \phi_j(x)_{h,w,c'} \quad (4)$$

The size of the feature map extracted by the VGG network is Cj×Wj×Hj. Then the Euclidean distance difference of the Gram matrix is calculated in the corresponding layer and added to obtain the style reconstruction loss function, as follows:

$$L_{style/j}^{\phi} = \left\| G_j^{\phi}(Y) - G_j^{\phi}(G_{\theta_G}(x)) \right\|^2 \quad (5)$$

The SRGAN network proposes to combat loss for the first time, and its basic form is as follows:

$$L_{Gen}^{SR} = \sum_{n=1}^{N} -lbD_{\theta_D}(G_{\theta_G}(x)) \quad (6)$$

Among them, N represents the number of images, θ D epresents the parameters of the identi-fication network, θ G represents the parameters of the generating network, and represents the probability that the generated image is a real HR image. The final optimization goal of the network is a minimum-maximization problem:

$$\min_{\theta_G} \max_{\theta_D} E_{I^{HR} \square P_{train}(I^{HR})} \left[lbD_{\theta_D}(I^{HR}) \right] + E_{I^{HR} \square P_{G}(I^{HR})} \left[1 - lbD_{\theta_D}(G_{\theta_G}(I^{LR})) \right] \quad (7)$$

Among them, P train（IHR） represents the distribution of HR images, P G（ILR） represents the distribution of original LR images. The adversarial training makes the generated SR image highly similar to the original HR image, which makes it difficult for the discriminant network to distinguish, and finally obtains a fake SR image. The discrimination target of SRGAN is whether the input image is true or not. Unlike SRGAN, ESRGAN replaces the discriminator of the discriminant network with a relative average discriminator. The discrimination target is to predict the probability that the real HR image is more realistic than the generated SR image。The discriminant network is shown in equation (8) and equation (9):

$$\begin{aligned} D(x_r) &= \sigma(C(real)) \to 1 \\ D(x_f) &= \sigma(C(fake)) \to 0 \end{aligned} \quad (8)$$

$$\begin{aligned} D_{Ra}(x_r, x_f) &= \sigma(C(real) - E\left[C(fake)\right]) \to 1 \\ D_{Ra}(x_f, x_r) &= \sigma(C(fake) - E\left[C(real)\right]) \to 0 \end{aligned} \quad (9)$$

Among them, real represents the real HR image, fake represents the generated SR image, C (real) represents the judgment result of the discriminator, and E[C (fake)] represents the average value of the judgment result of the authentication network. Among them, σ represents the Sigmoid activation function, which helps the network learn sharper edges and more texture details by improving the discriminator.

IV. IMAGE SUPER-RESOLUTION QUALITY EVALUATION

The higher the similarity between the high-resolution image reconstructed by super-resolution technology and the real high-resolution image, the better the performance of the image super-resolution algorithm. Generally speaking, two objective quantitative indicators are mainly used for evaluation, including peak signal-to-noise ratio (PSNR) [20] and structural similarity (Structual similarity, SSIM) [21]. For the fairness of comparison, all PSNR (dB) and SSIM metrics are calculated on the y-channel with the center drooping. The higher the evaluation index value, the smaller the difference between the reconstruction result and the original image and the higher the fidelity.

PSNR is one of the most commonly used image reconstruction quality evaluation methods, based on the mean square error (MSE) of the image, MSE is shown in the following equation (10):

$$MSE = \frac{1}{mn}\sum_{i=0}^{m-1}\sum_{j=0}^{n-1}\left[I(i,j) - K(i,j)\right]^2 \quad (10)$$

Among them, I represents the original high-resolution image, K represents the reconstructed image, and m and n represent the length and width of the image, respectively. PSNR is defined as follows:

$$PSNR = 10\log\left(\frac{MAX_I^2}{MSE}\right) \quad (11)$$

Among them, the logarithm base is the natural base e, which is used unless otherwise stated. MAXI is the maximum number of colors that can be represented by each pixel in the current picture, that is, the number of bits in the picture. t can be seen from equation (11) that minimizing MSE is equivalent to maximizing PSNR. PSNR calculates the error between pixels at the same position. In the calculation process, the influence of visual perception characteristics on the image quality is not considered, so occasionally, although the PSNR value is high, the image quality that people subjectively feel is not improved. Due to the inability to quantitatively analyze the image perception quality, PSNR is still the most commonly used evaluation method in the field of image super-resolution [15].

Structural similarity (Structural SIMilarity, SSIM) can simultaneously compare the similarity of image brightness, contrast, and structure. The value range of SSIM is [0,1]. The higher the value, the more similar the reconstruction result is to the original image structure.

$$\begin{aligned} l(x,y) &= \frac{2\mu_x\mu_y + c_1}{\mu_x^2 + \mu_y^2 + c_1} \\ c(x,y) &= \frac{2\sigma_x\sigma_y + c_2}{\sigma_x^2 + \sigma_y^2 + c_2'} \\ s(x,y) &= \frac{\sigma_{xy} + c_3}{\sigma_x\sigma_y + c_3} \end{aligned} \quad (12)$$

$$SSIM(x,y) = (x,y)^{\alpha}\cdot c(x,y)^{\beta}\cdot s(x,y)^{\gamma} \quad (13)$$

Among them,μx and μy are the mean values of the two images respectively,σx and σy respectively represent the variances of the two images,σxy represents the covariance between the two images.SSIM can better evaluate the image perception quality, so it is widely used in the field of image super-resolution.

V. DATA SET INTRODUCTION

Nowadays, there are many public data sets that can be used for image SR tasks. These data sets contain different image content and quantity, and can comprehensively test the performance of super-resolution reconstruction methods. Table 1 lists the commonly used data sets and briefly introduces them. Some data sets have been divided into fixed training sets, validation sets and test sets, and some larger data sets are often used as training sets, such as ImageNet [22], DIV2K [23] and Flickr2K [15]. Researchers can also divide the training set, validation set, and test set on the data set according to different usage requirements, or increase the training set through image rotation and other methods, or combine multiple data sets for training. In the experiment, the images in the data set need to be correspondingly cropped to adapt to different SR network training.

TABLE I. INTRODUCTION TO COMMONLY USED DATA SETS

Data set name	Number of pictures	Image Format	Brief description of the data set
Set5[24]	5	PNG	The pictures included are from babies, birds, butterflies, children's heads, and a lady.
Set14[25]	14	PNG	The included pictures come from characters, animals, insects, flowers, vegetables, comedians, etc.
Berkeley segmentation[26]	500	JPG	Referred to as BSD500. The pictures included are from animals, buildings, food, people and plants, etc. One of the 100 or 300 pictures is often used, which is called the BSD100 BSD300 data set.
Urban100[27]	100	PNG	The pictures included are mainly different types of urban buildings.
Manga109[28]	109	PNG	The pictures included are all from Japanese manga.
T91[29]	91	PNG	The included pictures come from vehicles, flowers, fruits and human faces. Often used as a training set.
General-100[30]	100	BMP	The pictures included are from animals, daily necessities, food, plants, people, etc. It is also often used as a training set.
DIV2K[23]	1000	PNG	A dataset of high-definition pictures, with pictures from natural environments, landscapes, handicrafts and people, etc. Among them, 800 pictures are often used as training sets.
Flickr2K[15]	2650	PNG	The included pictures come from people, animals, landscapes, etc., and are often used as larger training sets.

VI. SUMMARY AND OUTLOOK

With the development of deep learning, the network of super-resolution reconstruction algorithms is becoming more and more complex, and the reconstruction effect is getting better and better, and it has reached a very high level. The super-resolution method based on deep learning can automatically extract image features, acquire prior knowledge and learn from massive training data, and have a variety of distinctive training models and support from a large number of public data sets. The reconstructed image is in various evaluation indicators. All performed well. The rapid development of deep learning and the continuous improvement of hardware facilities provide very good development opportunities for the field of image super-resolution. Undoubtedly, it has become the most popular research direction in the field of super-resolution research.

Although the performance of existing deep learning image super-resolution reconstruction algorithms has been greatly improved compared to before, far surpassing traditional algorithms, there is still much room for improvement. Looking to the future, research on super-resolution can be carried out from the following aspects:

1) Improve network performance. Improving the image effect after reconstruction has always

been a hot issue for researchers, but for different usage requirements, the performance requirements of the network are also different. For example, in video surveillance images, the reconstructed image needs to have a good visual perception effect and high reconstruction efficiency; in medical image reconstruction, the reconstructed image needs to have better texture details, while ensuring authenticity and credibility. Therefore, improving the reconstruction efficiency, obtaining better visual perception effects, better texture details, higher magnification and other aspects are the focus of future research to continue to improve the performance of super-resolution networks.

2) Application of image super-resolution in various fields. Super-resolution has high application value in video surveillance, medical images, satellite remote sensing imaging, criminal investigation analysis, and face recognition. It optimizes the reconstruction effect in the corresponding scene and has a practical application value for improving image super-resolution. Significant

3) Model evaluation problem. For the problem of image super-resolution reconstruction, the evaluation index directly affects the model optimization measures. At present, PSNR and SSIM are usually used as objective evaluation indicators. Although the calculation is simple and convenient, the consistency with the visual perception effect is poor. Subjective evaluation indicators require high costs, consume a lot of manpower and material resources, and have greater limitations in practical applications. Therefore, according to the characteristics of different reconstruction methods and the needs of different scenes, corresponding evaluation indicators should be designed, which is of great significance for improving the application value of image super-resolution.

REFERENCES

[1] Tsai R. Multiframe image restoration and registration [J]. Advance Computer Visual and Image Processing, 1984, 1: 317-339.

[2] Viet Khanh Ha,Jin-Chang Ren,Xin-Ying Xu,Sophia Zhao,Gang Xie,Valentin Masero,Amir Hussain.Deep Learning Based Single Image Super-resolution:A Survey [J]. International Journal of Automation and Computing,2019, 16(04) :413-426.

[3] Huang Jian, Zhao Yuanyuan, Guo Ping, Wang Jing. A review of single image super-resolution reconstruction methods based on deep learning [J]. Computer Engineering and Applications, 2021, 57(18): 13-23.

[4] Zhang Kaibing, Zhu Danni, Wang Zhen, Yan Yadi. A review of super-resolution image quality evaluation [J]. Computer Engineering and Applications, 2019, 55(04): 31-40+47.

[5] Dong C, Loy C C, He K, et al. Learning a deep convolutional network for image super-resolution[C]//European conference on computer vision. Springer, Cham, 2014: 184-199.

[6] Xu Ran, Zhang Junge, Huang Kaiqi. Image super-resolution algorithm using dual-channel convolutional neural network [J]. Journal of Image and Graphics, 2016, 21(5): 9.

[7] Tang Yanqiu, Pan Hong, Zhu Yaping, Li Xinde. A review of image super-resolution reconstruction research [J]. Chinese Journal of Electronics, 2020, 48(07): 1407-1420.

[8] Liu Yuefeng, Yang Hanxi, Cai Shuang, Zhang Chenrong. Single image super-resolution reconstruction method based on improved convolutional neural network [J]. Computer Applications, 2019, 39(05): 1440-1447.

[9] Shi W , Caballero J , F Huszár, et al. Real-Time Single Image and Video Super-Resolution Using an Efficient Sub-Pixel Convolutional Neural Network[C]// 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2016.

[10] Xie Haiping, Xie Kaili, Yang Haitao. Research progress of image super-resolution methods [J]. Computer Engineering and Applications, 2020, 56(19):34-41.

[11] Kim J, Lee J K , Lee K M . Accurate Image Super-Resolution Using Very Deep Convolutional Networks[C]// IEEE Conference on Computer Vision & Pattern Recognition. IEEE, 2016.

[12] Wang Jiaming, Lu Tao. Satellite image super-resolution algorithm based on multi-scale residual deep neural network [J]. Journal of Wuhan Institute of Technology, 2018, 40(04): 440-445.

[13] Zhang Y, Li K, Li K, et al. Image Super-Resolution Using Very Deep Residual Channel Attention Networks[C]// 2018.

[14] Kim J, Lee J K, Lee K M. Accurate Image Super-Resolution Using Very Deep Convolutional Networks[C]// IEEE Conference on Computer Vision & Pattern Recognition. IEEE, 2016.

[15] Wang X, Yu K, Wu S, et al. ESRGAN: Enhanced Super-Resolution Generative Adversarial Networks [J]. Springer, Cham, 2018.

[16] Johnson J, Alahi A, Fei-Fei L. Perceptual Losses for Real-Time Style Transfer and Super-Resolution[C]// European Conference on Computer Vision. Springer, Cham, 2016.

[17] Mishiba K, Suzuki T, Ikehara M. Edge-adaptive image interpolation using constrained least squares[C]// IEEE International Conference on Image Processing. IEEE, 2010.

[18] Simonyan K, Zisserman A. Very Deep Convolutional Networks for Large-Scale Image Recognition [J]. Computer Science, 2014.

[19] Gatys L A, Ecker A S, Bethge M. Image Style Transfer Using Convolutional Neural Networks[C]// 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2016.

[20] Zhou W, Bovik A C. A universal image quality index [J]. IEEE Signal Processing Letters, 2002, 9(3):81-84.

[21] Zhou W, Bovik A C, Sheikh H R , et al. Image quality assessment: from error visibility to structural similarity [J]. IEEE Trans Image Process, 2004, 13(4).

[22] Russakovsky O, Deng J, Su H , et al. ImageNet Large Scale Visual Recognition Challenge [J]. International Journal of Computer Vision, 2014:1-42.

[23] Agustsson E, Timofte R . NTIRE 2017 Challenge on Single Image Super-Resolution: Dataset and Study[C]// 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). IEEE, 2017.

[24] Bevilacqua M, Roumy A, Guillemot C, et al. Low-Complexity Single Image Super-Resolution Based on Nonnegative Neighbor Embedding [J]. Bmvc, 2012.

[25] Zeyde R. On single im-age scale-up using sparse repre-sentation [J]. Curves&Surfaces, 2010.

[26] 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[C]// IEEE International Conference on Computer Vision. IEEE, 2002.

[27] Huang J B, Singh A, Ahuja N. Single image super-resolution from transformed self-exemplars[C]// IEEE. IEEE, 2015.

[28] Fujimoto A, Ogawa T, Yamamoto K, et al. Manga109 dataset and creation of metadata[C]// the 1st International Workshop. ACM, 2016.

[29] Yang, J.Wright, J.Huang, T.Ma, Y. Image Super-Resolution Via Sparse Representation [J]. IEEE Transactions on Image Processing, 2010, 19(11):2861-2873.

[30] Chao D, Chen C L, Tang X. Accelerating the Super-Resolution Convolutional Neural Network[C]// European Conference on Computer Vision. Springer International Publishing, 2016.

基于深度学习的图像超分辨率重建研究

蒋婧宇
计算机科学与工程学院
西安工业大学
西安，中国
邮箱：jjy1030@126.com;

焦炎
计算机科学与工程学院
西安工业大学
西安，中国
邮箱：jiaoyan@st.xatu.edu.cn;

赵莉
计算机科学与工程学院
西安工业大学
西安，中国
邮箱：332099732@qq.com;

***摘要：*图像超分辨率重建（Super-resolution,SR）旨在使用特定算法将同一场景中的低分辨率模糊图像恢复成高分辨率清晰图像。由于广泛的应用价值与理论价值,图像超分辨率重建技术成为计算机视觉与图像处理领域的一个研究热点,引起了研究者的广泛关注。与传统方法相比，深度学习方法在图像超分辨率重建领域展现出了更好的重建效果，已逐渐发展成为主流技术。因此，本文将图像超分辨率重建问题按照网络模型结构的不同进行系统分类，分为基于卷积神经网络模型的超分方法和基于生成对抗网络模型的超分方法两大类。梳理了主要的图像超分辨率重建方法，阐述了几种较为重要的深度学习超分辨率重建模型，分析比较了不同算法的优缺点及适应的应用场景，讨论了各不同类别超分辨率算法的互相融合和图像视频质量评价的方法，并对常用数据集进行了简单介绍。最后讨论了目前图像超分辨率重建技术所面临的潜在问题，并对未来的发展方向做出了全新的展望。**

关键字：图像超分辨率重建；深度学习；卷积神经网络；生成对抗网络

1. 引言

图像超分辨率重建(Super-resolution Reconstruction，SR)是指采用图像处理和机器学习方法，从同一场景中的一张或多张低分辨率(Low-Resolution，LR)图像重建具有丰富图像细节和清晰纹理的高分辨率(High-Resolution，HR)图像的过程[1]。其在视频、遥感、医学和安全监控等领域都有着重要的应用价值。随着机器学习在计算机视觉领域的迅猛发展，深度学习技术被广泛应用于解决 SR 问题中并取得很好的重建效果，如今已逐渐成为主流。

现有的超分辨率重建算法通常分为三大类：基于插值的方法，这类算法虽简单但提供过于平滑的重建图像，失去了部分细节，产生了振铃效应；基于建模的方法，相较于插值法该类算法重建效果较好，但当面临很大的计算量时，计算过程耗时长，求解困难且受放大因子的影响较大；基于学习的方法，该类算法解决了对尺度缩放因子敏感的问题且重建效果最好，是目前研究的主流方向[3]。

卷积神经网络（convolution neural network，CNN）和生成对抗网络（generative adversarial network，GAN）是目前主流的网络模型，当缩放因子较大时采用这两个网络模型都可以很好

的恢复图像的高频信息，使输出的图像更接近原始真实图像[3]。

2. 超分辨重建方法的网络模型

根据网络模型结构的不同，基于深度学习的图像超分辨率重建方法可以分为以下两大类：①基于卷积神经网络（Convolutional Neural Network，CNN）模型的超分方法；②基于生成对抗网络（Generative Adversarial Networks，GAN）模型的超分方法。针对图像超分辨率的各类需求，产生了具有各种不同特点的超分辨率网络模型（如图 1）[4]。

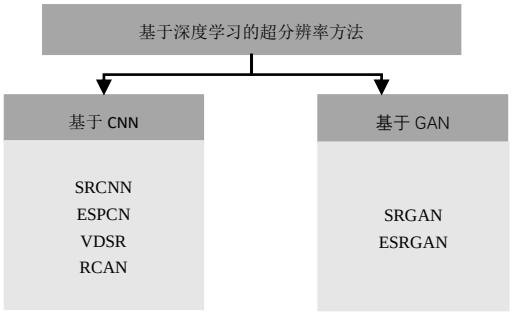

图 1 基于深度学习的超分辨率方法

2.1 基于 CNN 网络模型的超分方法

1) SRCNN

SRCNN[5]将深度学习与传统稀疏编码之间的关系作为依据，将 3 层网络划分为特征提取、非线性映射以及最终的重建。对于一个低分辨率图像，如图 2 所示，该方法先使用双三次（bicubic）插值将其放大至与目标等大小，然后进行图像块提取与表示，再通过三层卷积网络做非线性映射，重建得到的结果作为高分辨率图像输出。SRCNN 首次将卷积神经网络引入到 SR 任务中，与传统 SR 算法的分步处理不同，SRCNN 将各阶段整合到一个深度学习模型中，大幅简化了 SR 工作流程，可以视为基于深度学习的超分辨率方法的里程碑[6]。

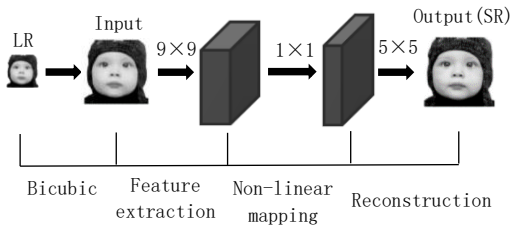

图 2 SRCNN 网络模型

虽然 SRCNN 的网络结构设计简单，在图像重建质量和速度上均优于传统的超分辨率算法，但存在以下问题： ①没有利用任何先验知识；②仅适用于单放大因子的 SR 任务，针对不同的放大因子，网络模型需要再次进行训练；③由于输入图像需要先经过插值放大至与目标等大小，图像重建全过程均在 HR 空间中进行，占用大量的内存空间，增加了计算量，同时，由插值过程产生的误差也会对重构效果产生影响，模型收敛速度较慢，训练耗时较长；④网络层数较少，卷积核感受野也较小(13 ×13)，不能很好地应用图像上下文相关信息，导致最终重建的 HR 图像纹理不清晰，算法适应性也会受限[7]。

最初是采用较小的数据集 Set5 和 Set14 训练 SRCNN 算法，训练后学习到的知识相对较少，图像重建性能受到约束，当采用相对较大的数据集 BSD200 后，重建性能显著提升，由此可见图像的重建性能受数据集大小的影响也较大[8]。

在此之后，虽然专家们又提出了各种网络模型，但是 SRCNN 仍作为一个用于评估其他网络模型性能的基准实验。

2) ESPCN

2016 年，Shi 等人提出了一种基于像素重排的 ESPCN（Efficient Sub-Pixel Convolutional Neural Network）网络模型[9]，ESPCN 的核心概念是亚像素卷积层(sub-pixel convolutional layer)，在 LR 图像上进行卷积操作来获取 LR 图像特征，再通过亚像素卷积层将 LR 空间中的特征扩充到 HR 空间，将卷积后得到的 HR 特征图进行通道重组，得到 HR 图像[10]。

ESPCN 网络主要是对 SRCNN 的重建层进行了改进，将 LR 图像作为网络输入，在重建层采用亚像素卷积层使得网络训练速度成倍提升，简单的网络结构和极高的重建速度使其非常适用于高速且对于重建性能要求相对较低的实时系统。在 ESPCN 网络中，图像尺寸放大使用的插值函数被隐式地包含在前面的卷积层中，可以进行自动学习得到。由于都是在低分辨率图像尺寸大小上进行卷积操作，因此模型效率较高。ESPCN 模型提出的亚像素卷积层在之后被广泛应用，与 FSRCNN 模型中提出的反卷积层相比，学习到的从低分辨率图像到高分辨率图像的上采样的非线性效果更好。值得注意的是，该模型还对激活函数进行了修改，采用 tanh 函数替代了 ReLU 函数，且损失函数为均方误差。

3) VDSR

VDSR(Very Deep CNN for SR)[11]是第一个提出用全局残差学习思想来解决图像 SR 问题的深度模型，是基于 SRCNN 的改进网络，其通过采用多层卷积核进行深层次卷积，既减少了参数量，又使得后面的网络层拥有更大的感受野，能够更好的利用更大区域图像上下文信息，与 SRCNN 相比，获得更好的重建效果。VDSR 的最大特点是网络层数深，图像重建效果好，训练速度比较快。由于作者发现输入的 LR 图像和输出的 HR 图像极其相似，即 LR 图像携带的低频信息与 HR 图像的低频信息极其相似[12]，所以只需要学习 HR 图像和 LR 图像之间的高频残差部分即可。VDSR 网络结构如图 3 所示，它将插值后得到与目标等大小的 LR 图像输入网络，再将这个图像与网络学到的残差相加得到最终的 HR 重建图像。自适应梯度裁剪策略是以更高的学习率来训练网络，尽管架构巨大，但仍可加快收敛速度。因此，在增加网络深度的基础上，结合残差网络和自适应梯度裁剪来加速模型训练，可以提升网络性能且重建效果更好，同时通过对不同尺度大小图像进行混合训练，VDSR 网络可以实现单一模型的多尺度 SR 重建。

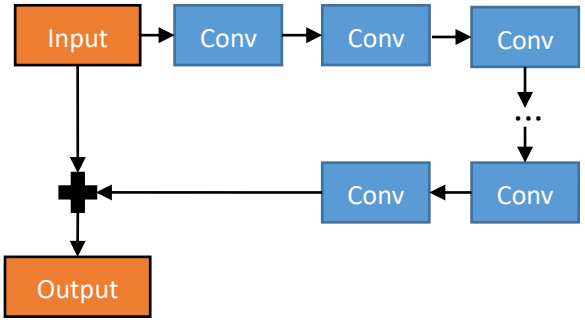

图 3 VDSR 网络结构

4) RCAN

2018 年，Zhang 等人[13]认为输入图像含有大量低频信息，现有的 SR 网络同等对待这些信息的所有通道，缺乏对这些通道的分辨学习能力，阻碍了网络的表征能力，因此提出了深度残差通道注意力网络 RCAN[13]，RCAN 是首个将注意力机制应用于 SR 问题的网络。该算法通过学习不同通道的重要性得到一个权重值，这相当于对信道间特征的相互关系进行建模，自适应调整每个信道特征，从而在有效地强化有用特征通道的同时抑制无用特征通道，更加充分利用计算资源。该模型使用局部嵌套残差结构(residual in residual)，该结构由残差组（RG）和跳远连接（LSC）组成，通过简单堆叠残差块来搭建更深的网络，并通过特征通道之间的依赖关系选择包含更多关键信息的特征通道，增强整个网络的辨识学习能力。

2.2 基于 GAN 网络模型的超分方法

1) SRGAN

生成对抗网络(Generative Adversarial Network，GAN）由 Goodfellow 等提出，它启发自博弈论中的二人零和博弈，GAN 模型中的两位博弈方分别由生成式模型（generative model）和判别式模型（discriminative model）充当[14]。SRGAN 首次将对抗训练应用于图像超分辨率重建问题中，结果显示引入对抗训练能够使网络生成更加精细的纹理细节，GAN 可以完成很多匪夷所思的生成问题，在图像生成、语音转换、文本生成领域均占有很重要地位。

如图 4 所示，SRGAN 将 LR 图像输入至生成器 G 中进行图像重建，由判别器 D 将生成图像与 HR 图像对抗训练，最后输出训练生成的图像。生成器和判别器的协同训练，使网络不仅在像素域判断生成图像与实际高分辨率图像的相似度，且更加关注其在像素空间中的分布相似度。与之前的算法相比，虽然 SRGAN 在客观评价指标（如 PSNR）上相对较低，但是在视觉效果、图像细节等直观方面重建效果更佳。这与其独特的网络结构以及将感知损失(perceptual loss)和对抗损失(adversarial loss)相结合的损失函数有关，其中感知损失是利用卷积神经网络提取出的特征，通过比较生成图片和目标图片经过卷积神经网络后的特征差别，使生成图片和目标图片在语义和风格上更相似。对抗损失由 GAN 提供，根据图像是否可以成功欺骗判别网络进行训练。SRGAN 是追求视觉体验发展中的一个里程碑，SRGAN 显著提高了基于 PSNR 的方法重建的整体视觉质量。

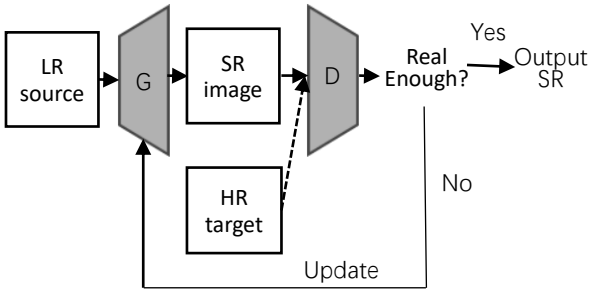

图 4 SRGAN 基本结构示意图

2) ESRGAN

为了提升 SRGAN 模型的重建精度，在此基础上，ESRGAN[15]模型在三个方面上进行了改进：对生成器架构进行了更改，通过去除批量归一化（BN）层以降低重建图像中产生的伪影，并引入一种有更高重建精度的密集残差块（RDDB）来提升模型的结构，使之具有更大容量且更易于训练，有助于提高泛化能力，降低计算复杂度和内存使用；改进感知域损失函数，使用激活前的 VGG 特征，这个改进会提供清晰的边缘和更符合视觉的结果，可以更好地保持图像亮度一致性和恢复更好的细节纹理；增强了判别器的判别能力，使用相对判别器，使其判别不再是真伪绝对值而是预测生成图像与真实图像的相对距离。

与 SRGAN 模型相比，在大部分的标准测试图像上，ESRGAN 模型能在保持较好细节的同时使重建图像中的噪声变少。大量实验表明，增强型 SRGAN，称之为 ESRGAN，在锐度和细节上始终优于最先进的方法。

2.3 小结

近年来，深度学习技术蓬勃发展，基于深度学习的超分辨率重建方法已逐步成为图像超分辨率的主流，并发展出了一系列以 CNN 和 GAN 为基础的网络模型，2.1-2.2 节分别介绍了基于 CNN 网络模型和基于 GAN 网络模型的超分方法。整体来看，算法的重建性能在不断提升，重建后的图像与原始图像越来越接近。基于 GAN 网络模型的算法重建得到的图像纹理细节较优，提升了图像的真实感，人眼感知效果较好。与之相比，基于 CNN 网络模型的算法网络复杂度更低，训练难度更小，重建结果精度更高，但会产生伪影且重建速度更慢。

3. 损失函数构建

在图像超分辨领域，损失函数是用来度量生成的图像与实际高分辨率图像之间的差异。目前，在超分辨领域中，损失函数发挥着重要的作用，常用的损失函数有基于像素的损失函数和基于感知的损失函数。

3.1 像素损失函数

像素损失主要是度量预测图像和目标图像之间的像素差异，主要包括了 L_1（平均绝对误差）损失和 L_2（均方误差）损失。

L_1 损失，也被称为最小绝对偏差（LAD）和最小绝对误差（LAE），计算的是实际值与目标值之间绝对差值的总和。在有监督的图像超分辨率任务中，目标是使得生成的图像（SR）尽可能接近真实的高分辨率图像（HR），使用 L_1 损失计算的则是 SR 和 HR 对应像素位置的值

之间的误差。平均绝对值误差（MAE）的基本表达式如式（1）所示：

$$L_{MAE}(\theta) = \frac{1}{n}\sum_{i=1}^{n}\left|Y_i - F(X_i,\theta)\right| \quad (1)$$

其中，L(θ)表示网络需要优化的损失函数，n 表示训练样本的数目，θ 表示深度神经网络的参数，$F(X_i, \theta)$为网络重建后的图像，Y_i 是表示对应的 HR 图像。L_2损失，也叫均方误差损失（MSE），计算的是实际值与目标值之间绝对差值的平方总和，使得基于深度学习的图像超分辨率模型在性能方面有了很大的提升，是最常用的损失函数。但是 MSE 损失能够对大的损失进行惩罚，对于小的损失上却无所作为，因此会产生模糊的图像。其基本表达式如式（2）所示：

$$L_{MSE}(\theta) = \frac{1}{n}\sum_{i=1}^{n}\left\|Y_i - F(X_i,\theta)\right\|^2 \quad (2)$$

其基本符号意义与式（1）中一致。最小均方误差的应用有效解决了 SR 重建图像与目标 HR 图像之间差值衡量问题，使得基于深度学习的图像 SR 模型相对传统基于学习的 SR 重建模型有了较大的提高。

与 MSE 相比，MAE 有个优点，即它对离群点有更好的鲁棒性，更具有包容性。由式（1）可知，MAE 计算的是误差 Yi-F(Xi, θ)的绝对值，所以无论是 Yi-F(Xi, θ)>1 还是 Yi- F(Xi, θ)<1，没有平方项的作用，惩罚力度都是相同的，所占权重相同。

由于这些损失函数分别对每个像素向量的类预测进行评估，然后对所有像素进行平均，因此它们断言图像中的每个像素都具有相同的学习能力。

与 L_1 损失函数相比，L_2 损失函数会放大最大误差和最小误差之间的差距（比如 2*2 和 0.1*0.1），另外 L_2 损失函数对异常点也比较敏感。

由于峰值信噪比（PSNR）的定义（见 4.1）与像素差相关度较高，且最小化像素损失等同于将 PSNR 最大化，所以像素损失函数逐渐成为被普遍使用的损失函数。但由于其并不探究图像质量问题(例如，感知质量[16]，纹理细节[17])，因此结果往往欠缺高频细节内容，产生过于平滑的纹理细节。

3.2 感知损失函数

对于图像风格化，图像超分辨率重建等任务来说，早期都使用了图像像素空间的 L_2 损失，但是 L_2 损失与人眼感知损失的图像质量并不契合，恢复出来的图像细节往往表现较差。如今的研究中，L_2 损失逐步被人眼感知损失所替代。人眼感知损失也被称为感知损失（perceptual loss），其与 MSE 损失采用图像像素进行求差的不同之处在于所计算的空间不再是图像空间。基于感知的损失函数可以恢复更多的高频信息，使重建性能更好。目前，基于感知的损失函数主要有内容损失和对抗损失。

内容损失函数又分为特征重建损失函数和风格重建损失函数。Bruna 等人[18]首次提出特征重建损失函数，通过预训练的 VGG19 网络分别提取重建 SR 图像与 HR 图像在特征空间中对应的特征映射且进行比较，其表达式如下：

$$L_{VGG/j} = \frac{1}{W_j H_j}\sum_{h=1}^{H_j}\sum_{w=1}^{W_j}\left(\phi_j(Y)_{x,y} - \phi_j(G_{\theta_G}(X))_{x,y}\right)^2 \quad (3)$$

其中，W_j 和 H_j 表示第 j 幅特征图的宽与高，ϕ_j 表示 VGG19 网络内第 j 次卷积获得的特征映射，X 表示原始 LR 图像，Y 为重建后的 HR 图像，G_{θ_G} 表示网络生成的 SR 图像。Gatys 等人[19]为了使重建 SR 图像与 HR 图像的纹理细节等保持一致，在特征重建损失函数的基础上又提出

了风格重建损失函数，该函数定义了一个 Gram 矩阵，计算公式如下：

$$G_j^{\phi}(x)_{c,c'} = \frac{1}{C_j W_j H_j} \sum_{h=1}^{H_j} \sum_{w=1}^{W_j} \phi_j(x)_{h,w,c} \phi_j(x)_{h,w,c'} \quad (4)$$

经过 VGG 网络提取到的特征图大小为 $C_j \times W_j \times H_j$。随后在对应层中计算 Gram 矩阵的欧式距离差并相加得到风格重建损失函数，如下式：

$$L_{style/j}^{\phi} = \left\| G_j^{\phi}(Y) - G_j^{\phi}(G_{\theta_G}(x)) \right\|^2 \quad (5)$$

SRGAN 网络首次提出对抗损失，其基本形式如下：

$$L_{Gen}^{SR} = \sum_{n=1}^{N} -lbD_{\theta_D}(G_{\theta_G}(x)) \quad (6)$$

其中，N 表示图像数量，θ_D 表示鉴别网络的参数，θ_G 表示生成网络的参数，表示生成图像，是真实的 HR 图像的概率。网络最终的优化目标为一个最小最大化问题：

$$\min_{\theta_G} \max_{\theta_D} E_{I^{HR} \square P_{train}(I^{HR})} \left[lbD_{\theta_D}(I^{HR}) \right] + E_{I^{HR} \square P_G(I^{HR})} \left[1 - lbD_{\theta_D}(G_{\theta_G}(I^{LR})) \right] \quad (7)$$

其中，P_{train}（I^{HR}）表示 HR 图像分布，P_G（I^{LR}）表示原始 LR 图像分布，对抗训练使得生成的 SR 图像与原始的 HR 图像高度相似，使得判别网络难以辨别，最终得到能够以假乱真的 SR 图像。SRGAN 的判别目标为输入图像是否为真，与 SRGAN 不同，ESRGAN 用相对平均判别器替代了判别网络的判别器，判别目标为预测真实 HR 图像比生成 SR 图像更真实的概率。判别网络如式（8）和式（9）所示：

$$\begin{aligned} D(x_r) &= \sigma(C(real)) \to 1 \\ D(x_f) &= \sigma(C(fake)) \to 0 \end{aligned} \quad (8)$$

$$\begin{aligned} D_{Ra}(x_r, x_f) &= \sigma(C(real) - E\left[C(fake)\right]) \to 1 \\ D_{Ra}(x_f, x_r) &= \sigma(C(fake) - E\left[C(real)\right]) \to 0 \end{aligned} \quad (9)$$

其中，real 表示真实 HR 图像，fake 表示生成 SR 图像，C(real)表示鉴别器判断结果，E[C(fake)]表示鉴别网络判断结果的平均值。其中 σ 表示 Sigmoid 激活函数，通过改进判别器帮助网络学习更敏锐的边缘和更多的纹理细节。

4. 图像超分辨质量评价

通过超分辨技术重构后得到的高分辨率图像，与真实高分辨率图像的相似度越高，则表示图像超分辨算法性能越好。一般来说，主要通过两种客观的量化指标进行评价，包括峰值信噪比(Peak signal-to-noise ratio, PSNR)[20]和结构相似度(Structual similarity, SSIM)[21]。为了比较的公平性，在中心下垂的 y 通道上对所有 PSNR（dB）和 SSIM 度量进行计算。评价指标值越高，说明重建结果和原始图像差异性越小，逼真度越高。

PSNR 是一种最常用的图像重构质量评价方式，基于图像的均方误差（MSE），MSE 如下式(10)所示：

$$MSE = \frac{1}{mn} \sum_{i=0}^{m-1} \sum_{j=0}^{n-1} \left[I(i,j) - K(i,j) \right]^2 \quad (10)$$

其中，I 表示原始高分辨率图像，K 表示重建后的图像，m 和 n 分别表示图像的长和宽。PSNR 定义如下：

$$PSNR = 10 \log\left(\frac{MAX_I^2}{MSE}\right) \quad (11)$$

其中，对数底为自然底数 e，若不额外说明均采用此底数。MAX_I 为当前图片中每一个像素能表示的颜色最大个数，即图片的 bit 数。由式

(11)可以看出，最小化 MSE 相当于最大化 PSNR。PSNR 计算的是同位置像素之间的误差，在计算过程中未考虑可视化感知特性对图片质量的影响，所以偶尔会产生虽然 PSNR 值很高但人主观感受到的图像质量并没有提高的现象。由于无法对图像感知质量进行定量分析，PSNR 仍作为图像超分辨领域最常用的评价方式。

结构相似度（Structural SIMilarity，SSIM）可以同时比较图像亮度、对比度、结构这三方面的相似度。SSIM 的取值范围为［0，1］，值越高就越表明重建结果和原始图像结构越相似。

$$\begin{aligned} l(x,y) &= \frac{2\mu_x\mu_y + c_1}{\mu_x^2 + \mu_y^2 + c_1} \\ c(x,y) &= \frac{2\sigma_x\sigma_y + c_2}{\sigma_x^2 + \sigma_y^2 + c_2'} \\ s(x,y) &= \frac{\sigma_{xy} + c_3}{\sigma_x\sigma_y + c_3} \end{aligned} \tag{12}$$

$$SSIM(x,y) = (x,y)^{\alpha} \cdot c(x,y)^{\beta} \cdot s(x,y)^{\gamma} \tag{13}$$

其中，μ_x，μ_y分别为两张图像的均值，σ_X，σ_Y分别代表两张图像的方差，σ_{xy}表示两张图像之间的协方差。SSIM 能够较好地评估图像感知质量，所以在图像超分辨领域被广泛使用。

5. 数据集介绍

如今有很多公开数据集可用于图像 SR 任务，这些数据集包含的图像内容、数量等各不相同，可对超分辨重建方法的性能进行综合测试。表 1 列出了常用的数据集，并进行了简单介绍。一些数据集已被划分为固定的训练集、验证集和测试集，也有一些较大的数据集常被用做训练集，如 ImageNet[22]、DIV2K[23]和 Flickr2K[15]等。研究人员也可根据不同的使用需求在数据集上自行划分训练集、验证集和测试集，或者通过图像旋转等方式进行训练集扩增，或者联合多个数据集进行训练。在实验中，需要将数据集中的图像进行对应裁剪，以适应不同 SR 网络训练。

表 1 常用数据集介绍

数据集名称	图片数量	图片格式	数据集简单描述
Set5[24]	5	PNG	包含的图片分别来自婴儿、鸟、蝴蝶、小孩的头部和一个女士。
Set14[25]	14	PNG	包含的图片来自人物、动物、昆虫、花、蔬菜和喜剧演员等。
Berkeley segmentation[26]	500	JPG	简称 BSD500。包含的图片来自动物、建筑、食物、人和植物等。经常使用其中的 100 张或 300 张图片，称为 BSD100 BSD300 数据集。
Urban100[27]	100	PNG	包含的图片主要是不同类型的城市建筑物。
Manga109[28]	109	PNG	包含的图片均来自日本漫画。
T91[29]	91	PNG	包含的图片来自车辆、花、水果和人脸等。常被用作训练集。
General-100[30]	100	BMP	包含的图片来自动物、日常用品、食物、植物和人物等。也常用作训练集。
DIV2K[23]	1000	PNG	高清图片数据集，图片来自自然环境、风景、手工艺品和人物等。其中的 800 张图片经常作为训练集使用。
Flickr2K[15]	2650	PNG	包含的图片来自人物、动物和风景等，常作为较大规模训练集使用。

6. 总结与展望

随着深度学习的发展，超分辨率重建算法的网络越来越复杂，重建效果也越来越好，已经达到了很高的水平。基于深度学习的超分辨率方法能够自动提取图像特征，从海量训练数据中获取先验知识并进行学习，拥有各类各具特色的训练模型和大量公开数据集的支持，重建图像在各评价指标上都表现良好。深度学习的快速发展，以及硬件设施的不断完善，为图像超分辨领域提供了非常好的发展机遇。毋庸置疑，其已成为超分辨率研究领域的最热门研究方向。

虽然现有的深度学习图像超分辨率重建算法的性能较之前已经有了很大提升，远超越传统算法，但还有很大提升空间。展望未来，超分辨率的研究可以从以下几个方面开展：

（1）提升网络性能。提升重构后的图像效果一直是研究者们关注的热点问题，但对于不同的使用需求，对网络的性能要求也不同。例如，视频监控图像中，需要重建图像视觉感知效果好，重建效率高；医学图像重建中，需要重建图像具有较优的纹理细节，同时保证真实可信。因此，提升重建效率、获得更好的视觉感知效果、更优的纹理细节、更高的放大倍数等方面，是未来继续提升超分辨率网络性能的研究重点。

（2）图像超分辨率在各领域的应用。超分辨率在视频监控、医学图像、卫星遥感成像、刑侦分析和人脸识别等方面有很高的应用价值，实现对应场景下重构效果最优化，对于提升图像超分辨率的实际应用价值具有重大意义。

（3）模型评价问题。对于图像超分辨率重建问题，评价指标直接影响着模型优化举措。目前通常采用 PSNR 和 SSIM 作为客观评价指标，虽计算简单方便，但与视觉感知效果一致性较差。主观评价指标所需成本较高，花费大量人力、物力，在实际应用中局限性较大。因此，应针对不同重建方法的特点和不同场景需求，设计相应的评价指标，对于提升图像超分辨率的应用价值具有重要意义。

参考文献

[1] Tsai R. Multiframe image restoration and registration [J]. Advance Computer Visual and Image Processing, 1984, 1: 317-339.

[2] Viet Khanh Ha, Jin-Chang Ren, Xin-Ying Xu, Sophia Zhao, Gang Xie, Valentin Masero, Amir Hussain. Deep Learning Based Single Image Super-resolution:A Survey [J]. International Journal of Automation and Computing, 2019, 16(04):413-426.

[3] 黄健,赵元元,郭苹,王静.深度学习的单幅图像超分辨率重建方法综述[J].计算机工程与应用,2021,57(18):13-23.

[4] 张凯兵,朱丹妮,王珍,闫亚娣.超分辨图像质量评价综述[J].计算机工程与应用,2019,55(04):31-40+47.

[5] Dong C, Loy C C, He K, et al. Learning a deep convolutional network for image super-resolution[C]//European conference on computer vision. Springer, Cham, 2014: 184-199.

[6] 徐冉,张俊格,黄凯奇.利用双通道卷积神经网络的图像超分辨率算法[J].中国图象图形学报，2016, 21(5):9.

[7] 唐艳秋,潘泓,朱亚平,李新德.图像超分辨率重建研究综述 [J]. 电子学报, 2020, 48(07):1407-1420.

[8] 刘月峰,杨涵晰,蔡爽,张晨荣.基于改进卷积神经网络的单幅图像超分辨率重建方法[J].计算机应用,2019,39(05):1440-1447.

[9] Shi W, Caballero J, F Huszár, et al. Real-Time Single Image and Video Super-Resolution Using an Efficient Sub-Pixel Convolutional Neural Network[C]// 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2016.

[10] 谢海平,谢凯利,杨海涛.图像超分辨率方法研究进展[J].计算机工程与应用,2020,56(19):34-41.

[11] Kim J, Lee J K, Lee K M. Accurate Image Super-Resolution Using Very Deep Convolutional Networks[C]// IEEE Conference on Computer Vision & Pattern Recognition. IEEE, 2016.

[12] 汪家明,卢涛.多尺度残差深度神经网络的卫星图像超分辨率算法[J].武汉工程大学学报，2018, 40(04):440-445.

[13] Zhang Y, Li K, Li K, et al. Image Super-Resolution

Using Very Deep Residual Channel Attention Networks[C]// 2018.

[14] Kim J, Lee J K, Lee K M. Accurate Image Super-Resolution Using Very Deep Convolutional Networks[C]// IEEE Conference on Computer Vision & Pattern Recognition. IEEE, 2016.

[15] Wang X, Yu K, Wu S, et al. ESRGAN: Enhanced Super-Resolution Generative Adversarial Networks [J]. Springer, Cham, 2018.

[16] Johnson J, Alahi A, Fei-Fei L. Perceptual Losses for Real-Time Style Transfer and Super-Resolution[C]// European Conference on Computer Vision. Springer, Cham, 2016.

[17] Mishiba K, Suzuki T, Ikehara M. Edge-adaptive image interpolation using constrained least squares[C]// IEEE International Conference on Image Processing. IEEE, 2010.

[18] Simonyan K, Zisserman A. Very Deep Convolutional Networks for Large-Scale Image Recognition [J]. Computer Science, 2014.

[19] Gatys L A, Ecker A S, Bethge M. Image Style Transfer Using Convolutional Neural Networks[C]// 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2016.

[20] Zhou W, Bovik A C. A universal image quality index [J]. IEEE Signal Processing Letters, 2002, 9(3):81-84.

[21] Zhou W, Bovik A C, Sheikh H R , et al. Image quality assessment: from error visibility to structural similarity [J]. IEEE Trans Image Process, 2004, 13(4).

[22] Russakovsky O, Deng J, Su H, et al. ImageNet Large Scale Visual Recognition Challenge [J]. International Journal of Computer Vision, 2014:1-42.

[23] Agustsson E, Timofte R. NTIRE 2017 Challenge on Single Image Super-Resolution: Dataset and Study[C]// 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). IEEE, 2017.

[24] Bevilacqua M, Roumy A, Guillemot C , et al. Low-Complexity Single Image Super-Resolution Based on Nonnegative Neighbor Embedding [J]. bmvc, 2012.

[25] Zeyde R. On single im-age scale-up using sparse repre-sentation [J]. Curves&Surfaces, 2010.

[26] 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[C]// IEEE International Conference on Computer Vision. IEEE, 2002.

[27] Huang J B, Singh A, Ahuja N. Single image super-resolution from transformed self-exemplars[C]// IEEE. IEEE, 2015.

[28] Fujimoto A, Ogawa T, Yamamoto K, et al. Manga109 dataset and creation of metadata[C]// the 1st International Workshop. ACM, 2016.

[29] Yang, J. Wright, J. Huang, T. Ma, Y. Image Super-Resolution Via Sparse Representation [J]. IEEE Transactions on Image Processing, 2010, 19(11):2861-2873.

[30] Chao D, Chen C L, Tang X. Accelerating the Super-Resolution Convolutional Neural Network[C]// European Conference on Computer Vision. Springer International Publishing, 2016.