J. Shanghai Jiao Tong Univ. (Sci.), 2023
https://doi.org/10.1007/s12204-023-2654-3

Infrared Single Pixel Imaging Based on Generative Adversarial Network

JIANG Yilin[1] (蒋伊琳), *ZHANG Yilong*[1] (张怡龙), *ZHANG Fangyuan*[2*] (张芳园)
(1. College of Information and Communication Engineering, Harbin Engineering University, Harbin 150001, China;
2. College of Computer Science and Technology, Harbin University of Science and Technology, Harbin 150080, China)

Abstract: In the field of imaging, the image resolution is required to be higher. There is always a contradiction between the sensitivity and resolution of the seeker in the infrared guidance system. This work uses the rosette scanning mode for physical compression imaging in order to improve the resolution of the image as much as possible under the high-sensitivity infrared rosette point scanning mode and complete the missing information that is not scanned. It is effective to use optical lens instead of traditional optical reflection system, which can reduce the loss in optical path transmission. At the same time, deep learning neural network is used for control. An infrared single pixel imaging system that integrates sparse algorithm and recovery algorithm through the improved generative adversarial networks is trained. The experiment on the infrared aerial target dataset shows that when the input is sparse image after rose sampling, the system finally can realize the single pixel recovery imaging of the infrared image, which improves the resolution of the image while ensuring high sensitivity.

Key words: image resolution, rose sampling, generative adversarial networks, single pixel imaging
CLC number: TP 391 **Document code:** A

0 Introduction

Infrared imaging is to use the detector to image the target containing the background which uses the thermal radiation difference between the target and the background and the characteristics of the target to form the infrared image of the target and the surrounding background. Single pixel imaging is an imaging technology developed from ghost imaging. It is based on the principle of correlation measurement, and only relies on collecting the intensity information of light to image objects. The single pixel camera adopts structured light illumination at the end of lighting and a single pixel light intensity detector at the detection end to collect signals. When the lighting structure changes, the change of the corresponding light intensity reflects the degree of correlation between the lighting structure and the spatial information of the object. By constantly changing the lighting structure and accumulating correlation information, the final image of the object is achieved. Because the single pixel camera only needs light intensity detection at the detection end, its requirements for detectors are far lower than those of the area array detectors in ordinary imaging. Therefore, in some special conditions such as some bands with immature array detector technology, single pixel imaging technology has great application advantages. When the dimensions of the output and input of the detector are the same, the sensitivity can be understood as the magnification. Higher measurement accuracy can be obtained by improving the sensitivity. For the single-pixel imaging system with point scanning, the imaging range is fixed and limited. If the sensitivity is low, the field of vision is wider and the imaging range is larger. When the target is smaller, it is not convenient for subsequent information processing. With high sensitivity, the field of vision is smaller, the imaging range is smaller, and the target imaging is clearer. The sparse small target generated by the system with low sensitivity is supplemented and amplified so that it has a wide field of vision and has a clear target image with high sensitivity similarly. In the form of software, a system retains its advantages, but also improves its sensitivity, thus presenting a larger and clearer image.

Reducing the pixel size and increasing the detector array size are the most direct ways to improve the resolution of the imaging system, but these methods are limited with space and high cost, which are difficult to meet the actual needs[1]. At this time, super resolution reconstruction technology that can effectively

Received: 2022-10-11 **Accepted:** 2023-02-21
Foundation item: the Fundamental Research Funds for the Central Universities (No. 3072022CF0802)

*E-mail:** 18645010385@163.com

improve the image resolution has naturally become the focus of attention because of its low cost and large development space. Super resolution technology can effectively improve the spatial resolution of the image obtained by the imaging system without changing the composition and structure of the imaging system on a large scale[2], which means using super resolution reconstruction technology can obtain a higher resolution target image on the basis of the existing imaging system, so as to apply the existing imaging technology to a wider range of fields. Thus, it is expected to use the method of computational imaging to break this deadlock and solve the contradiction between high resolution and high sensitivity.

In the early stage of development, single pixel reconstruction technology is mainly based on the method of frequency domain[3]. The successful demonstration of the computational ghost imaging which is similar to single pixel imaging[4] proves that ghost imaging can be accurately described by the cross-correlation theory of light field, rather than relying on the imaging technology of quantum properties. With the gradual maturity of low light level motor system technology, researchers began to use the random mask pattern generated by the spatial light modulator controlled and encoded by the computer to replace the rotating diffuser, in which the mask pattern loaded on the spatial light modulator is known. The area array detector is no longer required to measure, so the ghost imaging light route is combined into one forming a classic light route for computing ghost imaging[5]. Since the spatial light modulator can realize the controllable modulation of the light field distribution, it can project the mask pattern with a high sampling efficiency and reduce the sampling time while ensuring the quality of the reconstructed image. Single pixel imaging technology saves the resolution of the detection end and the cost in the measurement end. It also reduces the volume, especially in the non-visible light field. Moreover, the robustness of traditional imaging is poor, and the ability to resist atmospheric turbulence and scattering is poor. Once there is a scattering medium in the light propagation path, the traditional camera cannot accurately image the target. Most natural scenes can be expressed sparsely in the orthogonal transform domain, so orthogonal mask patterns are often used for single pixel imaging, such as Hadamard base pattern[6], Fourier base pattern[7] and wavelet base pattern[8], which can reconstruct a clearer image in the case of under sampling. When the number of samples is equal to the number of image pixels, complete sampling can be achieved. Tsai and Huang[9] used the characteristics of Fourier transform in the process of super-resolution reconstruction. After deriving the mapping relationship between low-resolution image and high-resolution image, it solves the frequency domain coefficient of high-resolution image through the derived relationship and finally uses the inverse Fourier transform to obtain a high-resolution image close to the original image. Kim et al.[10] improved the super-resolution reconstruction method. The main improvement focus is to solve the nonconvergence problem of the reconstruction process caused by noise and blur by introducing regularization. Shallow learning super-resolution reconstruction method is represented by sparse representation algorithm[11-12]. In the process of single pixel imaging, it is necessary to use spatial light modulator to sample the mask of the scene. If the mask pattern used for each sampling is the row vector of the sampling matrix, compressed sensing can be realized, so compressed sensing can be naturally integrated into single pixel imaging technology. There are two core problems in single pixel imaging based on compressed sensing: the design of sampling matrix and speed and accuracy of reconstruction algorithm. The compressed sensing reconstruction method based on iterative optimization mainly depends on the sparsity in the transform domain. The most basic iterative optimization algorithm is the matching pursuit (MP) algorithm[13], which selects the sampling basis that contributes the most to the measured value from the sampling matrix in each iteration as the sparse approximation of the original signal and uses the residual as the evaluation index for multiple iterations. In order to improve the reconstruction accuracy and reduce the computational complexity, other improved algorithms have added some iterative modifications on the basis of this algorithm such as orthogonal matching pursuit (OMP) algorithm[14], and regularized orthogonal matching pursuit (ROMP) algorithm. Yang et al.[15] proposed a sparse representation super-resolution reconstruction algorithm. When the over complete dictionaries of low-resolution and high resolution are constructed, the sparse representation coefficient of the low-resolution image will be calculated. The sparse representation coefficient and high-resolution image dictionaries are used for convolution. Then, the mapping from low-resolution images to high-resolution images is formed, and the resolution of the image has become higher.

With the rapid development of people's computing power and the great potential of deep learning methods in the field of computer vision, the technology of super-resolution reconstruction based on deep learning has attracted more and more attention of researchers[16]. Due to the proposal of deep convolution neural network, the method of deep convolution neural network has developed vigorously[17-18]. It is significant to apply deep convolution neural network to image restoration[19], which can improve the peak signal-to-noise ratio (PSNR). On the one hand, because of the separation of training process and reconstruction process, the time-consuming training process can be carried out on the mainframe or server, and the trained model can be transplanted to the mobile terminal for fast operation after compression and quantification[20]. On the other hand, the advent of dedicated neural network chips in recent years has brought a huge computing power to mobile terminals[21]. The advantage of super-resolution reconstruction algorithm based on deep learning lies in the flexible and accurate introduction of external sample information, so as to increase the high frequency detail information of the image and complete the super-resolution reconstruction of low-resolution images. Because researchers can use different neural networks[22], at the same time, they can use different neural network layers and different numbers of convolution cores, which makes the types of super-resolution reconstruction algorithms based on deep learning more complicated. Taking the number of input images as the classification standard, the deep learning super-resolution reconstruction algorithm can be divided into single frame reconstruction method[23] and multi-frame reconstruction method[24]. Caballero et al.[25] proposed a generative adversarial network (GAN), which uses convolutional neural network in order to achieve the super-resolution of a single image. The bottleneck still lies in how recovering the fine texture information of the image. The generated image quality is often poor when a group of random noise is input into the generator in a GAN. Caballero et al.[25] proposed a super-resolution generative adversarial network (SRGAN) and defined a loss function to drive the model, which can finally generate high-resolution images. Sajjadi et al.[26] proposed the perceptual loss to optimize the super-resolution model in feature space rather than pixel space. Wang et al.[27] combined the prior information of semantic images to restore the texture details of images better. However, most of the current methods tend to excessively increase the PSNR value and ignore the high-frequency details of the output image. They do not take into account the research on the restoration of sparse images and the relationship between sparse algorithm and restoration algorithm, resulting in the result that compression algorithm and restoration algorithm are relatively independent and have little relevance. There are mainly two ways of single pixel imaging. Many researchers are using compressed sensing or deep learning for image processing and imaging, but the relevant research on combining them for image processing is still in the blank period. Based on the existing research status, technology and theory, this paper combines the advantages of single pixel high sensitivity with the mentioned neural network to achieve single pixel imaging. It solves the shortcomings of single pixel imaging and can form a fast target.

In this paper, a sparse super-resolution generative adversarial network (SSRGAN) is proposed. Three improvements are made on the basis of the SRGAN model. First, we use the residual dense block (RDDB) to improve the network structure and make the training simpler. The experimental simulation shows that removing the batch standardization (BN) layer will reduce the residual effect of the image generated by the generator and improve the quality of image restoration. Second, the relative average GAN (RAGAN) is used to improve the discriminator that is used to judge whether the generated image is more realistic than the label rather than whether the generated image is a real image. This improvement helps the generator recover more realistic texture details and higher resolution with a small amount of data. Third, the visual geometry group network (VGG) features are used to improve the perceptual loss before activation function instead after activation function. The adjusted perceptual loss provides clearer edges and better visual effects. An infrared imaging system with high sensitivity and high resolution is formed. The rosette scanning imaging method is used to simulate the work of the point scanning detector. Combined with physical reality, the network model generated by sparse sampling of images is more universal. The final infrared optical sparse compression module completes the thinning of the infrared image. The computational reconstruction algorithm based on the deep neural network realizes the restoration and generates the infrared image, improving the sensitivity and image quality of the single pixel infrared imaging system from the system level.

1 Related Work

1.1 System Process

The research hardware block diagram and process are as follows. The single pixel imaging system adopts the rose line scanning method to scan and image the infrared radiation. The basic composition of the scanning single pixel imaging system is shown in Fig. 1. First step is to train the generative adversarial network and the high-resolution image set is down sampled to generate a low-resolution image training set. Then the sparse image set is restored to form a model that can generate relatively accurate high-resolution images through the information of low-resolution images. According to the actual super-resolution image results, we train the countermeasure network again so that the actual effect of the generative countermeasure network is significantly improved. Infrared light machine scanning is used to image the target with rose sparse scanning. The reconstruction and restoration of low-resolution image are completed through the trained generation network.

The working principle is as follows. The infrared detection system collects the long-range infrared radiation transmitted from the atmosphere and passes it through the optical channel to the scanning mechanism. The scanning mechanism converges the infrared radiation

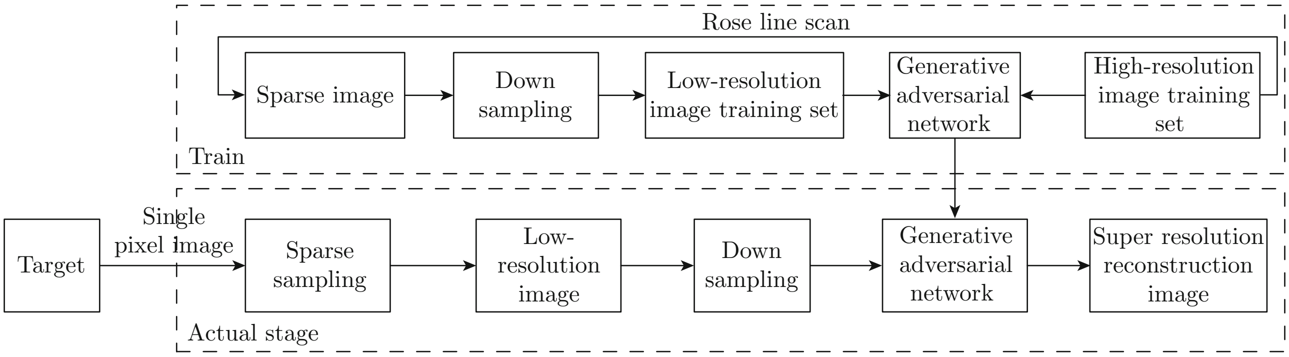

Rose line scan
Sparse image
Down sampling
Low-resolution image training set
Generative adversarial network
High-resolution image training set
Train
Target
Single pixel image
Sparse sampling
Low-resolution image
Down sampling
Generative adversarial network
Super resolution reconstruction image
Actual stage

Fig. 1 Block diagram of infrared super-resolution system

of the instantaneous field of view to the infrared detector located at the focus of the optical system according to a certain scanning mode. The detector converts the infrared radiation into a one-dimensional electrical signal modulated by azimuth and sends it to the signal processing module through pre-amplification.

The one-dimensional scanning electric signal is subject to azimuth conversion phase-locked imaging, which is sent to the digital image recognition and tracking unit for recognition and tracking processing, and the image processing results are output to form tracking instructions and sent to the infrared detection system.

The ideal Cassegrain system is composed of two perfectly consistent optical convex lenses with a certain distance apart and slightly deflected relative to the optical axis of the system. The primary and secondary mirrors rotate around the optical axis in the opposite direction at different frequencies so that the detector can receive the infrared radiation of the instantaneous field of view according to the rose line scanning pattern track. The trajectory equation of continuous rose line scanning is

$$\left.\begin{aligned} x(t) &= \frac{d_{\mathrm{rp}}}{2}(\cos(\omega_1 t+\theta_1)+\cos(\omega_2 t+\theta_2)) \\ y(t) &= \frac{d_{\mathrm{rp}}}{2}(\sin(\omega_1 t+\theta_1)-\sin(\omega_2 t+\theta_2)) \end{aligned}\right\}, \tag{1}$$

where, d_{rp} is the maximum length of rose petals, equal to the scanning field radius; $\omega_1 = 2\pi f_1$ and $\omega_2 = 2\pi f_2$ are the scanning angular frequencies; f_1 and f_2 are the rotational frequencies of two optical elements; θ_1 and θ_2 are the initial phases of the optical element rotation driving motor. The principle of Cassegrain optical system is shown in Fig. 2.

Figure 2 is the schematic diagram of optical mechanical scanning. In the simulation experiment of this paper, Cassegrain optical system is used to simulate the rose line scanning mode of infrared seeker with $d = 64$, $f_1 = 290\,\mathrm{rad/s}$, and $f_2 = 70\,\mathrm{rad/s}$.

Figure 3(a) shows the schematic diagram of rosette scanning for one frame, using the simulated rosette track combined with the label image. Because the point scan takes a long time to obtain all the information of the target, the multi-frame combination of different

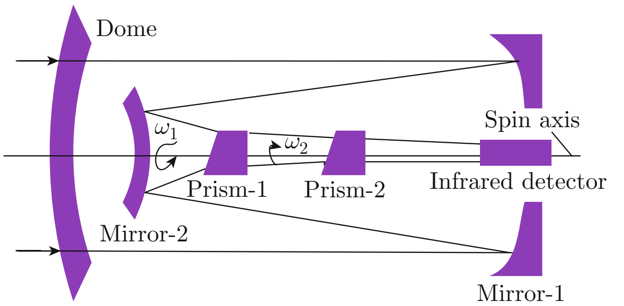
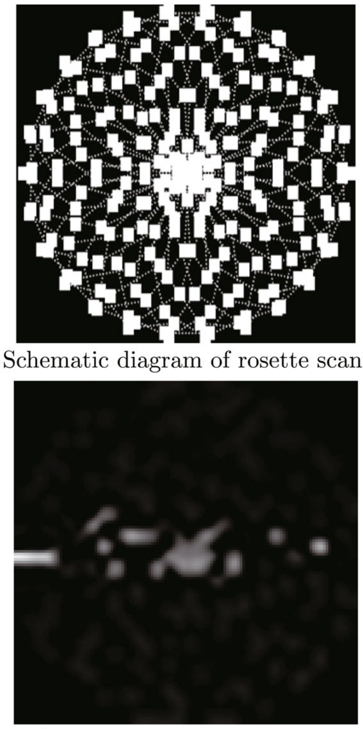


Fig. 2 Schematic diagram of optical mechanical scanning

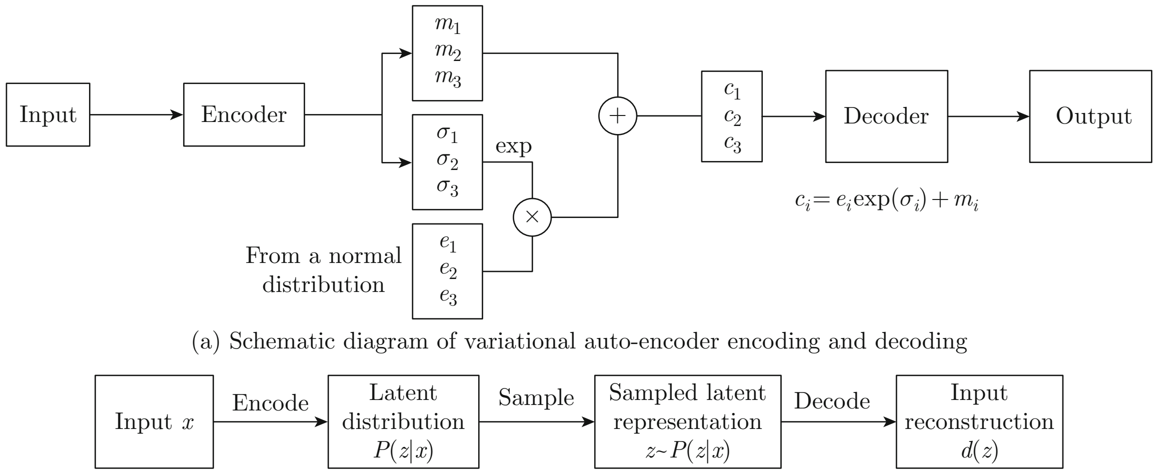

(a) Schematic diagram of rosette scanning

(b) Low-resolution sparse image

Fig. 3 Track map of rosette scanning for one frame and the sparse image after rosette scanning

angles is used to reduce the scanning time as much as possible to obtain more useful information. This paper uses five frames to obtain more pixels and each frame scans 1 000 points. The scanned part of the tag image is displayed as the gray value of the infrared aircraft image and the part not scanned is the black area. The image is compressed by twice bilinear interpolation so that the scanned pixel is the average of the four surrounding pixels. Figure 3(b) shows the obtained low-resolution sparse image with a size of 32×32 which is used as the input of the network. The generator extracts features from the input image and generates a group of output images with similar features of the input image. The discriminator then compares the output image with the tag image. If the discriminator thinks that they are not similar, the generator generates the image again and circulates in turn until the generator generates the image to deceive the discriminator to stop training when it thinks that the image is similar to the tag.

Variational auto-encoder (VAE) is a kind of important generation model similar to GAN. The purpose is to map the input to the late vector during training, and then map the late vector to a Gaussian distribution to get z, and then use the coder to get the same data as the input mode. As shown in Fig. 4(a), in order to add appropriate noise to the code in VAE, the encoder will output two codes. One is the original code (m_1, m_2, m_3) and the other is the code to control the noise interference level $(\sigma_1, \sigma_2, \sigma_3)$. The second coding is to assign weights to random noise codes (e_1, e_2, e_3) and then add $\exp(\sigma_i)$. The purpose of adding $\exp(\sigma_i)$ is to ensure that the assigned weight is a positive value. Finally, the original code and the noise code are added to get the output results of VAE in the code layer (c_1, c_2, c_3). Figure 4(b) is system block diagram of variational auto-encoder. The input image is encoded and mapped to the potential distribution of z, $P(z|x)$. Then the potential distribution is sampled and restored to obtain reconstructed image.


Input
Encoder
m_1 m_2 m_3
σ_1 σ_2 σ_3
exp
From a normal distribution
e_1 e_2 e_3
+
×
c_1 c_2 c_3
Decoder
Output
$c_i = e_i \exp(\sigma_i) + m_i$
Input x
Encode
Latent distribution $P(z|x)$
Sample
Sampled latent representation $z\sim P(z|x)$
Decode
Input reconstruction $d(z)$

(a) Schematic diagram of variational auto-encoder encoding and decoding

(b) System block diagram of variational auto-encoder

Fig. 4 Schematic diagram of variational auto-encoder

VAE is to do further variational processing on the self-coder model so that the output result of the encoder can correspond to the mean and variance of the target distribution. Therefore, it has two encoders: one for calculating the mean and the other for calculating the variance. Essentially, it is the result of encoder based on conventional self-encoder. It means Gaussian noise is added to the network for calculating the mean value, so that the result decoder can be robust to noise. The loss function uses the reconstruction loss and the Kullback Leibler divergence with the standard normal distribution. Adding the KL divergence is actually equivalent to a regular term for the encoder. It is hoped that the model trained by the encoder is more consistent with the standard normal distribution so as to prevent the variance trained by the model from approaching 0 during the training process. As a result, the model loses the ability to generate. Another encoder is used to dynamically adjust the intensity of noise. When the decoder has not been trained, the noise will be appropriately reduced to make fitting easier. On the contrary, if the decoder is trained well, the noise will increase, the fitting difficulty will increase, and the generation ability of the decoder will be improved.

1.2 Generative Countermeasure Network

It is necessary to train the generative network G when the generative adversarial network is applied to the super-resolution field. The image is generated by the forward propagation of the generated network G_{θ_G}. Herein, θ_G can be adjusted by learning the mapping relationship between $\boldsymbol{I}_n^{\mathrm{LR}}$ and $\boldsymbol{I}_n^{\mathrm{HR}}$ $(n = 1, 2, \cdots, N)$ to obtain the optimal parameters which can be expressed in the mathematical formula as

$$\hat{\theta}_G = \arg\min_{\theta_G} \frac{1}{N} \sum_{n=1}^{N} \boldsymbol{I}^{\mathrm{SR}} \left(G_{\theta_G}(\boldsymbol{I}_n^{\mathrm{LR}}), \boldsymbol{I}_n^{\mathrm{HR}} \right). \quad (2)$$

Additionally, $\theta_G = \{W_L; b_L\}$ is the parameter of the generated network representing the weight and deviation of the Lth layer neural network of the generated network, $\boldsymbol{I}_n^{\mathrm{LR}}$ is the input image with a size of 32×32, $\boldsymbol{I}^{\mathrm{SR}}$ is the output image with a size of 128×128, and $\boldsymbol{I}_n^{\mathrm{HR}}$is the label image with a size of 128×128.

The convolutional neural network is used as the generation model to improve the generation of confrontation network, and the loss function based on vision is used to guide the training.

Figure 5 shows the structure diagram of generator. The structure of each dense block is composed of five convolutions, as shown in Fig. 6. Three dense blocks constitute a residual-in-residual dense block (RRDB). Compared with SRGAN, the following two changes have been made in the generator. The first is to remove the BN layer, which can appropriately prevent excessive artifact effects. The BN layer uses the mean and variance of the data of a batch to normalize the characteristics of the batch during training. The mean and variance predicted by the data on the whole test set are used during testing. The BN layer tends to generate bad artifacts and limits the generalization ability of the model when the statistics of the training set and the test set are very different. The second is to change the residual block into RRDB, as shown in Fig. 6. The purpose is to increase the depth of the network. As the depth of the network increases, a deeper receptive field will be obtained, containing more convolutional information. It can use the surrounding environment to increase the amount of information, enough to build a better high-resolution image. Moreover, deeper networks often have more complex nonlinear mappings and complex calculations. The information of each convolution of the residual block is retained as the prior information of the convolution of the next layer or even deeper layer. The dense connection of residual blocks is to avoid the loss of information such as image information and gradient in the forward and reverse transmission process. Dense connection is carried out through channel fusion. The main performance is to get the output of the current layer, which means all the outputs of the previous layer are needed.

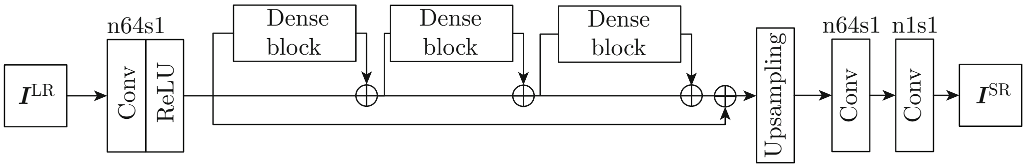


Fig. 5 Structure diagram of generator

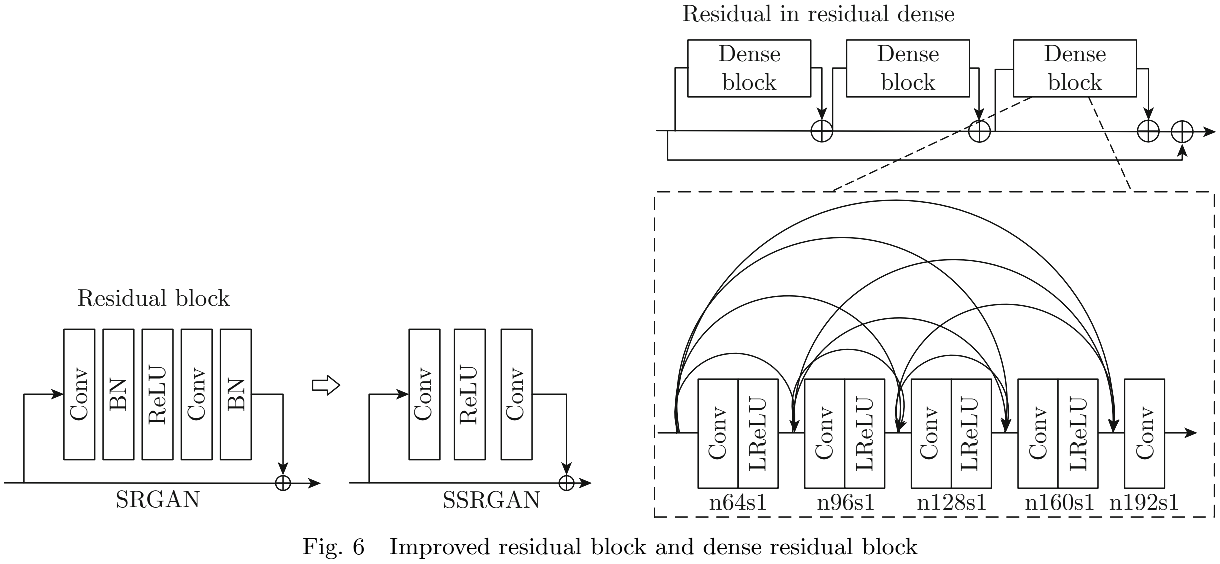


Fig. 6 Improved residual block and dense residual block

The input of the overall generation model is the low-resolution image $\boldsymbol{I}^{\mathrm{LR}}$, which can complete the following operations:

(1) Block mapping and representation. This operation can extract blocks from low-resolution image $\boldsymbol{I}^{\mathrm{LR}}$ and represent each block as high-dimensional vectors

which include a set of features with the same number and vector dimension. Choose to convolute through a set of convolution check images and then use PReLU as the activation function, that is

$$F_1(\boldsymbol{I}^{\mathrm{LR}}) = \max(0, W_1\boldsymbol{I}^{\mathrm{LR}} + b_1) + \alpha_1 \min(0, W_1\boldsymbol{I}^{\mathrm{LR}} + b_1), \tag{3}$$

where W_1 and b_1 are convolution kernel and deviation respectively, the size of W_1 is $1 \times k_1 \times k_1 \times n_1$, k_1 is the space size of convolution kernel, n_1 is the number of convolution kernel, and α_1 is the learnable parameter in the activation function PReLU.

(2) Nonlinear mapping. This operation can nonlinearly map each high-dimensional vector to another high-dimensional vector. Each mapping vector is the representation of super-resolution image blocks. We use two convolution network layers to realize nonlinear mapping:

$$F_2(\boldsymbol{I}^{\mathrm{LR}}) = \max(0, W_2F_1(\boldsymbol{I}^{\mathrm{LR}}) + b_2) + \alpha_2 \min(0, W_2F_1(\boldsymbol{I}^{\mathrm{LR}}) + b_2), \tag{4}$$

$$F_3(\boldsymbol{I}^{\mathrm{LR}}) = \max(0, W_3F_2(\boldsymbol{I}^{\mathrm{LR}}) + b_3) + \alpha_3 \min(0, W_3F_2(\boldsymbol{I}^{\mathrm{LR}}) + b_3), \tag{5}$$

where W_2 and b_2 represent the weights and biases of the second convolutional layer respectively, W_3 and b_3 represent the weights and biases of the third convolutional layer respectively, F_2 and F_3 are feature maps after passing through convolutional layers respectively, and α_2 and α_3 are the learnable parameter in the activation function PReLU.

(3) Reconstruction. This operation aggregates high-resolution blocks and produces super-resolution images $\boldsymbol{I}^{\mathrm{SR}}$ which are expected to be similar to $\boldsymbol{I}^{\mathrm{HR}}$. Two sub-pixel convolution layers are used to magnify the image four times. In order to distinguish the real high-resolution image from the generated $\boldsymbol{I}^{\mathrm{SR}}$ samples, a discrimination model is designed. The discriminant model is to solve the maximum problem as a two-class classifier which extracts image features through convolution layer and finally obtains the probability of sample classification by using full connection layer and sigmoid activation function.

1.3 Discriminator

For the improvement of the discriminator, the discriminator is improved on the basis of relative GAN. Unlike the discriminator in SRGAN, the original discriminator directly determines the probability of whether the generated image is true. In the relative discriminator, there is the probability that the real image is more real than the false image. The discrimination criteria of the original GAN are as follows:

$$\left.\begin{aligned} D(x_{\mathrm{r}}) = \sigma(C_{\mathrm{Real}}) \rightarrow 1 \\ D(x_{\mathrm{f}}) = \sigma(C_{\mathrm{Fake}}) \rightarrow 0 \end{aligned}\right\}, \tag{6}$$

where C_{Real} is the label data input by the discriminator, C_{Fake} is the pseudo data generated by the generator when it is though the discriminator, x_{r} is the label image, x_{f} is the generating image, and σ is the sigmoid function.

The criteria of relative GAN are

$$\left.\begin{aligned} D(x_{\mathrm{r}}, x_{\mathrm{f}}) = \sigma(C_{\mathrm{Real}} - E[C_{\mathrm{Fake}}]) \rightarrow 1 \\ D(x_{\mathrm{f}}, x_{\mathrm{r}}) = \sigma(C_{\mathrm{Fake}} - E[C_{\mathrm{Real}}]) \rightarrow 0 \end{aligned}\right\}, \tag{7}$$

where $E[\cdot]$ is the expected value. The original standard discriminator can be expressed as $D(x) = \sigma(C(x))$, where $C(x)$ is the output result of the discriminator. The purpose of the original GAN is to make the probability of the discrimination result of the real image closer to 1 and the probability of the discrimination result of the generated image closer to 0.

Figure 7 shows the structural diagram of the discriminator. Its input is a tag image or a false image generated by the generator. The output is a probability after a series of convolution networks and discrimination. The discriminator determines whether the input image is a tag image and gives the probability that the image is a tag. Then it sends the parameters back to the generator to prepare for subsequent training.

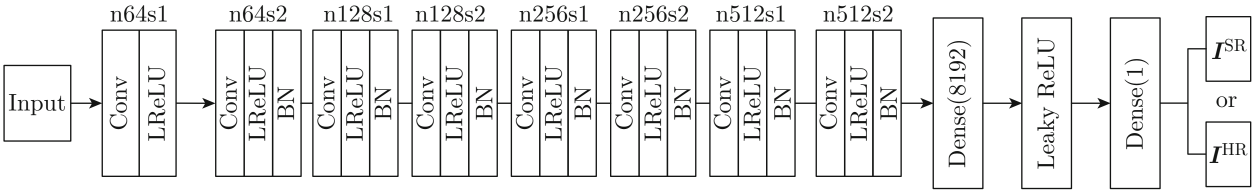


Fig. 7 Structure diagram of discriminator

The improved discriminator is explained as follows. The improved discriminator determines that the original value of the real data judgment is greater than the original value of the generated data judgment. So consider the first equation, which has

$$C_{\mathrm{Real}} - E[C_{\mathrm{Fake}}] > 0. \tag{8}$$

The greater the difference, the greater the difference between the two. That is, the value of the difference after sigmoid is closer to 1. Therefore, the purpose of the discriminator is to make the sigmoid value as close to 1 as possible, which distinguishes the real image from the generated image. Consider the second equation, which has

$$C_{\mathrm{Fake}} - E[C_{\mathrm{Real}}] < 0. \tag{9}$$

The smaller the difference, the farther the distance between the two. That is, the value of the difference after sigmoid is closer to 0. Therefore, the purpose of the discriminator is to make the sigmoid value as close to 0 as possible, which distinguishes the real image from the generated image. The mathematical expression of the loss function of the discriminator is as follows:

$$\begin{aligned} L_{\mathrm{D}} = &- E_{x_{\mathrm{r}}}[\lg(D(x_{\mathrm{r}}, x_{\mathrm{f}}))] - \\ &E_{x_{\mathrm{f}}}[\lg(1 - D(x_{\mathrm{f}}, x_{\mathrm{r}}))], \end{aligned} \tag{10}$$

where $E_{x_{\mathrm{r}}}$ and $E_{x_{\mathrm{f}}}$ represent the operation of averaging all real data and false data in a small batch respectively. Correspondingly, the loss function of the generator is

$$\begin{aligned} L_{\mathrm{G}} = &- E_{x_{\mathrm{r}}}[\lg(1 - D(x_{\mathrm{r}}, x_{\mathrm{f}}))] - \\ &E_{x_{\mathrm{f}}}[\lg(D(x_{\mathrm{f}}, x_{\mathrm{r}}))]. \end{aligned} \tag{11}$$

The average operation is obtained by averaging all the data in the mini-batch. It can be observed that the countermeasure loss includes x_{r} and x_{f}, so this generator benefits from the gradient between the generated data and the actual data in the countermeasure training. This adjustment enables the network to learn sharper edges and more detailed textures.

The discriminator is to better distinguish between reality and generation while the generator is to make it more difficult to distinguish between reality and generation. Because the optimization function of the generator involves both generated data x_{f} and real data x_{r}, it is more conducive to the generation of gradient and the generation of edge and detail information in the image. The algorithm flowchart is shown in Fig. 8. LR is the solution image after rose line sampling whose size is 32 × 32. LR generates high-resolution image SR after convolution layer and up-sampling of the generator whose size is 128 × 128. The gradient is calculated after the original high-resolution image HR and SR pass through the discriminator respectively and add the gradient. The parameters of the generator and the discriminator are updated by the added gradient sum and the discriminator determines whether the generated SR image is real or fake.

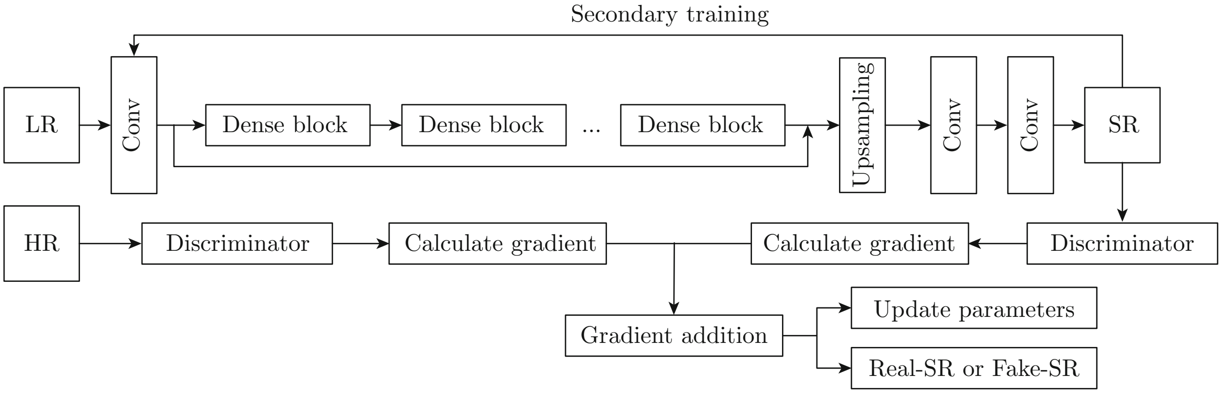


Fig. 8 Block diagram of generative adversarial network training

1.4 Perceptual Loss

A perceptual loss obtained in front of the VGG activation layer is adopted instead of the traditional SRGAN calculating the perceptual loss after the activation layer. The loss of perceptual domain is currently defined in the activation layer of a pre-trained deep network and the distance between two activated features in this layer will be minimized. The features used in this article are pre-activation features which can overcome two shortcomings. First, the features after activation are very sparse, especially in a very deep network. This sparse activation provides a very weak monitoring effect, which will cause low performance. Second, using the activated feature will cause the brightness of the reconstructed image to be inconsistent with the original image. The expression of the generator's loss function is as follows:

$$L_{\mathrm{G}} = L_{\mathrm{percep}} + \lambda L_{\mathrm{G}} + \eta L_1, \tag{12}$$

$$L_{\mathrm{percep}} = \frac{1}{C_j H_j W_j} \|\varphi_j(y) - \varphi_j(\dot{y})\|_2^2, \tag{13}$$

where L_{percep} is the perceived loss, φ represents the loss network, C_j represents the jth layer of the network, $C_j H_j W_j$ represents the size of the characteristic map of the jth layer, $L_1 = E_{x_i}\|G(x_i) - y\|_1$ represents the L_1 norm, λ and η are the correlation coefficients used to adjust the loss function, and $\|\varphi_j(y) - \varphi_j(\dot{y})\|_2^2$ represents the distance between the generated image and the real image. The network optimizes the back propagation of the network through loss. The optimization of the discriminator and generator is actually to adjust the probability that the generator generates the corresponding pixels and finally increase the probability that the generator generates images that are enough to confuse the false with the true.

At the same time, interpolation method is used for secondary training. That is, first use PSNR as the target to train the network GPSNR. Then finely tune the generator in GAN based on GPSNR. Then interpolate the parameters corresponding to the two networks to obtain the final model. The parameter relationship between models can be expressed as

$$\theta_G = (1 - \alpha)\theta_G^{\text{PSNR}} + \alpha\theta_G^{\text{GAN}}, \tag{14}$$

where θ is the parameter in the network, and $\alpha \in [0, 1]$ represents the interpolation parameter. The advantage of this interpolation network is that the interpolation network can produce more meaningful results and balance the perceived quality and fidelity without introducing artifacts.

2 Experiments

2.1 Training Dataset

LR and HR images of all experiments are based on ×4. By using MATLAB to scan the HR image with rose lines, the HR image is sparse and then five frames are combined to show more details. Finally, the HR image is down-sampled by bilinear interpolation to obtain LR image. The minimum batch size is set to 16. Considering that a larger receptive field is more beneficial for capturing semantic information, but it will take more training time at the same time, the size of HR image is finally determined to be 128×128.

The training process is divided into two stages. First, we train a PSNR oriented model with L1 loss. The learning rate is initialized to 2×10^{-4} and every 2×10^{-5} small batch updates. The learning rate decays by 2 times. Then we use the trained PSNR oriented model as the initialization of the generator. Use the loss function to train the generator with $\lambda = 5 \times 10^{-3}$ and $\eta = 1 \times 10^{-2}$. The learning rate is set to 1×10^{-4} and halved at 5×10^4, 1×10^5, 2×10^5, and 3×10^5 iterations separately. Pre-training using pixel by pixel loss is helpful for GAN to obtain visually more satisfactory results. The reason is that it can avoid the undesirable local optimization of the generator. At the same time, the discriminator receives relatively good super-resolution images at the beginning after pre-training rather than extreme false images, which helps it pay more attention to texture identification.

We use Adam for optimization: $\beta_1 = 0.9$ and $\beta_2 = 0.999$. We update the generator and discriminator networks alternately until the model converges. We use two settings for the generator. One contains 16 residual blocks with a capacity similar to SRGAN and the other is a deeper model with 23 RRDB blocks. We implemented our model using pytorch framework and trained with NVIDIA A30.

2.2 Dataset

For training, the self-built infrared aircraft image data set is mainly used which is a high-quality data set for image restoration tasks. A training set contains 13 735 images. Random translation and rotation are used to expand the training data set for data enhancement when training the model. At the same time, the traditional SRGAN and VAE algorithms are trained and the results are compared with SSRGAN.

2.3 Results

The experimental process is shown in Fig. 9. The low-resolution sparse image is used as the input of the system. The super-resolution complete image is generated as much as possible after the trained model.

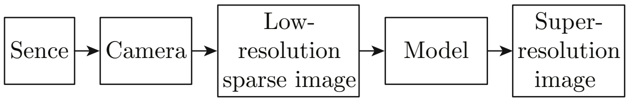
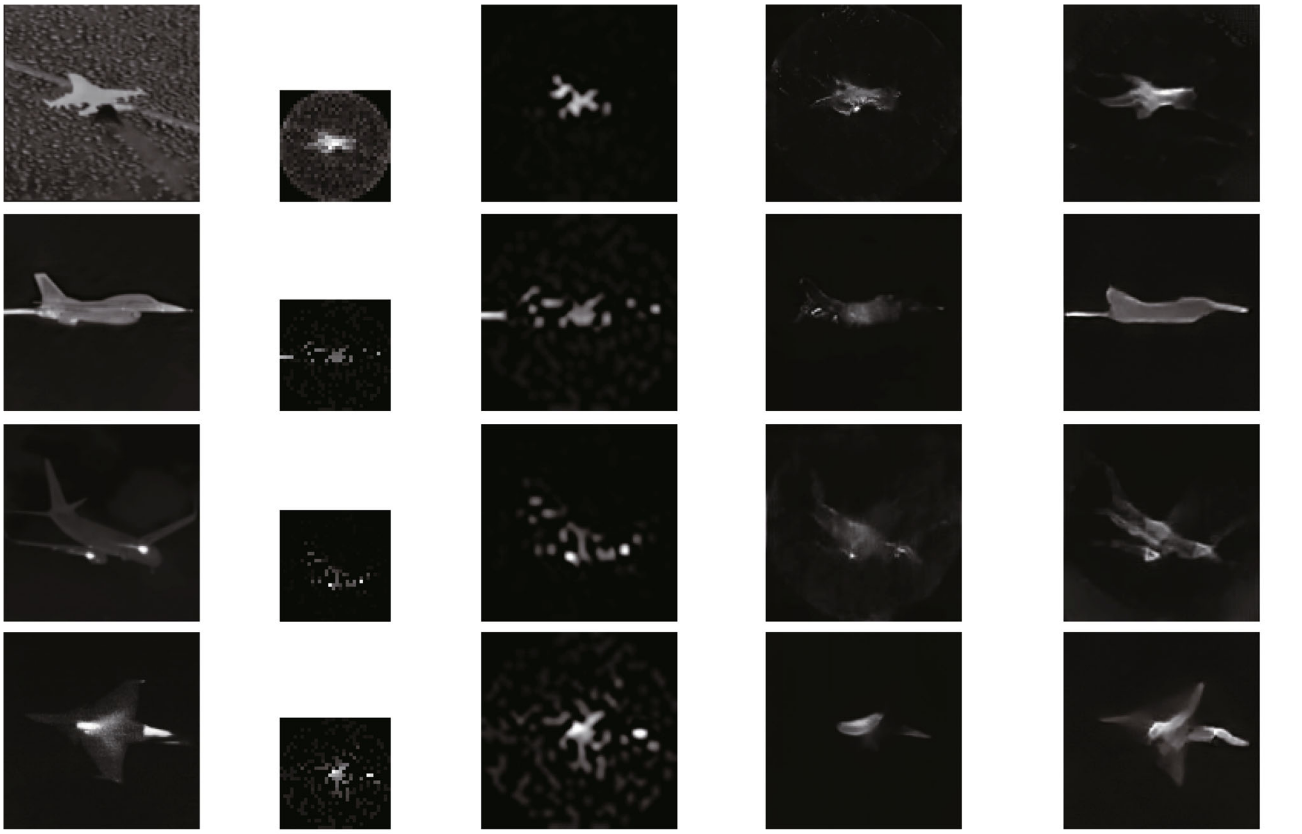


Fig. 9 Experimental testing phase block diagram

Figure 10(a) represents the original picture which is the object body. Figure 10(b) shows the picture after thinning the rose line and down sampling. The information of four adjacent pixels is fused on one pixel and then used as the input image of the network with a size of 32×32 pixels. The images in Figs. 10(a), 10(c)—10(e) are high-resolution images whose size is 128×128 pixels. Figure 10(c) shows the output picture of the VAE. It can be seen that the effect of using VAE to restore and reconstruct the original image or sparse image is not good and there is a great degree of distortion. Figure 10(d) shows the output picture of the traditional SRGAN. Some missing points in Fig. 10(b) are filled in the network and then restored. Considering the image in the second row, the PSNR of the reconstructed image in Fig. 10(c) is 15.23 dB as compared with Fig. 10(a). The PSNR of Fig. 10(d) generated by the network model is 16.18 dB as compared with Fig. 10(a), and the PSNR of the Fig. 10(e) generated based on SSRGAN method used in this paper is 19.26 dB as compared with Fig. 10(a).

It can be seen that the model generated by the improved generative adversarial network has a higher PSNR. At the same time, in terms of sensory quality it can also produce a clearer picture than the original generative adversarial network and also remove part of the redundant information in the picture. The outline



(a) Original image (b) Rose line sparse image (c) Reconstruction image after VAE (d) Super-resolution reconstruction image with SRGAN (e) Super-resolution reconstruction image with SSRGAN

Fig. 10 A series of image results of infrared super-resolution imaging by SRGAN and SSRGAN in this paper

Table 1 Comparison on average PSNR and structural similarity (SSIM) and mean square error (MSE) of different single pixel imaging

Model	PSNR/dB	SSIM	MSE
VAE	13.86	0.33	3 365.74
SRGAN	20.57	0.48	570.27
SSRGAN	21.94	0.56	415.99

is clearer and the object features are also closer to those of original image.

3 Conclusion

This paper uses the excellent remapping ability of deep learning from low-dimensional data to high-dimensional data to solve the inverse ill posed equation in compressed sensing and realizes high-resolution and high-sensitivity infrared imaging which means super-resolution imaging under low signal-to-noise ratio. Secondary training and secondary adaptation for the generative countermeasure network are conducted through the actual generated image set to realize rapid and accurate mapping from the virtual world to the real world.

This paper studies the process of super-resolution reconstruction of infrared single pixel imaging and generates an infrared imaging system with high sensitivity and high resolution. The design idea of combining algorithm with hardware is adopted. The software algorithm is used to improve the resolution while maintaining high sensitivity so as to strengthen the design of the sensory computing integrated system. The infrared image is collected and compressed by hardware system, and the image is restored by software. High resolution infrared image is improved and obtained by using computational imaging method. First, the image is sparsely sampled and then the generated super-resolution image is used for secondary adaptation of the generated countermeasure network. Among them, single pixel imaging is used to improve the sensitivity. In deep learning, the generative adversarial countermeasure network is used to train the sample data to form a neural network that can restore the original image according to the sparse image and achieve high-resolution imaging. The self-built infrared data set is used to train the generative adversarial network. Select the boundary actual data to realize network adaptation and improve the ability of network to adapt to the actual data. The neural network is secondarily adapted by the generated actual high-resolution image set to complete the mapping from virtual data to real data, which increases the accuracy of the learning network.

References

[1] ZHANG Z J, LIU L, LI X R, et al. Compressed sensing for rapid IR imaging [C]//*IET Colloquium on Millimetre-Wave and Terahertz Engineering & Technology 2016*. London: IET, 2016: 1-6.

[2] UZELER H, CAKIR S, AYTAÇ T. Image reconstruction for single detector rosette scanning systems based on compressive sensing theory [J]. *Optical Engineering*, 2016, **55**(2): 023108.

[3] XIE C, LU X, ZENG W. Single frame super-resolution reconstruction based on sparse representation [J]. *Journal of Southeast University* (*English Edition*), 2016, **32**(2): 177-182.

[4] BROMBERG Y, KATZ O, SILBERBERG Y. Ghost imaging with a single detector [J]. *Physical Review A*, 2009, **79**(5): 053840.

[5] SHAPIRO J H. Computational ghost imaging [J]. *Physical Review A*, 2008, **78**(6): 061802.

[6] WANG L, ZHAO S M. Fast reconstructed and high-quality ghost imaging with fast Walsh–Hadamard transform [J]. *Photonics Research*, 2016, **4**(6): 240.

[7] ZHANG Z B, LIU S J, PENG J Z, et al. Simultaneous spatial, spectral, and 3D compressive imaging via efficient Fourier single-pixel measurements [J]. *Optica*, 2018, **5**(3): 315.

[8] ROUSSET F, DUCROS N, FARINA A, et al. Adaptive basis scan by wavelet prediction for single-pixel imaging [J]. *IEEE Transactions on Computational Imaging*, 2017, **3**(1): 36-46.

[9] TSAI R, HUANG T S. Multiframe image restoration and registration [J]. *Computer Vision and Image Processing*, 1984, **1**: 317-339.

[10] KIM J, LEE J K, LEE K M. Deeply-recursive convolutional network for image super-resolution [C]//*2016 IEEE Conference on Computer Vision and Pattern Recognition*. Las Vegas: IEEE, 2016: 1637-1645.

[11] ZHANG D, HE J Z. Hybrid sparse-representation-based approach to image super-resolution reconstruction [J]. *Journal of Electronic Imaging*, 2017, **26**(2): 023008.

[12] TAN J, TAO Z Q, CAO A H, et al. An edge-preserving iterative back-projection method for image super-resolution [J]. *Proceedings of SPIE*, 2016, **10033**: 844-849.

[13] DAVENPORT M A, WAKIN M B. Analysis of orthogonal matching pursuit using the restricted isometry property [J]. *IEEE Transactions on Information Theory*, 2010, **56**(9): 4395-4401.

[14] TROPP J A, GILBERT A C. Signal recovery from random measurements via orthogonal matching pursuit [J]. *IEEE Transactions on Information Theory*, 2007, **53**(12): 4655-4666.

[15] YANG J C, WRIGHT J, HUANG T S, et al. Image super-resolution via sparse representation [J]. *IEEE Transactions on Image Processing*, 2010, **19**(11): 2861-2873.

[16] LIM B, SON S, KIM H, et al. Enhanced deep residual networks for single image super-resolution [C]//*2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops*. Honolulu: IEEE, 2017: 1132-1140.

[17] ZHANG Y L, TIAN Y P, KONG Y, et al. Residual dense network for image super-resolution [C]//*2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Salt Lake City: IEEE, 2018: 2472-2481.

[18] AN Z Y, ZHANG J Y, SHENG Z Y, et al. RBDN: Residual bottleneck dense network for image super-resolution [J]. *IEEE Access*, 2021, **9**: 103440-103451.

[19] ZHU Y, GEIß C, SO E. Image super-resolution with dense-sampling residual channel-spatial attention networks for multi-temporal remote sensing image classification [J]. *International Journal of Applied Earth Observation and Geoinformation*, 2021, **104**: 102543.

[20] ZHANG Y L, LI K P, LI K, et al. Image super-resolution using very deep residual channel attention networks [M]//Computer vision-ECCV 2018. Cham: Springer, 2018: 294-310.

[21] WANG Z H, CHEN J, HOI S C H. Deep learning for image super-resolution: A survey [J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021, **43**(10): 3365-3387.

[22] AYAS S, EKINCI M. Microscopic image super resolution using deep convolutional neural networks [J]. *Multimedia Tools and Applications*, 2020, **79**(21): 15397-15415.

[23] WANG Y F, PERAZZI F, MCWILLIAMS B, et al. A fully progressive approach to single-image super-resolution [C]//*2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*. Salt Lake City: IEEE, 2018: 977-97709.

[24] SHI W Z, CABALLERO J, HUSZÁR F, 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*. Las Vegas: IEEE, 2016: 1874-1883.

[25] CABALLERO J, LEDIG C, AITKEN A, et al. Real-time video super-resolution with spatio-temporal networks and motion compensation [C]//*2017 IEEE Conference on Computer Vision and Pattern Recognition*. Honolulu: IEEE, 2017: 2848-2857.

[26] SAJJADI M S M, SCHÖLKOPF B, HIRSCH M. EnhanceNet: single image super-resolution through automated texture synthesis [C]//*2017 IEEE International Conference on Computer Vision*. Venice: IEEE, 2017: 4501-4510.

[27] WANG X T, YU K, DONG C, et al. Recovering realistic texture in image super-resolution by deep spatial feature transform [C]//*2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Salt Lake City: IEEE, 2018: 606-615.