



Infrared image super-resolution reconstruction by using generative adversarial network with an attention mechanism

Qing-Ming Liu¹ · Rui-Sheng Jia^{1,2} · Yan-Bo Liu¹ · Hai-Bin Sun^{1,2} · Jian-Zhi Yu^{1,2} · Hong-Mei Sun^{1,2}

Accepted: 28 September 2020 / Published online: 23 October 2020
© Springer Science+Business Media, LLC, part of Springer Nature 2020

Abstract

Due to the limitations of infrared imaging principles and imaging systems, many problems are typically encountered with collected infrared images, such as low resolution, insufficient detail information, and blurred edges. In response to these problems, a method of infrared image super-resolution reconstruction that uses recursive attention and is based on a generative adversarial network is proposed. First, according to the characteristics of low-resolution infrared images such as uniform pixel distributions, low contrast, and poor perceived quality, a deep generator structure with a recursive-attention network is designed in this article. The recursive-attention module is used to extract high-frequency information from the feature maps, suppress useless information, and enhance the expressiveness of the features, which facilitates the reconstruction of texture details of infrared images. Then, to better distinguish the reconstructed images from the original high-resolution images, we designed a discriminator that was composed of a deep convolutional neural network. In addition, targeted improvements were made to the content loss function of GAN. We used the pre-trained VGG-19 network features before activation to calculate the perceptual loss, which helps recover the texture details of the infrared images. The experimental results on infrared image datasets demonstrated that the reconstruction performance of the proposed method is higher than those of several typical methods, and it realizes higher image visual quality.

Keywords Infrared images · Super-resolution reconstruction · Attention mechanism · Generative adversarial network

1 Introduction

With the development of infrared sensor technology, infrared images have been widely used in military, aerospace, medical, remote sensing, and other applications [1, 2]. Due to the limitations of infrared imaging principles and hardware performance, compared with natural images, infrared images have low spatial resolution and inconspicuous contrast. The industry mainly improves the hardware performances of infrared imaging systems by improving the manufacturing processes of infrared sensors to increase the spatial resolution of the obtained infrared images. Compared overcoming the

limitations of hardware, it is more economical and practical to increase the resolution of infrared images by using super-resolution reconstruction (SR) technology.

Image super-resolution reconstruction is a typical ill-posed problem [3], in which the objective is to reconstruct visually satisfactory high-resolution (HR) images from one or more low-resolution (LR) images. To solve this ill-posed problem, many scholars have proposed a variety of super-resolution reconstruction methods, which include interpolation-based [4] methods, reconstruction-based [5] methods and learning-based [6–12] methods. Limited by the lack of prior information inside low-resolution infrared images, infrared images that are reconstructed based on interpolation and reconstruction methods still cannot satisfy the requirements of the industry [13]. With the rapid development of deep learning and the emergence of high-performance GPUs, learning-based methods are widely used in image super-resolution reconstruction. Learning-based methods [14, 15] realize image super-resolution reconstruction by training many LR images and HR images in pairs to obtain the mapping relationship between the sample pairs. In 2014, Dong et al. [16] introduced a three-layer convolutional neural network into image super-resolution reconstruction and proposed super-resolution using

✉ Rui-Sheng Jia
jrs716@163.com

✉ Hong-Mei Sun
shm0221@163.com

¹ College of Computer Science and Engineering, Shandong University of Science and Technology, Qingdao 266590, China

² Shandong Province Key Laboratory of Wisdom Mine Information Technology, Shandong University of Science and Technology, Qingdao 266590, China

convolutional neural networks (SRCNNs). The model utilizes a convolution operation instead of traditional manual feature extraction, which can directly learn an end-to-end mapping between LR images and HR images, and has realized significant improvements over traditional methods. The structure of the network is simple, and the training is fast. However, the network model is too small and the range of pixels that can be used in the learning process is too small, thereby resulting in few feature extractions and limited reconstruction performance. In 2016, Kim et al. [17] proposed the deeply-recursive convolutional network for image super-resolution (DRCN), which realized significant progress compared with SRCNN. To further improve the visual quality of the reconstructed images, the generative adversarial network (GAN) [18] was introduced into the field of super-resolution reconstruction. In 2017, Ledig et al. [19] proposed super-resolution using a generative adversarial network (SRGAN) for the generation of more realistic images via the addition of adversarial losses. However, SRGAN is based on ordinary convolution, which treats the extracted image information equally, hence, it is highly inefficient in this calculation. Other algorithms without attention mechanisms all treat the extracted feature maps equally. When reconstructing infrared images, it is easy to produce false textures that are inconsistent with the original image. Methods that are based on deep learning have yielded more accurate results. However, the infrared image pixels are evenly distributed and the gradient range is small, thus, it is difficult for these methods to obtain sufficient effective feature information. The reconstructed image texture remains unsatisfactory in vision applications.

To increase the feature extraction performance and obtain satisfactory high-resolution infrared images, we introduced attention mechanisms [20–22] into SRGAN and proposed a super-resolution reconstruction method for infrared images that was based on the generative adversarial network. Due to the characteristics of low-resolution infrared images, such as low contrast and poor perceived quality, we introduced a recursive-attention network into the generator to extract more high-frequency features of the image and suppress the useless information, thereby enhancing the expressiveness of the features and contributing to the reconstruction of texture details. In addition, since infrared images are less visually recognizable than natural images and contain less key semantic information, a loss function is optimized. One strategy is to use pre-trained VGG-19 network features prior to activation to calculate the perceptual loss, which can increase the image reconstruction accuracy. The other is to use the Wasserstein [23, 24] distance to guide the adversarial training of GAN and ensure its convergence.

2 Related work

Generative adversarial network theory is derived from the two-person zero-sum game in game theory, which is a

probability generation model. It was proposed by Ian Goodfellow [18] in 2014. In recent years, GAN has been widely used in image generation [25], image inpainting [26], image super-resolution reconstruction [19], and other fields due to its powerful image generation capabilities. In image super-resolution reconstruction, SRGAN is the first mature network, which includes a generator (G) and a discriminator (D). The basic framework of SRGAN is illustrated in Fig. 1.

In Fig. 1, the generator is represented by function G , and the discriminator is represented by function D . The parameters of each function can be trained. First, a noise z , which was sampled from LR training images, is sent to G to generate a false-sample x that is similar to HR images that were input into D . At the same time, HR training samples are also sent to D as real samples, and D tries to distinguish the difference between the false-sample x and the real-sample HR. The objective of the generator G is to continuously produce high-resolution images that are as close to the real samples as possible until the generated high-resolution images can deceive the discriminator D , namely, to make $D(G(z))$ as close to 1 as possible. The objective of the discriminator D is to make $D(G)$ as close to 0 as possible, and eventually, it will realize a balance in the mutual game. By adding a discriminant loss to the traditional perceptual loss, SRGAN can generate texture details that are closer to those of the HR images, thereby making the images more realistic. However, due to the uniformity among the infrared image pixels, SRGAN still does not reconstruct some texture details sufficiently clearly, and it may produce fuzzy artifacts. Moreover, the perceptual loss that is based on the VGG-19 classification network cannot capture the high-frequency information that is required for infrared image super-resolution reconstruction.

3 Proposed method

Based on the basic framework of SRGAN, we redesigned the generator to generate infrared images with higher visual quality by introducing a recursion attention module. In addition, the loss function is improved as an optimization strategy to balance the objective evaluation index and the subjective visual quality. The objective of infrared image super-resolution reconstruction is to estimate a high-resolution reconstructed image I_{SR} according to a low-resolution input image I_{LR} such that I_{SR} is as close to the real high-resolution image I_{HR} as possible, which is pursued by training a generator network G to generate a high-resolution image that is as similar to the real high-resolution image I_{HR} as possible. I_{SR} is expressed as:

$$I_{SR} = G_{\theta}(I_{LR}) \quad (1)$$

where I_{LR} is obtained from the corresponding high-resolution image I_{HR} with scale down-sampling factor r and θ represents

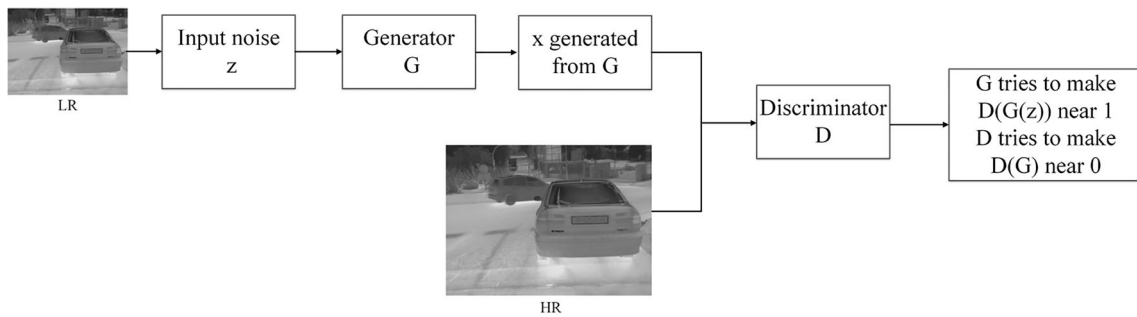


Fig. 1 Basic framework of SRGAN

the parameters of the network, which can be obtained through continuous optimization of the loss function in the adversarial training. The parameters must satisfy the following expression:

$$\theta = \arg \min \sum L^{SR}(I_{SR}, I_{HR}) \tag{2}$$

where $L^{SR}(I_{SR}, I_{HR})$ represents the reconstruction error.

3.1 Design of the generator

Due to its strong feature mapping performance, a neural network can map low-resolution feature maps to high-resolution space; hence, neural networks are often used in super-resolution reconstruction tasks. Infrared images often suffer from a uniform pixel distribution and a small gradient range, and some weak details are not easy to extract. Therefore, we introduced the recursion attention module into the generator structure. The main functions of the recursive attention module are to further extract the important texture information that is needed for image reconstruction and to suppress the useless information. These texture details are especially important for the generation of high-resolution images. The more high-frequency information that is extracted during training, the more accurate the reconstruction results will be. The generator structure is illustrated in Fig. 2.

In Fig. 2, I_{LR} and I_{SR} represent the input and output, respectively of the generator. First, the generator extracts the shallow information of the low-resolution image by using two-layer

convolution, which is accompanied by the PReLU [27] activation function and can be expressed as follows:

$$F_l(I_{LR}) = \max(\lambda * (W_l * F_{l-1}(I_{LR}) + B_l), W_l * F_{l-1}(I_{LR}) + B_l) \tag{3}$$

where W_l is convolution filter l , which can be expressed as $n * k * k * 1$, where n represents the number of convolution kernels and k represents the size of the convolution kernels; B_l is the bias of layer l , $F_{l-1}(I_{LR})$ is the feature map of the previous layer's output; and λ is the learnable parameter in the activation function PReLU.

Then, the extracted shallow feature maps are sent into the recursion attention module as new inputs. The attention layer structure is illustrated in detail in Fig. 3.

There are abundant low-frequency components and a few valuable high-frequency components in the low-resolution infrared image space. The low-frequency part is flatter, and the high-frequency part is typically full of edges, textures, and other details. In super-resolution tasks, high-frequency channel features are more important for reconstruction; thus, we introduced an attention mechanism to focus on such channel features. We can assign attention resources to each feature channel by recursively calling the attention layer. With the increase in the number of recursions, the trained attention maps can increasingly highlight the detailed texture and related structure.

In Fig. 3, GAP represents global average pooling. The attention layer uses this structure to average the information of all points in space into one value so that it can shield part of the

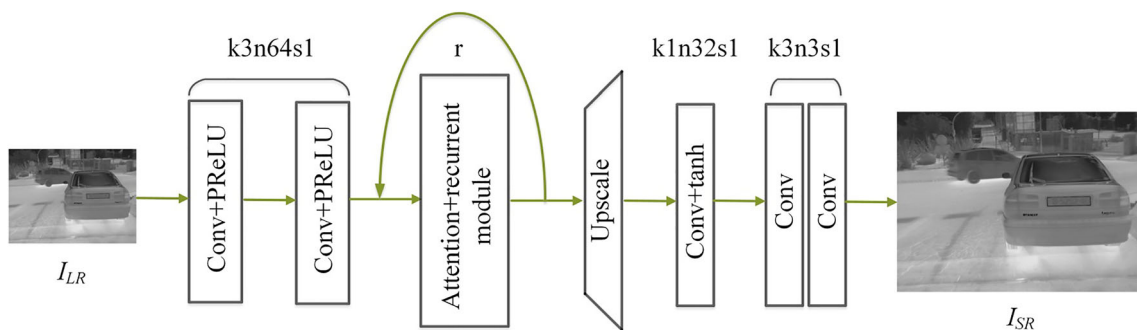


Fig. 2 Generator structure: k is the size of the convolution kernel, n is the number of convolution kernels, s is the convolution step, and r is the number of recursions

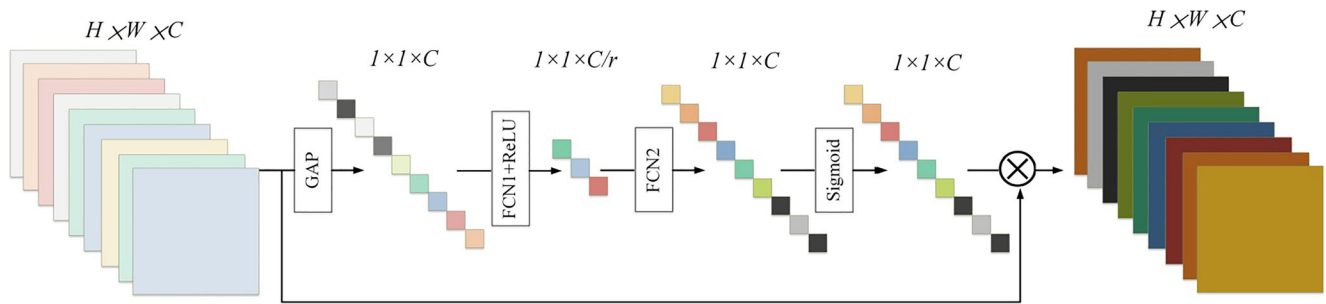


Fig. 3 Schematic diagram of the attention layer

smooth low-frequency information and retain more texture information of features. As illustrated in Fig. 3, feature maps with C channels and of size $H \times W$ undergo GAP to obtain the global information z among the C feature maps. The global information z_c of the c -th feature map is calculated as follows:

$$z_c = H_{GAP}(x_c) = \frac{1}{H \times W} \sum_{x=1}^H \sum_{y=1}^W x_c(x, y) \quad (4)$$

where $x_c(x, y)$ represents the value at position (x, y) in the c -th feature map x_c and H_{GAP} represents the global average pooling function. Such channel statistics help express the entire image information [28]. Then, z passes through a gating unit that is composed of two fully connected layers and a sigmoid function to fuse the feature map information of each channel. The unit is expressed as:

$$s = g(F_2 \delta(F_1 z)) \quad (5)$$

where g and δ represent the sigmoid function and the ReLU function, respectively, and $F_1 z$ represents a fully connected operation. After a ReLU, the dimension is reduced to C/r , where r is the scale factor. Then, after a fully connected layer F_2 , the dimension is increased back to C . Finally, the sigmoid function is used to obtain the weight value of the feature map, which is between 0 and 1, to reduce the disturbance to the feature information, and the attention layer obtains an enhancement matrix with the same size as the input feature. This enhancement matrix is multiplied pixel-by-pixel with the input feature map to realize effective feature enhancement and further suppress unimportant features. Then, the resulting attention feature maps are sent as new input to the next layer of the recursive network.

Assume that the output of the r -layer recursive attention module is F_r , which can be expressed as follows:

$$F_r = H_{ATT,r}(F_{r-1}) = H_{ATT,r}(H_{ATT,r-1}(\dots(H_{ATT,1}(F_0))\dots)) \quad (6)$$

where H_{ATT} represents the function of the attention module, and r represents the number of recursions.

Finally, the feature maps pass through up-sampling layers and three convolutional layers to realize up-sampling with

magnification factors of 2, 3, and 4 to obtain the final HR reconstructed image. After comparing several commonly used up-sampling methods such as transposed convolution, nearest up-sampling + convolution, and sub-pixel convolution [29], we select two sub-pixel convolutional layers, which are superior in terms of computational complexity and performance.

3.2 Design of the discriminator

In addition to improving the generator, we have also made targeted improvements to the discriminator. The improved discriminator network structure and parameter settings are presented in Fig. 4. In the figure, SR is a high-resolution image that is generated by the generator, and HR is the real high-resolution image. The discriminator adopts the deep convolution structure. A study [30] showed that in super-resolution reconstruction tasks, the batch norm (BN) layer tends to destroy image spatial information and reduce the reconstruction performance; hence, we removed the BN layer and used Leaky_ReLU ($\lambda = 0.2$) as the activation function. After the 7-layer convolutions, the feature maps were input into two fully connected layers and classified by the sigmoid activation function to judge the output high-resolution image as true or false.

3.3 Loss function

In the process of network training, the loss function has a substantial influence on the network model. To reconstruct high-resolution infrared images with clear texture, the loss function of SRGAN has been improved. The adversarial loss of GAN is:

$$\min_G \max_D V(G, D) = E[\log D_\theta(I_{HR})] + E[\log(1 - D_\theta(G_\theta(I_{LR})))] \quad (7)$$

where E is the probability expectation of the log function and $G_\theta(I_{LR})$ represents the reconstructed high-resolution image. The overall loss function of this article consists of the following parts.

3.3.1 Pixel-wise loss

To measure the similarity between the high-resolution image that is reconstructed by the network and the real high-resolution image, their mean square error (MSE) is typically calculated in pixels. The objective is to minimize the difference between the reconstructed image and the real high-resolution image in the pixel domain, which is referred to as the pixelwise loss. This loss can be expressed as:

$$L_{pixel} = \frac{1}{W \times H} \sum_{x=1}^W \sum_{y=1}^H \left((I_{HR})_{x,y} - G(I_{LR})_{x,y} \right)^2 \quad (8)$$

where $W \times H$ is the size of the image, and (x,y) is the pixel value in the image.

3.3.2 Perceptual loss

The perceptual loss is defined on the activation layer of the pre-trained deep network, and the corresponding loss function can be calculated from the activated feature values, which can prevent the network from generating blurred images. However, in deeper networks, after activation, the features will become highly sparse, thereby resulting in poor supervision performance. Hence, we used the feature values before activation to calculate the perceptual loss. The features before activation can better represent the feature information of the image; thus, the texture consistency between the reconstructed image and the original image can be well supervised. In this article, the trained VGG-19 network is used to obtain the feature values of the corresponding activation layer before activation. Then the Euclidean distance between the generated high-resolution image and the original image is calculated to obtain the perceptual loss:

$$L_{per} = \frac{1}{W \times H} \sum_{x=1}^W \sum_{y=1}^H \left(\phi_i(I_{HR})_{x,y} - \phi_i(G(I_{LR})_{x,y}) \right)^2 \quad (9)$$

where ϕ_i represents the feature values of the activated layer after the i -th convolutional layer of VGG-19.

3.3.3 Adversarial loss

We used the WGAN loss [31] to stabilize the training process. In WGAN, the loss function is defined as the Wasserstein distance between the target image I_{HR} and the reconstructed image I_{SR} . This is equivalent to adding a regularization term to the discriminator loss function as a gradient penalty strategy, which makes the training smoother. Via this approach, formula (7) can be optimized as follows:

$$L_{WGAN} = E[D_{\theta}(I_{HR}) - D_{\theta}(G(I_{LR}))] + \lambda E \left[\left\| \nabla_I D(I)_2 - 1 \right\|^2 \right] \quad (10)$$

where the second term on the right side of the equal sign is the gradient penalty term, $\nabla_I D(I)_2$ represents the two-norm of the

gradient of the discriminator D , and I represents an image that was randomly selected from the generated image samples and the real image samples.

In our method, the weighted sum function that is obtained by linearly combining three loss functions, namely, the pixel-wise loss, perceptual loss, and adversarial loss functions, is used as the global loss function of our method. It can be expressed as follows:

$$L_G = L_{WGAN} + \alpha L_{pixel} + \beta L_{per} \quad (11)$$

$$L_D = L_{WGAN} + \beta L_{per} \quad (12)$$

where α and β are the linear combination weights of the corresponding loss functions L_{pixel} and L_{per} , respectively, in the target function.

4 Experiment

4.1 Experimental environment and parameter settings

The platform of the experiments is a Windows 10 operating system with an Intel 2.2 GHz i7-8750h CPU with 8 G memory that is configured with an NVIDIA GTX 1080 GPU, and we trained the model under the GPU-based TensorFlow deep learning framework. In the article, the MSRA method [32], which was proposed in reference [30], was used to initialize the weights, and the Adam [33] optimizer with a momentum and weight decay of 0.9 and 0.0001, respectively, was used to optimize the network. The batch size was set to 16, and the initial learning rate was 10^{-4} . Since the generator is completely convolutional, it can be applied to images of any size. The training consisted of two parts: First, the generator was pre-trained with MSE to avoid the local optimization of the GAN network with direct training. The pre-training learning rate was 10^{-4} for 200 iterations. Then, the generator and the discriminator were alternately trained, and the learning rate was 10^{-4} 10,000 iterations. At the same time, the Wasserstein distance was used to optimize the adversarial training. For perceptual loss, the feature values before the activation of the second convolutional layer in the second module of the VGG-19 network were used to calculate the perceptual loss. All the above parameter values of our experiments are the values that yielded the best results in multiple sets of experiments.

4.2 Datasets

Few public infrared image data sets are available for selection. In the experiments, the infrared parts of the CVC-09 and CVC-14 datasets were selected for training and testing. We also built an infrared data set for directly evaluating the

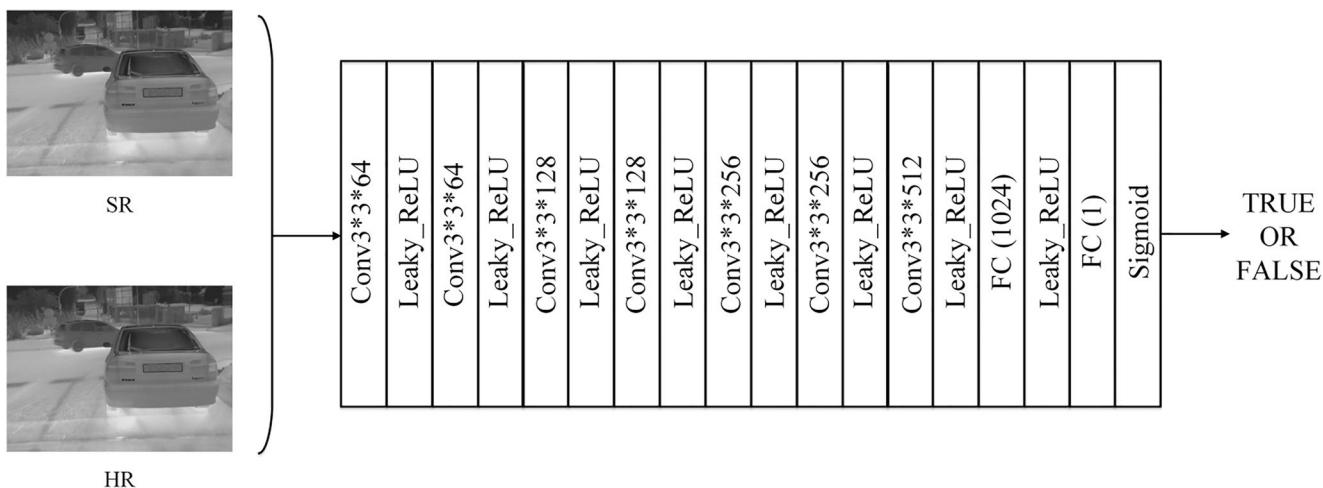


Fig. 4 Discriminator structure and parameter settings

performance of our model. Our self-built infrared data set was collected using refrigerated infrared cameras and the images included pedestrians, cars, traffic signs, and buildings. We selected 1000 infrared images from the CVC-09 dataset as the high-resolution images of the training set and down-sampled them using the bicubic interpolation method to obtain low-resolution images. The down-sampling factors are 2, 3, and 4, respectively. In addition, infrared images from the CVC-14 data set and the self-built data set were selected as the testing set. Due to the small number of images in the training set, it is necessary to enhance the training set to better match the proposed model. The enhancement method rotates to rotate the original images clockwise by 0°, 90°, 180°, and 270° and mirrors them to obtain 8000 images in total. Original images from the training set are shown in Fig. 5.

4.3 Evaluation indices

Image quality evaluation is important for judging the performances of super-resolution reconstruction algorithms. The

evaluation methods of super-resolution reconstruction can be divided into subjective evaluation and objective evaluation. Subjective evaluation refers to the evaluation of image quality according to the subjective feelings of the experimenters. SRGAN proposed a mean opinion score (MOS) [19], which requires a specified number of raters to score the resulting images of super-resolution reconstruction methods to evaluate their subjective visual quality. However, this index requires large labor cost and cannot be reproduced; hence, more accurate image quality evaluation methods are urgently needed. However, before proposing the new evaluation method, two commonly used full-reference image quality evaluation methods, namely, PSNR and SSIM, were used to objectively evaluate the model. The two evaluation indices differ in terms of visual perception, but both involve reference images for comparison. However, in a real infrared image super-resolution reconstruction scenario, only low-resolution images must be reconstructed, but no corresponding high-resolution reference images are available; hence, it is necessary to introduce quantitative free-reference image quality evaluation



Fig. 5 Sample images from the training dataset

Table 1 Average PSNR(dB)/SSIM values on Test1 and Test2 of the proposed method and five other SR methods

Dataset	Scale	Bicubic PSNR/SSIM	SRCNN PSNR/SSIM	DRCN PSNR/SSIM	SRGAN PSNR/SSIM	ESRGAN PSNR/SSIM	Ours PSNR/SSIM
Test1	2	35.229/0.980	37.823/0.934	38.382/0.937	37.553/0.931	37.653/0.930	38.689/0.948
	3	32.008/0.856	33.971/0.909	34.590/0.919	33.595/0.899	33.580/0.909	34.703/0.912
	4	30.242/0.795	31.503/0.840	32.142/0.867	31.039/0.856	31.033/0.861	32.230/0.872
Test2	2	31.280/0.835	32.001/0.845	32.905/0.869	31.824/0.844	32.258/0.855	33.176/0.874
	3	28.169/0.811	28.906/0.830	29.988/0.858	29.079/0.831	29.909/0.835	30.566/0.862
	4	22.683/0.758	24.067/0.794	25.005/0.821	24.055/0.791	24.066/0.807	25.697/0.826

methods. In this article, the average gradient [34] (AG) and the natural image quality evaluator [35] (NIQE) are used as methods of image quality evaluation without reference images.

The peak signal-to-noise ratio (PSNR) is an objective image evaluation index that is based on the error between corresponding pixels, which is defined as follows:

$$PSNR = 10 \times \lg \frac{(2^n - 1)^2}{MSE} \quad (13)$$

where MSE represents the mean square error between images, which is expressed as:

$$MSE = \frac{1}{H \times W} \sum_{x=1}^H \sum_{y=1}^W [I_{HR}(x, y) - I_{SR}(x, y)]^2 \quad (14)$$

The smaller the MSE value is, the larger the PSNR value is and, thus, the closer the image is to the comparison image.

The structural similarity (SSIM) measures the similarity of images from three aspects: brightness, contrast, and structure. Its value ranges from 0 to 1. The larger the value, the lower the distortion of the image.

$$SSIM(x, y) = L(x, y)C(x, y)S(x, y) \quad (15)$$

where $L(x, y)$, $C(x, y)$ and $S(x, y)$ represent the brightness, contrast, and structure, respectively.

The average gradient (AG) refers to the average value of the gray level change rate, which is often used to express the image clarity. The larger the value is, the clearer the image is. It is defined as:

$$G(x, y) = d_x(i, j) + d_y(i, j) \quad (16)$$

$$d_x(i, j) = I(i + 1, j) - I(i, j) \quad (17)$$



Fig. 6 Visual result comparison of various methods and the proposed model on Test1-car with an upscaling factor of 3. (a) Ground-truth HR; (b) Bicubic interpolation; (c) SRCNN; (d) DRCN; (e) SRGAN; (f) ESRGAN; and (g) Ours



Fig. 7 Visual result comparison of various methods and the proposed model on Test2-people with an upscaling factor of 4. **(a)** Ground-truth HR; **(b)** Bicubic interpolation; **(c)** SRCNN; **(d)** DRCN; **(e)** SRGAN; **(f)** ESRGAN; and **(g)** Ours

$$d_y(i, j) = I(i, j + 1) - I(i, j) \tag{18}$$

where I is the value of the image pixel, (i, j) is the coordinate of the pixel, and x and y represent the horizontal and vertical directions, respectively.

The natural image quality evaluator (NIQE) is an image quality evaluation algorithm that was proposed by the University of Texas. It constructs a series of features for measuring image quality and uses these features to fit a multivariate Gaussian model to measure the differences in the multivariate distribution of the image. The smaller the value, the clearer the image. It can be defined as:

$$\hat{T}(x, y) = \frac{I(x, y) - \mu(x, y)}{\sigma(x, y) + 1} \tag{19}$$

$$\mu(x, y) = \sum_{k=-M}^M \sum_{l=-N}^N \omega_{k,l} I(x + k, y + l) \tag{20}$$

$$\sigma(x, y) = \sqrt{\sum_{k=-M}^M \sum_{l=-N}^N \omega_{k,l} [I(x + k, y + l) - \mu(x, y)]^2} \tag{21}$$

where M and N are the length and width, respectively, of the image, and ω is a Gaussian weight function.

4.4 Results and analysis

We mainly investigated the reconstruction effect of the model on low-resolution infrared images with down-sampling factors of 2, 3, and 4. Test1 and Test2 were composed of 15 images from the CVC-14 dataset and 9 images from the self-built dataset, respectively, and were compared using bicubic interpolation, SRCNN [16], DRCN [17], SRGAN [19], and ESRGAN [30]. All methods are retrained and tested on our training set, and the codes are the source codes that were published by the authors. The quantitative results of the peak signal-to-noise ratio (PSNR) and structure similarity index (SSIM) are presented in Table 1. The larger the values, the better the reconstruction effect.

Table 1 quantitatively compares the proposed method with five other super-resolution methods. The PSNR and SSIM values of the proposed method mostly exceed those of the

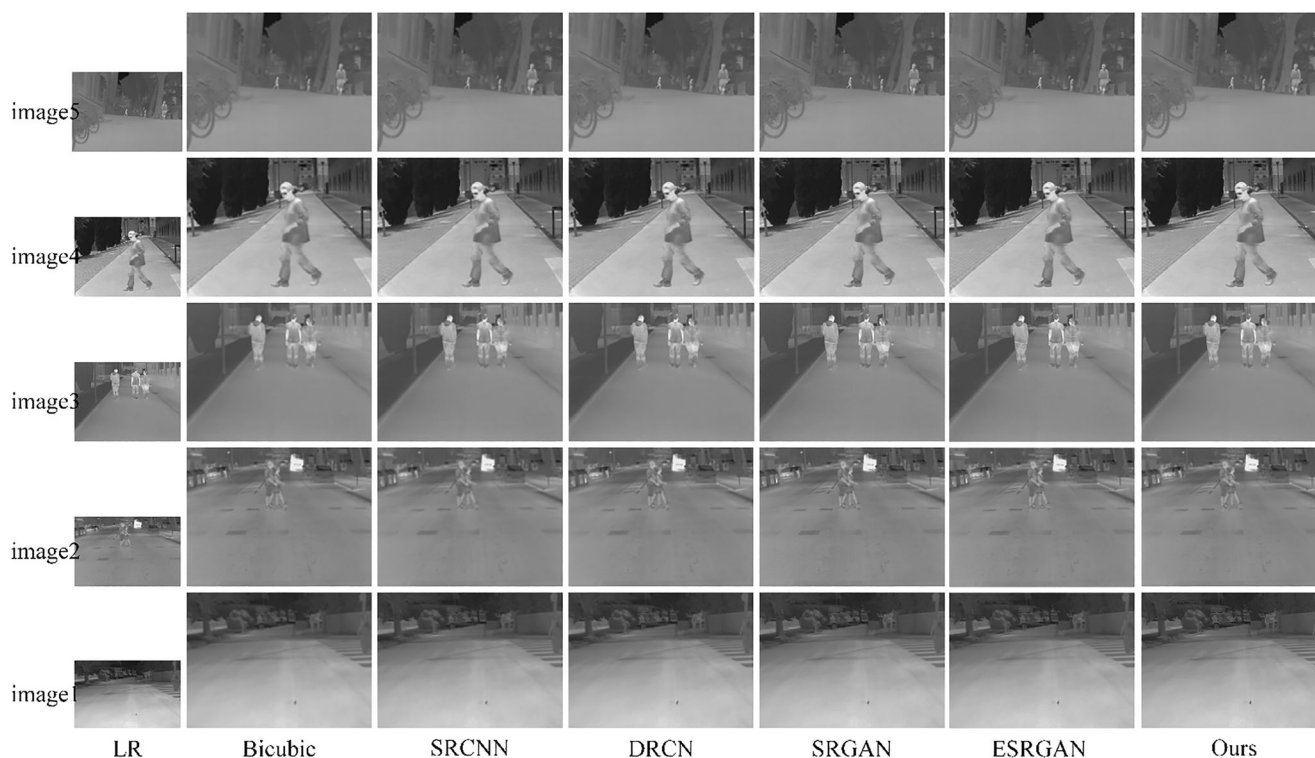


Fig. 8 Actual magnifications of 5 images in Test2 via various methods, the scale factor is 2

other methods under the same magnification factors and the same dataset. In Table 1, the PSNR and SSIM values of the proposed method are significantly improved compared to those of SRGAN and ESRGAN, and those of the DRCN model are the closest to ours. The DRCN model is based on the deep convolutional structure of the MSE loss, and it does not pursue the improvement of visual effects. Similar results are obtained in the article, which proved that the proposed recursive-attention structure and the optimization of the loss function that are adopted in the article are effective. When the scale factor is 3, compared with the ESRGAN model, the PSNR of the proposed method increased by approximately 0.89 dB and the SSIM by approximately 0.015.

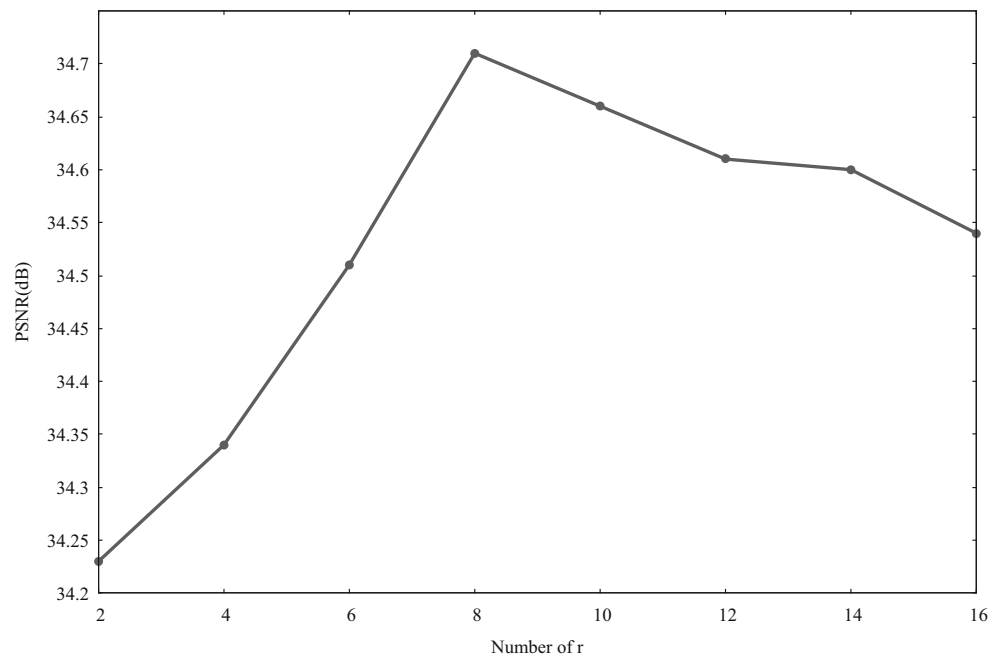
Figures 6 and 7 present the qualitative evaluation results of the above methods on test1-car and test2-people, respectively. To facilitate the visual evaluation, various areas in the above pictures are marked with boxes and enlarged separately.

According to Figs. 6 and 7, the bicubic interpolation method produces the worst visual effect, and the overall picture is relatively fuzzy. The reconstruction performance of SRCNN is only slightly higher than that of bicubic interpolation, and the details of the picture are still blurred. Compared with the previous two methods, the reconstruction quality of the DRCN method has significantly improved, and the image is clearer, but some of its details are too sharp. As shown in Fig. 6d, the license plate even produces pseudo textures, which is not in line with the visual perception of human eyes. The proposed method has realized the most satisfactory visual effect. Compared with SRCNN and DRCN, the addition of the adversarial loss can produce more visually satisfactory results. The model is no longer limited to the simple pixel-wise loss but generates high-quality images that are closer to human visual perception. Compared with SRGAN and ESRGAN, the recursive-attention module can coordinate the relationship

Table 2 Average AG/NIQE values of 5 images in Test2 for the proposed method and various SR methods, with a scale factor of 2

Data	Scale	Bicubic AG/NIQE	SRCNN AG/NIQE	DRCN AG/NIQE	SRGAN AG/NIQE	ESRGAN AG/NIQE	Ours AG/NIQE
image1	2	4.589/8.768	5.173/8.074	5.443/7.558	5.411/7.499	5.471/7.485	5.508/7.445
image2	2	4.819/6.855	6.506/6.491	6.801/5.886	7.010/5.690	7.110/5.685	7.125/5.305
image3	2	3.929/7.454	4.483/6.511	4.574/6.080	4.788/5.964	4.788/5.961	4.707/5.788
image4	2	6.425/6.876	8.147/5.846	8.311/5.506	8.808/5.377	8.882/5.067	8.881/4.998
image5	2	2.936/7.478	3.358/6.590	3.421/6.757	3.645/5.206	3.668/5.116	3.587/5.114
Average	2	4.540/7.486	5.533/6.702	5.710/6.357	5.932/5.947	5.984/5.863	5.962/5.730

Fig. 9 The number r of the attention module versus the performance for scale factor 3 on the Test1 dataset



between the feature information in the image well and can use the global features of the image to generate a clearer high-resolution image. In conclusion, compared with other methods, the proposed method is not only yields more realistic visual restoration results but also produces sharper edges and clearer textural fineness in the enlarged area. The proposed method has improved the visual quality of the reconstructed image while simultaneously reducing the generation of false textures. The reconstruction performance is better than those of the comparison methods; hence, the proposed method is practicable.

To further evaluate the efficiency of the proposed model in real scenes, we randomly selected five images from the Test2 to conduct another set of comparative experiments. In the experiment, we directly input the LR test images into the network without down-sampling, which is more in line with the actual degradation process. The results are shown in Fig. 8.

Due to the lack of reference high-resolution images, we used two representative indices (AG and NIQE) for evaluation, and the comparison results are presented in Table 2. AG and NIQE can reasonably evaluate the sharpness of images and can reflect the content sharpness, detail contrast, and

texture diversity of images. The larger the AG value, the smaller the NIQE value and the clearer the image. According to Table 2, ESRGAN has obtained the optimal value for AG, whereas the proposed method has obtained the optimal value for NIQE. That may be because the quality of our self-built dataset is relatively high and the pixel sharpness is relatively high. When we reconstruct directly, the visual effect is better. Hence, the significant improvement of the visual expression and quantitative indices of the reconstructed image has fully proven the effectiveness and practicability of the proposed method. The combination of recursive-attention learning and adversarial learning strategies has shown substantial advantages in the super-resolution reconstruction of infrared images.

4.5 Model analysis

4.5.1 Recursive learning

The recursive learning of the attention module can reduce the number of parameters of the network and the storage demand while extracting the feature information of infrared images. However, as the number of recursions increases, the gradient

Table 3 Average PSNR/SSIM values of Test1 and Test2 on various loss functions, with a scale factor of 3

Loss	Scale	Test1		Test2	
		PSNR	SSIM	PSNR	SSIM
$L_{pixel} + L'_{per}$	$\times 3$	34.148	0.886	29.937	0.829
$L_{pixel} + L_{WGAN} + L'_{per}$		34.586	0.908	30.421	0.853
$L_{pixel} + L_{WGAN} + L_{per}$		34.703	0.912	30.566	0.862

disappearance of the network will become increasingly severe, and the overall training difficulty will substantially increase. Therefore, we examined the relationship between the number of recursions r of the attention module and the reconstruction performance of the model. The performance evolution of the model with various values of r and a scale factor of 3 on dataset Test1 is presented in Fig. 9. The average PSNR of the model gradually increases with the increase of r initially, and the maximum is attained when $r = 8$, after which the PSNR begins to decline slowly. Therefore, based on the reconstruction speed and the training difficulty, the number of recursions r of the attention module in our implementation of the model was set to 8.

4.5.2 Loss functions

We conducted a set of experiments to evaluate the influence of the loss function on the super-resolution performance. The experimental results are presented in Table 3, where L'_{per} represents the perceptual loss using the feature values of VGG-19 after activation and L_{per} represents the loss using the feature values of VGG-19 before activation. The experimental results demonstrate that the average PSNR and SSIM values of the reconstructed images were severely degraded without the Wasserstein distance loss. The model that uses the feature values of VGG-19 before activation slightly outperforms the model that uses the feature values of VGG-19 after activation. This is because the perceptual loss is constrained by the use of the features of the VGG-19 network before activation, which can better supervise the texture restoration of infrared images. This proved the effectiveness of the improvement. The experimental results have demonstrated that the proposed model can reconstruct infrared images with realistic texture.

5 Conclusions

We present a novel model that uses recursive attention learning that is based on generative adversarial network to super-resolve infrared images. We design a generator network with recursive-attention modules, which can adaptively adjust the feature channel information and enhance the expressiveness of features. The recursive attention learning strategy can not only make the texture of infrared images more natural and realistic but also reduce the generation of pseudo textures. For the loss function, feature values of the VGG-19 network before activation are used to constrain the perceptual loss, which can be better monitored recursively. The Wasserstein distance is also used to optimize adversarial training and to increase the stability of the network training. The experimental results have demonstrated that the proposed model has a significant effect on the super-resolution reconstruction of

vehicle infrared images. Compared with those of several advanced models, the objective indices of the reconstructed images have improved significantly, along with the visual texture details. However, the proposed method did not perform well on other images, such as natural images and thermal infrared images. Increasing the general performance of our model will be the focus of our next study.

Acknowledgements The authors are grateful for collaborative funding support from the Natural Science Foundation of Shandong Province, China (ZR2018 MEE008) and the Key Research and Development Project of Shandong Province, China (2019JZZY020326, 2019GGX101066).

References

1. Chen Q (2013) The status and development trend of infrared image processing technology. *Infrared Tech* 35(6):311–318
2. Yu X, Ye X, Gao Q (2019) Pipeline image segmentation algorithm and heat loss calculation based on gene-regulated apoptosis mechanism. *Int J Press Vessel Pip*, 329–336
3. Baker S, Kanade T (2002) Limits on super-resolution and how to break them. *IEEE Trans Pattern Anal Mach Intell* 24(9):1167–1183
4. Zhang L, Wu X (2006) An edge-guided image interpolation algorithm via directional filtering and data fusion. *IEEE Trans Image Process* 15(8):2226–2238
5. Zhang K, Gao X, Tao D (2012) Single image super-resolution with non-local means and steering kernel regression. *IEEE Trans Image Process* 21(11):4544–4556
6. Timofte R, De Smet V, Van G (2014) A+: adjusted anchored neighborhood regression for fast super-resolution. In *Asian Conf Comput Vis (ACCV)*, 111–126
7. Huang J, Singh A, Ahuja N (2015) Single image super-resolution from transformed self-exemplars. In *Proceedings of the Proc IEEE Conf Comput Vis Pattern Recognit*, 5197–5206
8. Tong T, Li G, Liu X (2017) Image super-resolution using dense skip connections. In *Proc IEEE Int conf comput vis (ICCV)*, 4809–4817
9. Liu X, Jia R, Liu Q, Zhao C, Sun H (2019) Coastline extraction method based on convolutional neural networks—a case study of Jiaozhou Bay in Qingdao, China. *IEEE Access* 7:180281–180291
10. Zhang K, Zuo W, Zhang L (2018) Learning a single convolutional super-resolution network for multiple degradations. In *Proc IEEE Conf Comput Vis Pattern Recognit (CVPR)*, 3262–3271
11. Mao Q, Sun H, Liu Y, Jia R (2019) Fast and efficient non-contact ball detector for picking robots. *IEEE Access* 7:175487–175498
12. He K, Cao X, Shi Y (2019) Pelvic organ segmentation using distinctive curve guided fully convolutional networks. *IEEE Trans Med Imaging* 38(2):585–595
13. Cui Y, Schuon S, Chan D (2010) 3D shape scanning with a time-of-flight camera. In *Proc IEEE Conf Comput Vis Pattern Recognit (CVPR)*, 1173–1180
14. Tian Y, Jia R, Xu S, Hua R, Deng M (2019) Super-resolution reconstruction of remote sensing images based on convolutional neural network. *J Appl Remote Sens* 13(4):046502
15. Liu Q, Jia R, Zhao C, Liu X, Sun H (2020) Face super-resolution reconstruction based on self-attention residual network. *IEEE Access* 8:4110–4121
16. Dong C, Loy C, He K (2014) Learning a deep convolutional network for image super-resolution. *Eur Conf Comput Vis Springer*: 184–199

17. Kim J, Lee J, Lee K (2016) Deeply-recursive convolutional network for image super—resolution. In Proc IEEE Conf Comput Vis Pattern Recognit (CVPR), 1637-1645
18. Goodfellow I, Pougetabadie J, Mirza M (2014) Generative adversarial nets. In Proc Annu Conf Neural Inf Process Syst (NIPS), 2672–2680
19. Ledig C, Theis L, Huszár F (2017) Photo-realistic single image super-resolution using a generative adversarial network. In Proc IEEE Conf Comput Vis Pattern Recognit (CVPR), 105–114
20. Qian R, Tan R, Yang W (2018) Attentive generative adversarial network for raindrop removal from a single image. In Proc IEEE Conf Comput Vis Pattern Recognit (CVPR), 2482–2491
21. Zhang H, Goodfellow I, Metaxas D (2017) Self-attention generative adversarial networks. Mach Learn arXiv preprint arXiv: 1805.08318 :1805.08318
22. Zhang Y, Li K, Li K (2018) Image super-resolution using very deep residual channel attention networks. In Proc Eur Conf Comput Vis (ECCV), 294-310
23. Adler J, Lunz S (2018) Banach Wasserstein GAN. Neural Inf Process Syst (NIPS), 6755–6764
24. Gulrajani I, Ahmed F, Arjovsky M, Dumoulin V, Courville A (2017) Improved training of Wasserstein GANs In Adv Neural Inf Proces Syst (NIPS), 5769-5779
25. Berthelot D, Schumm T, Metz L (2017) BEGAN: boundary equilibrium generative adversarial networks. arXiv preprint arXiv: 1703.10717
26. Yeh R, Chen C, Lim T (2017) Semantic image inpainting with deep generative models. In Proc IEEE Conf Comput Vis Pattern Recognit (CVPR), 6882-6890
27. Xu C, Liu T, Tao D (2016) Local Rademacher complexity for multi-label learning. IEEE Trans Image Process 25(3):1495–1507
28. Hu J, Shen L, Sun G (2018) Squeeze-and-excitation networks. In Proc IEEE Conf Comput Vis Pattern Recognit (CVPR), 7132-7141
29. Shi W, Caballero J, Huszár F (2016) Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In Proc IEEE Conf Comput Vis Pattern Recognit (CVPR), 1874-1883
30. Wang X, Yu K, Wu S (2018) ESRGAN: enhanced super-resolution generative adversarial networks. In Proc Euro Conf Comp Vis (ECCV), 63-79
31. Arjovsky M, Chintala S, Bottou L (2017) Wasserstein generative adversarial networks. Int Conf Mach Learn (ICML), 214-223
32. He K, Zhang X, Ren S (2015) Delving deep into rectifiers: surpassing human-level performance on ImageNet classification. In IEEE Int Conf Compu Vis (ICCV), 1026–1034
33. Kingma D, Ba J (2014) Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980
34. Chen A, Chen B, Chai X, Bian R, Li H (2017) A novel stochastic stratified average gradient method: convergence rate and its complexity. arXiv preprint arXiv:1710.07783
35. Mittal A, Soundararajan R, Bovik A (2013) Making a “completely blind” image quality analyzer. IEEE Signal Process Lett 20(3):209–212



Qing-Ming Liu was born in Shandong, China, in 1995. He received the B. S. degree from Shandong University of Science and Technology, China, in 2018. He is currently pursuing the M.S degree with Shandong University of Science and Technology. His research interests include image processing and deep learning.



Rui-Sheng Jia is currently a Full professor at the College of Computer Science and Engineering, Shandong University of Science and Technology, China. He has more than 30 first-author publications and has more than 25 co-author publications. His research interest includes artificial intelligence, big data processing, information fusion, microseismic monitoring and inversion.



Yan-Bo Liu was born in Shandong, China, in 1996. He received the B. S. degree from Liaocheng University, China, in 2018. He is currently pursuing the M.S degree with Shandong University of science and technology. His research interest includes image processing and deep learning.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Hai-Bin Sun is currently an associate professor at the College of Computer Science and Engineering, Shandong University of Science and Technology, China. He has more than 10 first-author publications. His research interests include artificial intelligence, internet of things, embedded system and industry information system applications.



Hong-Mei Sun received the B.S. and M.S. degree in computer science from Shandong University of Science and Technology, China, in 1995 and 2005, respectively. She is currently a associate professor at the College of Computer Science and Engineering, Shandong University of Science and Technology, China, and is the leader of a Key Research and Development Projects of Shandong Province, China. She has four first-author publications

and has five co-author publications. Her research interests include micro-seismic monitoring technology and software engineering.



Jian-Zhi Yu is a teacher at the College of Computer Science and Engineering, Shandong University of Science and Technology, China. He received the PhD degree from Tongji University, China, in 2012. He is the leader of a Key R&D Program of Shandong Province, China. He has more than 10 first-author publications and has three patents for the invention of the country. His research interest includes artificial intelligence and high performance computing.