Journal of Network Intelligence ©2024 ISSN 2414-8105
Taiwan Ubiquitous Information Volume **9**, Number **4**, November **2024**

Image Super-Resolution Reconstruction Based on Self-Attention Mechanism and Deep Generative Adversarial Network

Yu-Feng Zhao*

School of Electronics and IoT Engineering
Chongqing Industry Polytechnic College, Chongqing 401120, P. R. China
zhaoyf@cqipc.edu.cn

Jie He

Lee Kong Chian Faculty of Engineering and Science
Tunku Abdul Rahman University, Kuala Lumpur 43000, Malaysia
hejie@cqipc.edu.cn

*Corresponding author: Yu-Feng Zhao

Received October 17, 2023, revised January 27, 2024, accepted April 29, 2024.

Abstract. *Both theoretically and practically, the study of super-resolution is of great importance and has become one of the hotspots for research in related fields. However, when image super-resolution reconstruction enlarges the pixels of an image by a larger magnification, the image looks less smooth, which leads to the edges of the target object in the interior of the image look blurred. Therefore, a super-resolution reconstruction method based on self-attention mechanism and deep generative adversarial network is proposed. First, in order to better utilize the low-level features in the deep generative adversarial network, an improved generator network module is designed, which adopts a dense connectivity structure to integrate the functions of each layer, and can make full use of the functions of multiple layers. Then, a pixel self-attention module was designed. The semantic dependencies on the spatial and channel dimensions are modeled, and the upsampling and downsampling modules are redesigned to reduce the loss of image detail information. Finally, the extracted shallow information is divided into three scales and feature encoding-decoding reconstruction is performed at different scales. Experimental results on three image super-resolution datasets, BSDS100, SET14 and Urban100, show that the proposed method exhibits better performance in all metrics compared to multiple existing super-resolution reconstruction algorithms.*

Keywords: image super-resolution; deep learning; generative adversarial networks; self-attention mechanism; multiscale

1. **Introduction.** Spatial resolution is a reflection of the quality of an image. High spatial resolution images can provide more detailed information about the imaged scene due to their high pixel density and high image quality. Super-Resolution (SR) of images [1, 2] is the process of generating a high spatial resolution image about an imaging scene based on one or more low-resolution images of the scene, and has become an active research direction in the fields of computer vision and image processing.

Spatial resolution is an important metric regarding the quality description of digital images [3, 4]. Higher spatial resolution of an image means higher pixel density of that image, richer texture details and sharper image. The need for high spatial resolution images is often faced for various fields based on image processing. Although it is possible to acquire digital images with high spatial resolution with the help of some hardware techniques, this way of acquiring high resolution images will lead to high cost [5, 6]. For the general image acquisition process, the limitations of the inherent resolution of the imaging device's own sensors, as well as the influence of different imaging environments, make the image quality lower, and it is impossible to directly obtain high spatial resolution images [7]. Super-resolution reconstruction of images is the process of combining one or more low spatial resolution images of the same scene with high spatial resolution images of the same scene based on signal processing theory by borrowing software technology.

With the wide range and depth of digital image application areas, it makes a wide range of needs for how to improve image quality with the help of image spatial resolution for existing images. For example, in mass entertainment, network video can be converted between SD, HD, and Ultra HD images [8, 9] to satisfy the needs of the public with videos of different resolutions. In medicine, based on super-resolution reconstruction, existing medical images can be processed to obtain clearer results, which can help doctors to accurately understand the patient's situation, so that the patient can get timely and accurate treatment [10].

However, there are many degradation factors in the imaging process of images, such as the disturbance of atmospheric flow, the relative motion between the object and the imaging device, the inaccurate focusing of the imaging device, the change of the surrounding environment, and the own quality of the sensing instrument [11, 12]. These degradation factors lead to the interference of noise in the whole imaging process, accompanied by image distortion, distortion, etc. To recover the image and super-resolution reconstruction, it is necessary to analyze the basic principle of image degradation first. Due to the complexity of the image degradation factors, it is difficult to use a perfect mathematical model to accurately describe the degradation system, which is often approximated by a linear system model. The goal of super-resolution image reconstruction is to recover high-resolution images based on these observed images using some a priori knowledge and assumed conditions of the image degradation model, which is the inverse process about the imaging process of low spatial resolution images.

1.1. **Related Work.**

With the popularization of digital image acquisition equipment and the progress of technology, we can easily acquire a large number of image data. However, due to the limitation of hardware equipment or other factors, sometimes we can only get low-resolution images. This limits the details and clarity of the image and affects our accurate understanding and analysis of the image [13]. Therefore, super-resolution image reconstruction has become a hot research direction.

The significance of super-resolution image reconstruction lies in improving the visual quality and information richness of images. By converting a low-resolution image into a high-resolution image, we can get more details and clarity, making the image more realistic and true. For the classification of super-resolution problems, the related super-resolution reconstruction algorithms can be roughly divided into three major categories: interpolation-based methods, reconstruction-based methods, and deep learning-based methods.

(1) Interpolation-based approach. This method works by mapping all available low-resolution images to a reference image, where each low-resolution image provides some additional information about the imaged scene, fusing the available information from different low-resolution images on the basis of alignment, and finally de-blurring the image.

Since single image interpolation algorithms do not produce high frequency components that are lost during image acquisition, they do not deal well with the super-resolution problem, and the quality of the interpolated image obtained by such single image interpolation based methods is inherently limited by the information of the available data. There are many kinds of interpolation methods that can be used. One of the simplest algorithms is nearest neighbor interpolation, where each unknown pixel corresponds to the luminance value of the pixel closest to it, as well as bilinear interpolation and cubic interpolation algorithms, but these methods produce a very bad block effect. Haldar and Setsompop [14] proposed a non-uniform interpolation algorithm that uses the generalized multi-channel sampling theorem for a interpolation of a series of low-resolution images with spatial transfer. The advantage of this method is that it is computationally small and suitable for real-time applications. However, the optimality of the reconstruction cannot be guaranteed because the interpolation error of the reconstruction process is not taken into account. Liu et al. [15] proposed a polynomial approximation based on the motion minimum squared error (MLS) to estimate the luminance value at each pixel of a high-resolution image. Moreover, the coefficients and orders of this polynomial approximation are adaptively adjusted for each pixel location.

Interpolation-based reconstruction algorithms are simple and fast, and therefore suitable for real-time applications. However, single-frame image interpolation is not good enough to recover the detail information lost due to downsampling in the imaging process of low-resolution images, and thus the interpolation-based reconstructed images do not look real and believable.

(2) Reconstruction-based methods. This type of approach relies on observed low-resolution image sequences to reconstruct high-quality, high-resolution images according to a specific degradation model. The low-resolution image sequence contains different information about the same scene, and fusion of these different information about the scene should result in an image with a more complete description of the scene information. This method generally requires image alignment and sub-pixel accuracy. Liu et al. [16] proposed an iterative inverse projection (IBP) algorithm, which is based on the difference image between the observed low-resolution image and the simulated low-resolution image, and obtains the high-resolution image with the help of iterative projection, which is intuitive and easy to understand. However, this method does not obtain a unique solution, and it is difficult to select the parameters of the operators used in the inverse projection and to introduce a priori constraints.

(3) Deep learning based methods. This type of method obtains about the correspondence between high and low resolution images (or image blocks) through a large number of training samples. Then, the low-resolution input image (or image block) is reconstructed in high resolution based on this relationship. Deep learning based methods generally require the construction of sample libraries, and the size of the samples, and their sample diversity are key factors affecting the quality of the reconstruction. One of the challenges that the super-resolution reconstruction model needs to face is how to get a more reasonable description about the structural content of high-resolution images with the help of machine learning. In addition, some important application requirements are often accompanied by more complex image degradation processes, and the need for greater super-resolution reconstruction, such as super-resolution reconstruction of low-quality/long-distance surveillance videos. Zhao et al. [17] proposed an end-to-end deep convolutional neural network that can directly learn the representation of high-resolution images. The network contains convolutional, nonlinear and reconstruction layers, which can effectively recover the high-frequency details of an image. This method is a significant improvement over previous methods and demonstrates the power of deep learning in super-resolution tasks. Shi et al. [18] proposed a super-resolution generative adversarial network (SRGAN), which contains a generator and a discriminator. The generator is

used to reduce the high-resolution image and the discriminator is used to distinguish the super-resolution image from the real image. Through realistic adversarial training, the model can generate sharper and texture-rich results. SRGAN achieved the state-of-the-art performance at that time.SRGAN adds the results of the discriminator network to the loss function of the generator network, which makes the final trained high-resolution image closer to the natural image.

1.2. **Motivation and contribution.** The SRGAN network model, although it will obtain the most realistic visual perception, will reduce the PSNR value [19]. Therefore, the larger the magnification of the pixels of the image, the less smooth the image looks (lower value of PSNR), which results in the edges of the target object in the interior of the image looking blurred. Therefore, in order to solve the problems of blurred details and low contrast of SRGAN network models, this work proposes a super-resolution reconstruction method SA-DGAN based on Self-Attention (SA) mechanism and Deep Generative Adversarial Networks (DGAN).The main innovations and contributions of this work include:

(1) An improved generator network is proposed in order to fully utilize the rich sampling block to increase the sampling rate of SR features at different depths. low-level features in DGAN can potentially provide additional information to reconstruct high-frequency details in high-resolution (High Resolution, HR) images in order to learn low resolution (LR) and non-linear relationships in HR images. A dense connectivity structure is also used to integrate the features of each layer, which can fully utilize the multi-level features.

(2) A pixel self-attention module is designed for capturing feature dependencies in spatial and channel dimensions. The semantic dependencies in spatial and channel dimensions are modeled and the up-sampling and down-sampling modules of DGAN are redesigned to reduce the loss of image detail information.

(3) In order to effectively avoid the superposition caused by errors in DGAN, the extracted shallow information is divided into three scales by utilizing the idea of multi-scale, and the newly designed coding and decoding units are used to perform feature encoding-decoding reconstruction of the extracted shallow features at different scales.

2. Super-resolution reconstruction models and common structures.

2.1. **Image degradation and reconstruction models.** In the SR reconstruction process, the image degradation model must be studied and mastered in detail in order to realize an efficient reconstruction algorithm [20]. The degradation model mainly includes four processes: motion degradation (generated by the relative motion between the data acquisition device and the object), blurring, downsampling and noise, and the flow chart is shown in Figure 1.

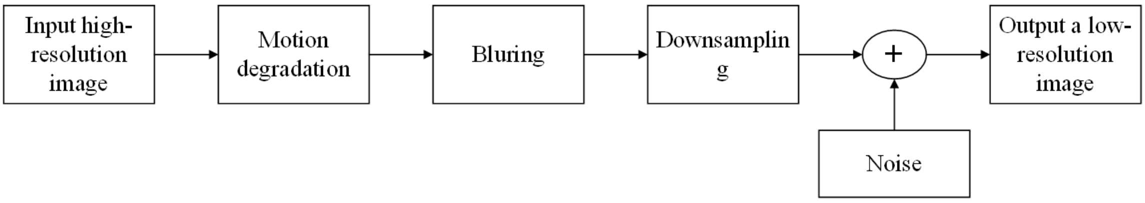


Figure 1. Image degradation process

The original HR image is $X_{L_1N_1\times L_2N_2}$, where L_1 and L_2 are the downsampling factors of the degradation process in horizontal and vertical directions, N_1 and N_2 are the dimensions of the LR image. The degradation process of the image can be represented as:

$$y_k = D_k H_k F_k X + V_k \tag{1}$$

where y_k denotes the kth degraded image, D_k is the downsampling operator corresponding to the image, H_k denotes the blurring factor consisting of a series of factors, F_k denotes the motion operator, X denotes the original HR image, V_k denotes the noise [21].

Image reconstruction techniques and image enhancement techniques are similar in that their main task is to recover the original low-resolution image, but image enhancement is subjective and evaluates the visual effect, while image reconstruction is objective and evaluates the image with objective evaluation indexes. The principle of image reconstruction is to try to recover the original image through the a priori information obtained in the degradation process, so the reconstruction technique is a reverse operation of the degradation model to recover the original image [22]. The super-resolution process makes the image recovery process irreversible due to the diversity of noise in the observation process, and secondly, the low-resolution dataset in the reconstruction process is generally scarce, thus making the recovery of image quality with multiple uncertainties. At the same time, the image degradation model is easily affected by external factors, which makes the reconstructed image become extremely unstable, and if the original image containing high-frequency noise is encountered, then the variation between the image pixel points will be very large, resulting in a lack of continuity of the image [23].

2.2. **Residual networks.** The main reason for using Residual Network in deep learning is to solve the problems of gradient disappearance and gradient explosion.

In deep neural networks, gradient disappearance and gradient explosion are two common problems. When information propagates in the network, with the increase of layers, the gradient will gradually become very small or very large, which makes the optimization process difficult and the network can not fully learn effective feature representation. The residual network solves this problem by introducing "jump connection". Jumping connection refers to introducing direct connection into the network and adding the input directly to the output of the network. In this way, the network can gradually adjust the output by learning the residual (the difference before and after skipping the connection) instead of using the traditional forward propagation. By jumping connection, the residual network can effectively transfer the gradient and avoid the disappearance or explosion of the gradient in the deep network, as shown in Figure 2.

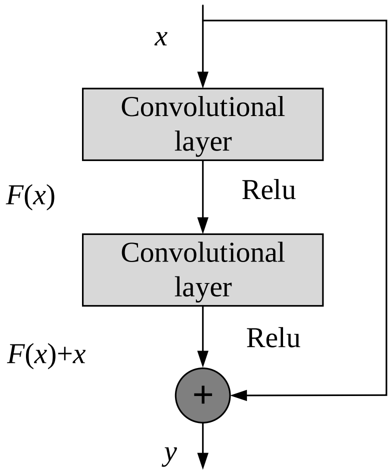


Figure 2. Structure of residual networks

Theoretically, the learning form of residual network is expressed as follow:

$$y = F(x) + x \tag{2}$$

where x and y are the inputs and outputs of the residual network, F represents the residual feature mapping learned during network training.

A residual network containing two convolutional layers is expressed as:

$$F(x) = \sigma(W_2\sigma(W_1 x)) \tag{3}$$

where σ is the activation function and W is the representation bias.

Since jump connections introduce neither additional parameters nor almost no computational complexity, yet they allow the network to converge more easily and solve the degradation problem associated with increasing depth [24]. The goal of SR image reconstruction is to improve the detail and clarity of the image. The residual network can reconstruct an image by learning the residual between a low-resolution image and its corresponding high-resolution image. This means that the network mainly pays attention to the differences of image details, and does not need to learn the representation of the whole image from scratch [25]. In this way, we can learn and reconstruct images more efficiently. Jumping connection in residual network allows fast information transmission path, which can help the network capture the nonlinear transformation of images better. For super-resolution image reconstruction, this means that the network can better model the high-frequency details and texture information in the image.

2.3. **Densely connected networks.** Conventional convolution network can only capture local information, while dense connection can learn the global feature representation of the image, thus improving the overall quality of the reconstructed image. Compared with the network with a large number of convolution layers stacked, the densely connected network has fewer parameters and is easy to train. Compared with the iterative optimization algorithm, the end-to-end dense connection network converges faster and the reconstruction results can be obtained quickly [26]. Generally speaking, the dense connection network can effectively extract the global information of the image, and is computationally efficient and easy to train and migrate, so it is very suitable for the task of super-resolution image reconstruction.

Densely connected neural networks can solve the problem of information vanishing. Unlike the summation operation of residual networks, the structure of dense connected networks is shown in Figure 3.

In SR reconstruction, cascading convolutional layers have different receptive fields from shallow to deep. This means that the convolutional layers at different locations are affected by pixel points in different size ranges in the LR image. When extracting features in different size ranges, localized features in small ranges are especially important for recovering texture details in the HR image, but often the feature maps in the forward position disappear as they flow through the whole network, resulting in the reconstructed HR image being too smooth. By introducing dense connectivity, the forward-positioned convolutional layer can be directly connected to the backward-positioned convolutional layer and the up-sampling operator, and the information in the forward-positioned feature maps can be directly applied to the process of feature extraction and up-sampling.

3. **Super-resolution reconstruction based on SA-DGAN.**

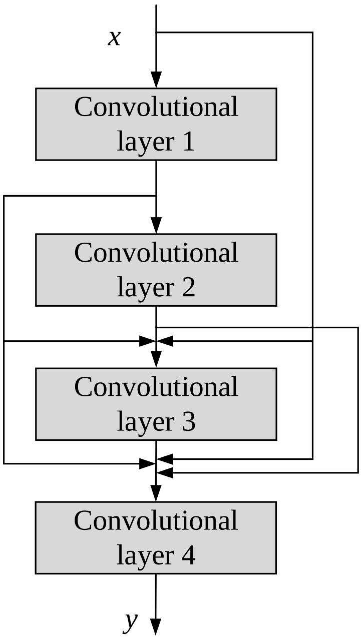


Figure 3. Structure of densely connected networks

3.1. **Improved generator network module.** Two consecutive OctConv [27] are used in DGAN for initial feature extraction. The kernel sizes are 3×3 and 1×1 for extracting low and high frequency feature maps, respectively. In the first OctConv layer, the original feature representation is converted to a multi-frequency feature representation. In the second OctConv layer, feature maps are constructed for the low-frequency inputs and high-frequency inputs, respectively, and the high-frequency feature maps will be used as inputs in the next stage to understand the nonlinear relationship between the LR and HR images. By skipping connections, the low-frequency feature maps are directly connected to the reconstruction module. The expression of the output feature map $Y = \{Y^H, Y^L\}$ is shown as follow:

$$Y^H = Y^{H \to H} + Y^{L \to H} = f(X^H; W^{H \to H}) + upsample(f(X^L; W^{L \to H}), 2) \tag{4}$$

$$Y^L = Y^{L \to L} + Y^{H \to L} = f(X^L; W^{L \to L}) + f(pool(X^H, 2); W^{H \to L}) \tag{5}$$

where X^H is the high-frequency feature detail and X^L is the low-frequency feature detail.

The structure of OctConv is shown in Figure 4. X and Y denote the input and output tensors, respectively. h and w denote the spatial dimensions, and the number of channels is denoted by c.

After the initial feature extraction, this work utilizes four improved sampling blocks to feed the proposed high-frequency features to the sampling, sample blocks, and connects all the results from the intermediate blocks to the next block in a densely connected manner. The sampling blocks in the module consist of up and down sampling operations [28]. In this work, six sets of up and down sampling operations are chosen to be constructed for the purpose of enriching the high-level representation and controlling the computational effort at the same time.

The improved dense sampling network is shown in Figure 5. The input of the s-th sampling block is $F_0, F_1, \ldots, F_{s-1}, F_s$, where F_0 is the high-frequency feature Y^H extracted

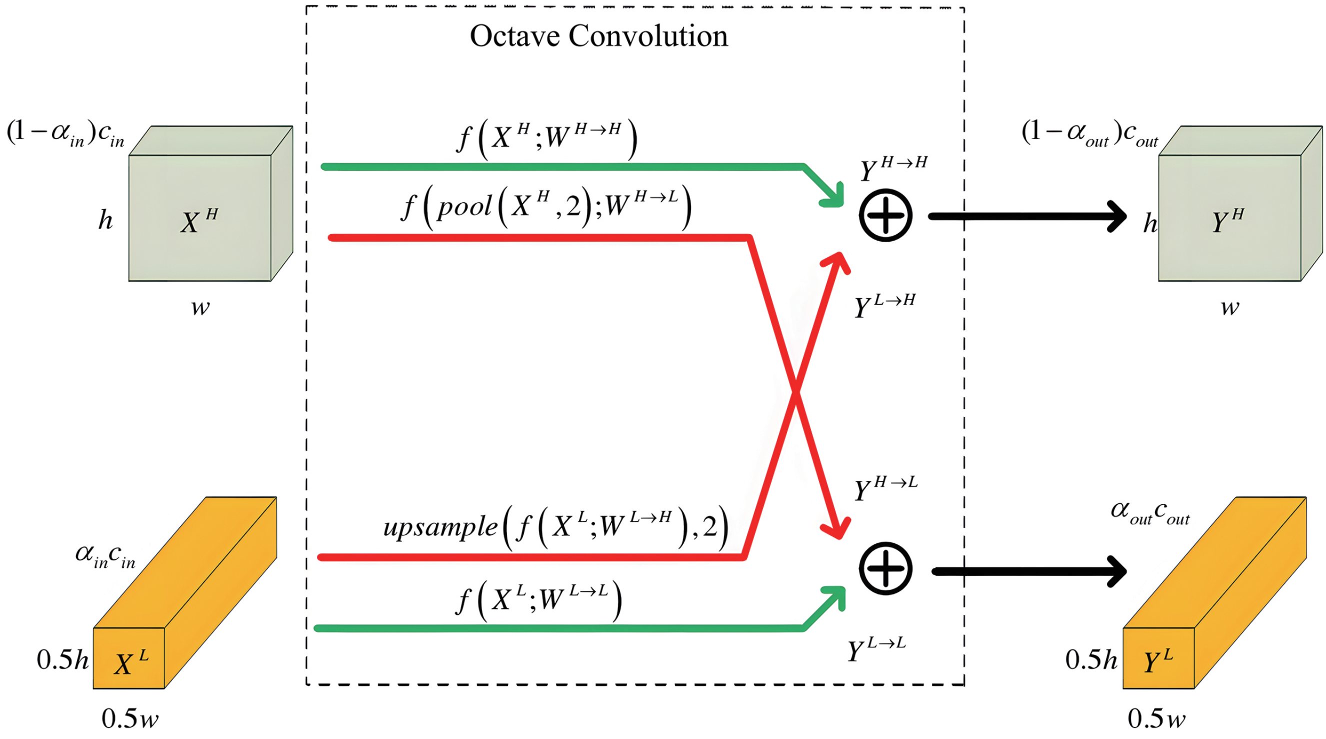


Figure 4. Structure of OctConv

from the previous module. First, the input image is concatenated and compressed by a 1×1 convolution.

$$L_0 = C_0\left([F_0, F_1, \ldots, F_s]\right) \tag{6}$$

where C_0 denotes the initial compression operation.

Then, let L_d denote the LR feature map obtained by the d-th upsampling group, and H_d denote the HR feature map obtained by the downsampling group.

$$L_d = C_d^{\downarrow}\left([H_1, H_2, \ldots, H_d]\right) \tag{7}$$

$$H_d = C_d^{\uparrow}\left([L_0, I_1, \ldots, L_d]\right) \tag{8}$$

where $C_d^{\downarrow}$ and $C_d^{\uparrow}$ denote the downsampling and upsampling used in the d-th sample block, respectively. The purpose of adding samples to the sample block is to reduce the number of parameters and increase computational efficiency.

Finally, the LR features generated from each upper and lower sampling block are fused using a 1×1 convolution to facilitate dense connectivity of useful information. F_s represents the output of the sampling block.

$$F_s = C_s\left([L_0, I_1, \ldots, L_d]\right) \tag{9}$$

Improved sampling blocks are tightly connected and used to create short paths between layers and other layers. All feature mappings from the previous layer are first connected to each other and then passed to all subsequent layers. Transferring elemental information to deeper layers through the deep network increases the information flow, introduces multi-stage feature extraction blocks for image super-resolution, and mitigates the problem of gradient vanishing. In addition, dense connectivity can significantly reduce the number of

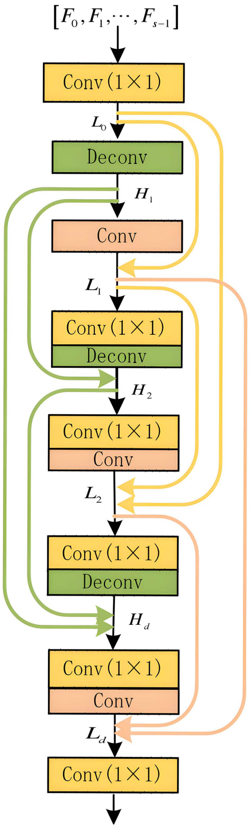


Figure 5. Improved dense sampling networks

parameters through feature reuse, resulting in a dramatic improvement in computational performance.

3.2. **Pixel SA module.** In this work, a pixel self-attention module consisting of a cascade of a channel attention [29] and a spatial attention [30] is added to the DGAN generator to fuse multilevel cross-modal features to enhance the compatibility of the network with the image features and to improve the extraction of high-frequency information from the image. The structure of the pixel SA module is shown in Figure 6.

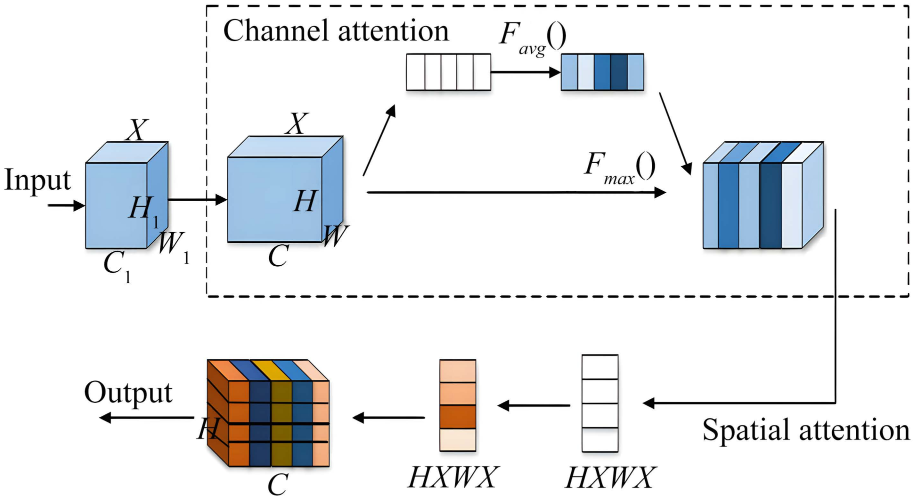


Figure 6. Pixel SA module

The number of input and output channels does not change in the network, and the feature map learns the degree of dependence on each channel by global averaging and global maximum pooling through channel attention, and subsequently improves the extraction of high-frequency image feature information while making corresponding spatial transformations of the information in the spatial domain to increase the diversity of feature information.

The input feature map is first denoted by F. The feature map is compressed in the spatial dimension. Complementarity in feature extraction is achieved by using average pooling and maximum pooling. A total of two one-dimensional vectors can be obtained after two pooling functions. Then the global average pooling is used for feature mean extraction in terms of feature maps, so that the network has feedback for each pixel point of the feature maps. The global maximum pooling is also a complement to the global average pooling.

The obtained weight feature map $M_C(F)$ after passing the channel attention is shown as follows:

$$M_C(F) = \sigma\left(R_1\left(R_0\left(F^c_{avg}\right)\right) + R_0\left(R_1\left(F^c_{max}\right)\right)\right) \tag{10}$$

where F^c_{avg} denotes the feature map after global maximum pooling computation, F^c_{max} denotes the feature map after global average pooling computation, R_0 and R_1 denote the parameters in the multilayer perceptual model, and σ denotes the activation function.

The feature maps form the attention model on the channel. In order to make the part of the input feature maps in the spatial dimension get higher weights accordingly, so compression operations are done on the input feature maps in the channel dimension by using average pooling and maximum pooling. The average and maximum sampling operations were done on the input features in the channel dimension to obtain two 2D feature maps respectively. Then a feature map with a channel number of 2 is obtained by stacking them together in the channel dimension. Finally, to ensure that the final feature map obtained agrees with the input feature map in the spatial dimension, a hidden layer containing a single 7 × 7 convolutional kernel is used. A Sigmoid activation function is used between this part of the convolutional layers.

3.3. **Multi-scale and cross-layer connection.** In order to effectively avoid the superposition caused by errors in DGAN, the extracted shallow information is divided into three scales by utilizing the idea of multi-scale, and the newly designed coding and decoding units are used to perform feature encoding-decoding reconstruction of the extracted shallow features at different scales

Considering that the coarse-scale information extracted from the image under the larger sensory field is often able to assist the image in the smaller sensory field of the slightly finer scale information extraction, the proposed network in this chapter combines the reconstruction information obtained from the decoding under the finer scale with the feature information obtained from the coding under the coarser scale, and jointly carries out the decoding and reconstruction operation of the current (i.e., the finer scale layer), which can better complete the reconstruction work. The output feature map of the coding and decoding unit at the coarse scale is shown as follow:

$$I_{de2}(x) = f_{de2}(f_{en2}(I_{en2}(x))) \tag{11}$$

where I_{en2} is the coarse-scale input feature map, f_{en2} is the encoding unit in the 1/8 branch, and f_{de2} is the decoding unit in the 1/8 branch.

$$I_{de1}(x) = f_{de1}(f_{en1}(I_{en1}(x)) + I_{de2}(x)) \tag{13}$$

where I_{en1} is the mesoscale input feature map, f_{en1} is the encoding unit in the 1/4 branch, and f_{de1} is the decoding unit in the 1/4 branch.

$$I_{de}(x) = f_{de}(f_{en}(I_{en}(x)) + I_{de1}(x)) \tag{14}$$

where I_{en} is the fine-scale input feature map, f_{en} is the encoding unit in the 1/2 branch, and f_{de} is the decoding unit in the 1/2 branch.

The proposed network fuses fine-scale feature information with slightly coarser size feature information layer by layer from bottom to top, while the features at the coarser scale also help to re-add some of the features at the finer scale that may have been lost step by step.

3.4. **Loss Function.**

DGAN needs to learn the mapping from X to Y. If the mapping relation is set to G, then the learned image is G_X. A discriminator is then used to determine whether it is a real image or not to form a generative adversarial network. The loss functions used include:

$$L_{adv} = \mathbb{E}_{X_{LR}}[logD(x_{HR})] + \mathbb{E}_{X_{HR}}[log(1 - D(G(x_{LR})))] \tag{15}$$

where x_{LR} denotes low resolution image, x_{HR} denotes high resolution image, G denotes generator and D denotes discriminator.

The adversarial loss makes the generated image closer to the distribution of the real high resolution image.

(2) Content loss.

$$L_{content} = \mathbb{E}_{X_{HR},X_{LR}}[\|x_{HR} - G(x_{LR})\|_1] \tag{16}$$

Content loss minimizes the difference between the generated image $G(x_{LR})$ and the real high-resolution image x_{HR}.

(3) Edge feature loss (Perceptual loss).

$$L_{perceptual} = \mathbb{E}_{X_{HR},X_{LR}}[\|\phi(x_{HR}) - \phi(G(x_{LR}))\|_1] \tag{17}$$

where ϕ denotes a pre-trained feature extraction network (e.g. VGG19). This preserves the edge and texture features of the generated image.

Therefore, the total loss function is shown as follows:

$$L_{total} = \lambda_{adv}L_{adv} + \lambda_{content}L_{content} + \lambda_{perceptual}L_{perceptual} \tag{18}$$

where λ is used to balance the weight of each loss. The generated high-resolution image can be obtained by minimizing this loss function.

4. **Experimental results and analysis.**

4.1. **Experimental environment and experimental dataset.**

The experimental hardware environment is: Intel Core i5 2.2GHz processor, 6G RAM, 400G hard disk, GTX1060 discrete graphics card. The experimental software environment is: Windows 7 operating system, Matlab 2012 (R2012a) simulation software.

An objective quantitative assessment is performed based on the peak signal-to-noise ratio (PSNR) and structural similarity metric (SSIM) between the reconstructed and original high-resolution images. Three image super-resolution datasets, BSDS100, SET14 and Urban100, were used for the experimental dataset, and the main parameters are shown in Table 1.

It can be seen that BSDS100 has the largest number of images with high resolution and rich scenes. SET14 has a small number of images but contains a variety of resolutions.

Table 1. Main parameters of the three datasets

Data set	**Number of pictures**	**Resolution range**	**Scene type**	**Hallmark**
Bsds100	100	321×481∼481×321	Portrait of an animal, nature Scenarios	High resolution
Set14	14	4 differentiation per picture Hasty	Generic scenario	Large resolution span
Urban100	100	4 differentiation per picture Hasty	Urban architecture, street view	Complexity of detail

urban100 focuses on urban scenes and has a complex method of details. These three datasets complement each other to comprehensively evaluate the recovery effect and generalization ability of super-resolution algorithms on different types of images. Researchers often use these datasets simultaneously to compare and analyze image super-resolution reconstruction algorithms.

4.2. **Objective experimental results.** In order to show the effectiveness of SA-DGAN network more objectively, this experiment selects a variety of existing SR algorithms as comparison methods, including Bicubic Interpolation, SRCNN, FSRCNN [31], VDSR [32] and DBPN [33]. The 2x and 4x scaling experiments are performed on the three datasets, and the experimental results are shown in Table 2 and Table 3.

Table 2. Results of 2x amplification experiments

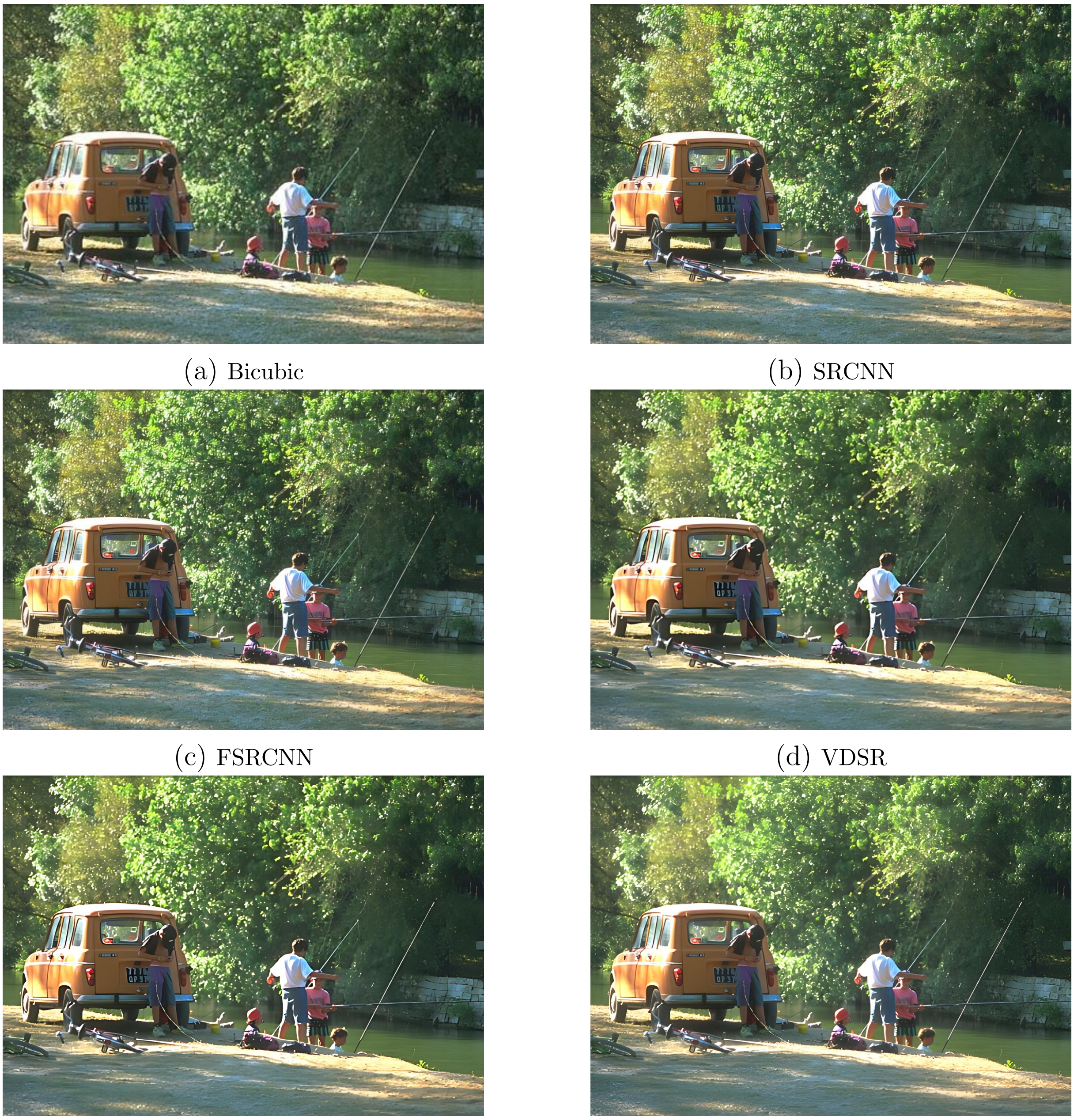
Model	**BSDS100 PSNR(dB)/SSIM**	**SET14 PSNR(dB)/SSIM**	**Urban100 PSNR(dB)/SSIM**
Bicubic	26.59/0.7728	26.03/0.7495	25.76/0.7258
SRCNN	27.42/0.8063	27.06/0.7793	26.90/0.7706
FSRCNN	27.55/0.8130	27.20/0.7801	26.97/0.7741
VDSR	27.84/0.8409	27.48/0.7989	27.19/0.7760
DBPN	28.69/0.8833	28.13/0.8268	27.62/0.7900
SA-DGAN	29.15/0.9024	28.64/0.8405	28.15/0.7922

Table 3. Results of 4x amplification experiments

Model	**BSDS100 PSNR(dB)/SSIM**	**SET14 PSNR(dB)/SSIM**	**Urban100 PSNR(dB)/SSIM**
Bicubic	24.68/0.7237	24.12/0.7019	23.65/0.6667
SRCNN	25.51/0.7574	25.15/0.7202	24.71/0.7115
FSRCNN	25.64/0.7640	25.39/0.7310	24.86/0.7150
VDSR	25.93/0.7918	25.57/0.7498	25.38/0.7179
DBPN	26.78/0.8342	26.22/0.7774	25.81/0.7310
SA-DGAN	27.24/0.8533	26.73/0.7914	25.84/0.7331

We averaged the PSNR and SSIM data of the test sample set to derive the final results for each evaluation metric. As can be seen from Tables 2 and 3, the super-resolution effect decreases with increasing magnification. The experimental results show that the experimental results of the deep learning-based methods are generally higher than those of the traditional methods, and the SA-DGAN network generally outperforms the other networks when tested on various types of datasets, and the SA-DGAN network also exhibits better results in various metrics.

4.3. **Subjective experimental results.** In order to visualize the reconstruction effect of various super-resolution methods, a randomly selected image sample with rich texture details from BSDS100 was used for the experiments in this work. The 4x zoom experiments were conducted using six super-resolution methods respectively, as shown in Figure 7.



(a) Bicubic (b) SRCNN

(c) FSRCNN (d) VDSR

(e) DBPN (f) SA-DGAN

Figure 7. Results of subjective experiments

In subjective visual comparison, the images generated by SA-DGAN are clearer in terms of texture, shadow, occlusion and pixel distribution.The images reconstructed by Bicubic, SRCNN, FSRCNN and VDSR networks are distorted to varying degrees.The images reconstructed by DBPN and SA-DGAN are almost free of distortion and have clearer outlines.The images reconstructed by SA-DGAN contain more detailed information and the visual effect is closer to the original HR images. DGAN reconstructed images contain more details and the visual effect is closer to the original HR image.

5. **Conclusions.** In this work, a super-resolution reconstruction method SA-DGAN is proposed.First, an improved generator network module is designed, which uses a dense connection structure to integrate the functions of each layer and can fully utilize the functions of multiple layers. Then, a pixel SA module is designed. The semantic dependencies on the spatial and channel dimensions are modeled, and the up-sampling and down-sampling modules are redesigned so as to reduce the loss of image detail information. Finally, the extracted shallow information is divided into three scales and feature encoding-decoding reconstruction is performed at different scales. Experimental results on three image super-resolution datasets, BSDS100, SET14 and Urban100, show that SA-DGAN has higher PSNR and SSIM compared to multiple existing super-resolution reconstruction algorithms.However, the self-attention is limited in its ability to model remote dependencies, and it cannot integrate global information effectively. Therefore, subsequent studies will try to use hierarchical self-attention structure to gradually expand the sensory field through multiple iterations of inference.

Acknowledgment. This work is partially supported by the Science and Technology Research Project of Chongqing Education Commission (No. KJQN202203213 ,No. KJQN202403214, No.KJZD-K202403206), Technology Innovation and Application Development Shandong-Chongqing Science and Technology Cooperation Special Project of Chongqing Science and Technology Bureau(No.CSTB2024TIAD-LDX0019).

REFERENCES

[1] H. Chen, X. He, L. Qing, Y. Wu, C. Ren, R. E. Sheriff, and C. Zhu, "Real-world single image super-resolution: A brief review," *Information Fusion*, vol. 79, pp. 124-145, 2022.

[2] C. Saharia, J. Ho, W. Chan, T. Salimans, D. J. Fleet, and M. Norouzi, "Image super-resolution via iterative refinement," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 4, pp. 4713-4726, 2022.

[3] S. Liu, P. Hoess, and J. Ries, "Super-resolution microscopy for structural cell biology," *Annual Review of Biophysics*, vol. 51, pp. 301-326, 2022.

[4] H. Li, Y. Yang, M. Chang, S. Chen, H. Feng, Z. Xu, Q. Li, and Y. Chen, "Srdiff: Single image super-resolution with diffusion probabilistic models," *Neurocomputing*, vol. 479, pp. 47-59, 2022.

[5] C. Tian, Y. Yuan, S. Zhang, C.-W. Lin, W. Zuo, and D. Zhang, "Image super-resolution with an enhanced group convolutional neural network," *Neural Networks*, vol. 153, pp. 373-385, 2022.

[6] Y. Chen, R. Xia, K. Yang, and K. Zou, "MFFN: Image super-resolution via multi-level features fusion network," *The Visual Computer*, pp. 1-16, 2023.

[7] W. Zhao, S. Zhao, L. Li, X. Huang, S. Xing, Y. Zhang, G. Qiu, Z. Han, Y. Shang, and D.-e. Sun, "Sparse deconvolution improves the resolution of live-cell super-resolution fluorescence microscopy," *Nature Biotechnology*, vol. 40, no. 4, pp. 606-617, 2022.

[8] Y. Wang, L. Wang, G. Wu, J. Yang, W. An, J. Yu, and Y. Guo, "Disentangling light fields for super-resolution and disparity estimation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 1, pp. 425-443, 2022.

[9] Y. Huang, S. Li, L. Wang, and T. Tan, "Unfolding the alternating optimization for blind super resolution," *Advances in Neural Information Processing Systems*, vol. 33, pp. 5632-5643, 2020.

[10] T.-Y. Wu, X. Fan, K.-H. Wang, C.-F. Lai, N. Xiong, and J. M.-T. Wu, "A DNA Computation-Based Image Encryption Scheme for Cloud CCTV Systems," *IEEE Access*, vol. 7, pp. 181434-181443, 2019.

[11] T.-Y. Wu, X.-N. Fan, K.-H. Wang, J.-S. Pan, C.-M. Chen, and J. M.-T. Wu, "Security Analysis of a Public Key Authenticated Encryption with Keyword Search Scheme," *Journal of Information Hiding and Multimedia Signal Processing*, vol. 9, no. 4, pp. 1050-1057, 2018.

[12] K. Wang, X. Zhang, F. Wang, T.-Y. Wu, and C.-M. Chen, "Multilayer Dense Attention Model for Image Caption," *IEEE Access*, vol. 7, pp. 66358-66368, 2019.

[13] T.-Y. Wu, X.-N. Fan, K.-H. Wang, J.-S. Pan, and C.-M. Chen, "Security analysis and improvement on an image encryption algorithm using Chebyshev generator," *Journal of Internet Technology*, vol. 20, no. 1, pp. 13-23, 2019.

[14] J. P. Haldar, and K. Setsompop, "Linear Predictability in MRI Reconstruction: Leveraging Shift-Invariant Fourier Structure for Faster and Better Imaging," *IEEE Signal Processing Magazine*, vol. 37, no. 1, pp. 69-82, 2020.

[15] Z. Liu, S. Yang, Z. Feng, Q. Gao, and M. Wang, "Fast SAR autofocus based on ensemble convolutional extreme learning machine," *Remote Sensing*, vol. 13, no. 14, 2683, 2021.

[16] H. Liu, Y. Lin, B. Ibragimov, and C. Zhang, "Low dose 4D-CT super-resolution reconstruction via inter-plane motion estimation based on optical flow," *Biomedical Signal Processing and Control*, vol. 62, 102085, 2020.

[17] W. Zhao, C. Persello, and A. Stein, "Building outline delineation: From aerial images to polygons with an improved end-to-end learning framework," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 175, pp. 119-131, 2021.

[18] J. Shi, Y. Ye, H. Liu, D. Zhu, L. Su, Y. Chen, Y. Huang, and J. Huang, "Super-resolution reconstruction of pneumocystis carinii pneumonia images based on generative confrontation network," *Computer Methods and Programs in Biomedicine*, vol. 215, 106578, 2022.

[19] Y. Xiong, S. Guo, J. Chen, X. Deng, L. Sun, X. Zheng, and W. Xu, "Improved SRGAN for remote sensing image super-resolution across locations and sensors," *Remote Sensing*, vol. 12, no. 8, 1263, 2020.

[20] K. Li, S. Yang, R. Dong, X. Wang, and J. Huang, "Survey of single image super-resolution reconstruction," *IET Image Processing*, vol. 14, no. 11, pp. 2273-2290, 2020.

[21] Y. Chen, L. Liu, V. Phonevilay, K. Gu, R. Xia, J. Xie, Q. Zhang, and K. Yang, "Image super-resolution reconstruction based on feature map attention mechanism," *Applied Intelligence*, vol. 51, pp. 4367-4380, 2021.

[22] H. Kim, J. Kim, S. Won, and C. Lee, "Unsupervised deep learning for super-resolution reconstruction of turbulence," *Journal of Fluid Mechanics*, vol. 910, 29, 2021.

[23] Y. Zou, L. Zhang, C. Liu, B. Wang, Y. Hu, and Q. Chen, "Super-resolution reconstruction of infrared images based on a convolutional neural network with skip connections," *Optics and Lasers in Engineering*, vol. 146, 106717, 2021.

[24] D. Qiu, S. Zhang, Y. Liu, J. Zhu, and L. Zheng, "Super-resolution reconstruction of knee magnetic resonance imaging based on deep learning," *Computer Methods and Programs in Biomedicine*, vol. 187, 105059, 2020.

[25] J. Du, Z. He, L. Wang, A. Gholipour, Z. Zhou, D. Chen, and Y. Jia, "Super-resolution reconstruction of single anisotropic 3D MR images using residual convolutional neural network," *Neurocomputing*, vol. 392, pp. 209-220, 2020.

[26] Y. Jin, Y. Zhang, Y. Cen, Y. Li, V. Mladenovic, and V. Voronin, "Pedestrian detection with super-resolution reconstruction for low-quality image," *Pattern Recognition*, vol. 115, 107846, 2021.

[27] L. Guo, and M. Woźniak, "An image super-resolution reconstruction method with single frame character based on wavelet neural network in internet of things," *Mobile Networks and Applications*, vol. 26, pp. 390-403, 2021.

[28] N. Q. Truong, P. H. Nguyen, S. H. Nam, and K. R. Park, "Deep learning-based super-resolution reconstruction and marker detection for drone landing," *IEEE Access*, vol. 7, pp. 61639-61655, 2019.

[29] W. Wang, F. Wei, L. Dong, H. Bao, N. Yang, and M. Zhou, "Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers," *Advances in Neural Information Processing Systems*, vol. 33, pp. 5776-5788, 2020.

[30] M. Hahn, "Theoretical limitations of self-attention in neural sequence models," *Transactions of the Association for Computational Linguistics*, vol. 8, pp. 156-171, 2020.

[31] Z. Chen, J. Liu, J. Yang, and W. Yang, "Super-resolution network-based fractional-pixel motion compensation," *Signal, Image and Video Processing*, vol. 15, no. 7, pp. 1547-1554, 2021.

[32] J. Cong, X. Wang, X. Lan, M. Huang, and L. Wan, "Fast target localization method for FMCW MIMO radar via VDSR neural network," *Remote Sensing*, vol. 13, no. 10, 1956, 2021.

[33] Y. Yu, X. Li, and F. Liu, "E-DBPN: Enhanced deep back-projection networks for remote sensing scene image super-resolution," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 58, no. 8, pp. 5503-5515, 2020.