Hindawi
Computational Intelligence and Neuroscience
Volume 2021, Article ID 2464648, 16 pages
https://doi.org/10.1155/2021/2464648

Research Article

A Generative Adversarial Network Fused with Dual-Attention Mechanism and Its Application in Multitarget Image Fine Segmentation

Jian Yin,[1] **Zhibo Zhou**,[2] **Shaohua Xu**,[1] **Ruiping Yang**,[1] **and Kun Liu**[1]

[1]*College of Computer Science and Engineering, Shandong University of Science and Technology, Qingdao 266 590, China*
[2]*Qingdao Ruisi Intelligent Technology Co., Ltd., Qingdao 266 590, China*

Correspondence should be addressed to Shaohua Xu; xush62@163.com

Received 31 August 2021; Accepted 20 November 2021; Published 18 December 2021

Academic Editor: Yugen Yi

Aiming at the problem of insignificant target morphological features, inaccurate detection and unclear boundary of small-target regions, and multitarget boundary overlap in multitarget complex image segmentation, combining the image segmentation mechanism of generative adversarial network with the feature enhancement method of nonlocal attention, a generative adversarial network fused with attention mechanism (AM-GAN) is proposed. The generative network in the model is composed of residual network and nonlocal attention module, which use the feature extraction and multiscale fusion mechanism of residual network, as well as feature enhancement and global information fusion ability of nonlocal spatial-channel dual attention to enhance the target features in the detection area and improve the continuity and clarity of the segmentation boundary. The adversarial network is composed of fully convolutional networks, which penalizes the loss of information in small-target regions by judging the authenticity of prediction and label segmentation and improves the detection ability of the generative adversarial model for small targets and the accuracy of multitarget segmentation. AM-GAN can use the GAN's inherent mechanism that reconstruct and repair high-resolution image, as well as the ability of nonlocal attention global receptive field to strengthen detail features, automatically learn to focus on target structures of different shapes and sizes, highlight salient features useful for specific tasks, reduce the loss of image detail features, improve the accuracy of small-target detection, and optimize the segmentation boundary of multitargets. Taking medical MRI abdominal image segmentation as a verification experiment, multitargets such as liver, left/right kidney, and spleen are selected for segmentation and abnormal tissue detection. In the case of small and unbalanced sample datasets, the class pixels' accuracy reaches 87.37%, the intersection over union is 92.42%, and the average Dice coefficient is 93%. Compared with other methods in the experiment, the segmentation precision and accuracy are greatly improved. It shows that the proposed method has good applicability for solving typical multitarget image segmentation problems such as small-target feature detection, boundary overlap, and offset deformation.

1. Introduction

Image multitarget segmentation is one of the hotspots in the field of image processing and artificial intelligence. Essentially, it can be expressed as pixel classification of multitarget with semantic labels, that is, segmenting and describing multitarget of interest in the image by using a set of object categories to classify and mark images at the pixel level [1, 2]. With the continuous development of image analysis theory and deep learning technology, researchers have proposed many effective multitarget image segmentation models and algorithms [3–5]. However, in some complex scenes, the image is affected by noise, offset deformation, gray value distortion, local position effect, and other factors, so the existing methods still have problems such as target omission, position offset, unclear boundary, and so on [6, 7]. At the same time, some image segmentation methods have good results for single-target segmentation, but still not applicable to the complex multitarget image segmentation. Especially, in multi-instance, the accurate detection and segmentation

accuracy of small-size targets are difficult to guarantee [8, 9]. There are still challenges in the research of image multitarget detection and segmentation.

At present, image segmentation can be divided into two categories: traditional methods and deep learning methods. Traditional image segmentation algorithms mostly use gray-scale features to segment images [10]. Typical methods include threshold-based method [11], edge-based method [12], and region-based method [13]. However, there are problems such as susceptibility to the contrast of image gray features, poor segmentation of low-resolution and blurred images, and prone to oversegmentation of images. In general, traditional image segmentation methods are greatly affected by subjective factors, and the preprocessing process is complicated. For example, strong prior knowledge is needed to solve the problems such as the selection of seed points during segmentation and the selection of the threshold of the segmentation boundary. At the same time, small errors in the selection of key parameters have a greater impact on segmentation accuracy. These problems make traditional image segmentation algorithms still have many restrictions when applied to image semantic segmentation.

Deep convolutional network is an effective image feature extraction and analysis method, which has been widely used in image classification, image generative, and target detection. Many excellent algorithms have emerged, such as AlexNet [14], ResNet [15], Faster R-CNN [16], and so on. Although deep convolutional networks have powerful feature extraction capabilities, they still have many limitations when applied to image segmentation problems. Take multitarget segmentation of medical images as an example. Medical images intuitively reflect the 2D and 3D morphological characteristics of organs and tissues in specific areas of the human body. The human body has a relatively complex organ and tissue structure, and the anatomical structure of different human bodies is also different among individuals. At the same time, it is easy to be affected by factors such as noise, illumination, and local posture effect, and the image shape and various organ tissue regions are soft boundaries, showing regular or irregular dynamic periodicity with cardiac contraction or relaxation. These factors increase the difficulty of medical image feature differentiation, target detection, and segmentation [17, 18]. Moreover, the pooling layer in the convolutional neural network (CNN) will downsample the input image size and reduce the resolution of the image; the fully connected layer will turn the image features into vectors, destroying the spatial information of the image. Therefore, it still has inadequacy in solving the problem of complex image segmentation. Although these convolution-based methods have achieved certain segmentation results, they ignore the spatial correlation of medical images such as CT and MRI, which is easy to produce nonsmooth and discontinuous segmentation results [19].

In order to solve the limitations of the structure and information processing mechanism of traditional convolutional neural networks in image segmentation problems, Long et al. [20] proposed the fully convolutional network (FCN) model. FCN abandons the fully connected layer of CNN and replaces it with the fully convolution structure. At the same time, it uses the deconvolution operator to restore the image size and introduces a shortcut-connection structure to fuse the high-level characteristics of the network with the low-level features to optimize the segmentation results. At present, most of the deep learning models used in image segmentation are based on the idea of the FCN model. Ronneberger et al. [21] proposed the U-Net model, which retains the convolution and deconvolution structure of the FCN model, but changes the way of fusion of high-level feature map and low-level feature map. The encoding part and the decoding part of U-Net are completely symmetrical, and the connection is realized through channel splicing and then convolution. At present, there are many variants of U-Net models, such as 3D U-Net [22], Res U-Net [23], Dens U-Net [24], and Attention U-Net [25]. In addition, the SegNet model proposed by Badrinarayanan et al. [26] replaces the deconvolution of FCN with an up-pooling method, which makes the upsampling part no longer participate in training while ensuring the segmentation accuracy, reducing the computational complexity. The above methods and mechanisms effectively solve the principle and strategy problems of image segmentation, but in complex image segmentation with noise and content diversity, there are still problems of unstable segmentation effect on low-resolution and fuzzy images and low accuracy of target pixel classification [27].

In the research of low-resolution image high-definition processing, Zhang et al. [28] (2021) build a hierarchical correlation filters model based on the multilevel convolutional features, which can suppress interference of background and similar objects. Chen et al. [29] (2021) proposed an image super-resolution reconstruction method using attention mechanism with the feature map. It uses the information extraction block of feature map attention mechanism to adaptively adjust the channel characteristics, enhance the feature expression ability, and facilitate reconstruction from original low-resolution images to multiscale super-resolution images. For the image inpainting, Chen et al. [30] (2021) proposed a novel image embedding algorithm based on encoder and similarity constraint, which effectively solved the problem of joint context awareness loss in image inpainting and improved the utilization of features. The above works provide good support for fine segmentation of complex images. However, the proposed method will still be affected by obvious blur, and the training time required will also be longer. Then, Chen et al. [31] (2021) proposed image completion algorithm based on the improved total variation minimization method, which can solve the issue mismatching and structure disconnecting in exemplar-based image inpainting.

The generative adversarial network (GAN) proposed by Goodfellow et al. (2014) [32] is a deep learning method based on Nash equilibrium in game theory, including two parts: generator and discriminator. The generator in the GAN model, that is, the generator network model, is mainly used to generate target data. The typical generator structure is a neural network based on deconvolution, which restores the input image size through multiple deconvolution layers'

upsampling and finally obtains generated image data. In image segmentation, the generator is essentially a segmentation model. For example, FCN [20], U-Net [21], and SegNet [26] can be selected as the generative network, which receives the input of the original image and takes the predictive segmentation as the output. The discriminator, that is, the adversarial network in GAN, usually uses CNNs as the basic model. Its inputs include the real data and the generator's generated data. The generated data are judged as false, and the real data are judged as true. Through the authenticity judgment, the game learning is performed to optimize the generator's ability to generate data. In mechanism, GAN can reconstruct low-resolution images into super-resolution high-definition images [33] and train the generative model based on the surrounding pixels of the missing part of the image to repair the complete image [34]. If applied to the image segmentation problems in the case of blur, offset, and small target, it can effectively improve the quality and accuracy of image segmentation. Attention mechanism is a target feature enhancement method that is widely studied and applied at present. It can be used as a module of the deep learning model to focus attention on objects of interest [35, 36]. However, most of the existing methods mainly focus on the local pixels of the target area and have low relevance to the image content with a large receptive field [37]. In order to capture the dependence of spatial long-distance information in the image, Wang et al. [38] proposed a nonlocal attention mechanism, the strategy of which is that the characteristic response value at a pixel is equal to the weighted average of the characteristic values at all receptive field points, that is, all points in the larger receptive field are connected to realize global information fusion. The nonlocal attention mechanism connects the understanding of global content with the semantics of local targets, which improves the enlightenment and restriction of target pixel classification. In complex image multiobject segmentation, if the image segmentation mechanism of GAN is combined with nonlocal attention feature enhancement method, it can effectively improve the accuracy of complex image multiobject segmentation and optimize segmentation boundary in mechanism.

Aiming at the poor accuracy of multitarget instance segmentation and small-scale target segmentation in complex images, a generative adversarial network fused with nonlocal attention mechanism is proposed in this paper. The generative network module of the GAN uses the residual network as the basic model for preliminary target segmentation. The nonlocal spatial-channel dual-attention mechanism is added to the output feature map of the residual network to capture the long-distance dependence information of each feature point on the output feature map. The adversarial network module is constructed based on CNNs, which performs masking operations on the original image with prediction segmentation and label segmentation, respectively, and inputs the masking result into the adversarial network. The adversarial network judges the mask result of the predicted segmentation as false, and the mask result of the label segmentation is judged as true. And the generative network judges the mask result of the predicted segmentation as true. Through the game learning between the generative network and the adversarial network, the image segmentation ability of the generative network is optimized. Specifically, firstly, a generative network module is constructed based on residual network and nonlocal attention mechanism for preliminary image segmentation. On this basis, masking operations on the original image with the prediction and label segmentation, respectively, are performed, and the results are input to the adversarial network to optimize the segmentation results. AM-GAN strengthens the extraction and fusion of multiscale features, as well as the distinguishing ability of each instance boundary pixels, to achieve the fine segmentation of multitarget instance regions in complex image and improve the accuracy of small-target segmentation.

Medical images are often blurred due to the influence of the imaging environment and detecting equipment. Meanwhile, the complexity of human organ and tissue structure, the differences of different individual anatomical structures, and the influence of local posture effect also increase the difficulty of medical image feature differentiation and segmentation. In this paper, taking the segmentation of abdominal MRI images as a verification experiment, multitarget tissues such as liver, left/right kidney, and spleen are selected for segmentation and abnormal detection to verify the effectiveness of the model and algorithm.

In this paper, the problems in image segmentation, such as insignificant morphological characteristics of the target image, easy to be affected by noise, gray value distortion and local position effect, and unclear boundary of the target region, are studied. The main motivation is to establish a novel image segmentation model, which can improve the accuracy of complex image fine segmentation in mechanism.

The novelty and main contributions of this paper are as follows:

(1) A novel generative adversarial network fused with the attention mechanism (AM-GAN) multitarget image segmentation model is proposed. In mechanism, the image segmentation mechanism of GAN is organically combined with the feature enhancement method of nonlocal attention so that the model and algorithm can automatically learn to focus on the target structure with different shapes and sizes, highlighting the feature usefulness for specific tasks. It can effectively improve the problems of existing image segmentation methods, such as insufficient utilization of correlation information between image voxels, imprecise detection of small-target area, unclear boundary, and overlapping multitarget boundary, and effectively improve the accuracy of complex image multitarget segmentation.

(2) The generative module of AM-GAN combines the feature enhancement of nonlocal attention with the feature fusion method of the residual network. In the generative network, the understanding of the global content is linked with the semantics of the local target, and the low-level and high-level feature maps

are added through the shortcut-connection structure to achieve feature fusion. In mechanism, it can give play to the guidance and heuristics of target segmentation and refine the segmentation results.

(3) In this paper, AM-GAN is proposed as the image multitarget segmentation model. In mechanism, it can use GAN's high-definition processing and repair capabilities to reduce the effects of noise, bias deformation, and gray value distortion. Through the information association and restriction of the nonlocal attention large receptive field to the local target, the continuity of target segmentation is maintained. Meanwhile, the Nash game strategy between generative network and adversarial network is adopted in AM-GAN algorithm to optimize the segmentation results and improve the continuity, smoothness, and accuracy of multitarget segmentation results.

In Section 1, the current challenges and current research status of the complex image multitarget segmentation are reviewed and analyzed. The ideas and algorithm strategies of a novel AM-GAN segmentation model established in this paper are pointed out. In Section 2, the AM-GAN model is established and its theoretical properties are analyzed. In Section 3, the comprehensive learning algorithm of AM-GAN is designed and proposed. In Section 4, multitarget segmentation experiments and result analysis are conducted based on medical images. Finally, the work of the paper is summarized, and the advantages and limitations are pointed out.

2. The Generative Adversarial Network Fused with the Dual-Attention Mechanism Segmentation Model

2.1. The Generative Adversarial Network Basic Model. The generative adversarial network (GAN) basic model in this paper includes two parts: generator and discriminator. The basic structure and information processing flow are shown in Figure 1.

In Figure 1, the generative network module in GAN is mainly used to generate target data. The typical generative network generally is usually the neural network based on deconvolution, such as FCN, U-Net, and SegNet, which recovers to the size of the input image through sampling on multiple deconvolution layers and finally obtains the generated image data. For image segmentation, the generative network is used as an image segmentation model, which receives the input of the original image and takes the predictive segmentation as the output. The adversarial network in GAN usually adopts convolutional neural networks. Its input includes real data and generated data from the generator. Through authenticity judgment and game learning with the generator, the ability of the generator to generate data is optimized.

The gray value of the input image to generative network is recorded as the random variable x, the distribution it obeys is set to $P(x)$, and the noise data z obeys the uniform distribution. The generator maps the noise data z to $G(z)$, and the authenticity discrimination probability of the image random variable x is $D(x)$.

In learning, GAN is optimized according to the principle that the generative model maximizes $\log D(x)$ and the discrimination model minimizes $\log(1 - D(G(z)))$. The objective function is defined as follows:

$$\min_G \max_D V(D,G) = E_{x\sim P(x)}[\log D(X)] + E_{z\sim P(x)}[\log(1 - D(G(Z)))]. \tag{1}$$

2.2. Nonlocal Attention Mechanism. The core idea of nonlocal attention mechanism is that the characteristic response value at a pixel is equal to the weighted average of the characteristic values at all points, that is, all points in the receptive field are connected, and the dependency relationship between pixels based on spatial distance is established to realize global information fusion. The calculation formula [38] is

$$y_i = \frac{1}{C(x)} \sum_{\forall j} f\left(x_i, x_j\right) g\left(x_j\right), \tag{2}$$

where i represents the output position index, j represents any position of the input, x represents the input, y represents the output with the same size as x, $f(\cdot)$ is the similarity measurement function between pixels, $g(\cdot)$ represents the feature mapping of j, and $C(x)$ is the normalization factor. (f) and g can be implemented in many ways. If g is a Gaussian function, its expression is

$$f\left(x_i, x_j\right) = e^{x_i^T x_j}, \tag{3}$$

$$C(x) = \sum_{\forall j} f\left(x_i, x_j\right). \tag{4}$$

The operation in equation (3) represents the difference amplified exponentially after multiplying the two vector matrices, and equation (4) is the mathematical expression of the normalization function $C(x)$.

Consider an extended form of equation (3). Firstly, the vector x is embedded into spatial mapping, that is, the two vectors are mapped to different feature spaces, and then, the Gaussian function is used to measure the similarity. The specific form is

$$f\left(x_i, x_j\right) = e^{\theta(x_i)^T \varphi\left(x_j\right)}. \tag{5}$$

The vector dot product of function f can be expressed as

$$f\left(x_i, x_j\right) = \theta(x_i)^T \varphi\left(x_j\right). \tag{6}$$

Now, the normalization function $C(x)$ is equal to N, and N is the number of positions of x.

Based on the paired function form in the relational network proposed by Santoro et al. [39], the function f can be expressed as a cascaded form:

$$f\left(x_i, x_j\right) = Relu\left(w_f^T\left[\theta(x_i), \varphi\left(x_j\right)\right]\right), \tag{7}$$

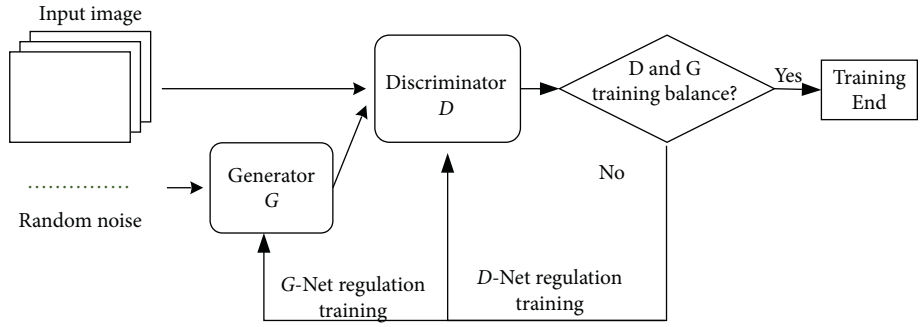

Figure 1: The generative adversarial network.

where [,] represents cascade and w_f^T represents the weight vector that projects the cascade vector to the scalar.

The nonlocal attention mechanism is integrated into convolutional neural network to form a general nonlocal block. The module can be integrated into any neural network structure. The definition of nonlocal block is as follows:

$$z = w_z y + x, \tag{8}$$

where z is the output of the attention module, y is the output of equation (2), w_z is the weight matrix of the number of restored input channels, and x is the input.

The structure of nonlocal block is shown in Figure 2.

In Figure 2, the nonlocal block first performs feature mapping on the input matrix $I(C \times H \times W)$ with a 1×1 convolution, that is, the mapping operation of the functions g, θ, and φ in (3) and (5), to get matrices w_r, w_k, and w_v. Then, the matrix w_k is deformed and transposed into a matrix $F_k(HW \times T)$, and w_r is transformed into a matrix $F_r(T \times HW)$. F_k is multiplied by F_r and then normalized by Softmax to obtain the similarity measure matrix $F(HW \times HW)$. Then, multiply the deformed matrix $F_v(T \times HW)$ from w_v and the matrix F to obtain the characteristic response matrix $F_s(T \times HW)$. Finally, F_s is deformed and multiplied by the convolution kernel w_z to restore the original number of channels. The operation result is added to the input matrix $I(C \times H \times W)$ to obtain the output $O(C \times H \times W)$, and the final output feature map is a feature map with enhanced global information dependence.

2.3. Residual Generation Network Based on Dual-Attention Mechanism. In the complex images' multitarget segmentation, if the number of image instance targets is large and the difference of individual gray value is small, the neural network needs to have strong ability of feature extraction, fusion, and recognition. In this paper, the residual network is used as the main body of the image segmentation model, and the shortcut-connection structure is added between different convolution layers, that is, the input of the upper network is directly superimposed with the output of the lower network at the element level, so as to reduce the loss of feature information and realize feature fusion of different levels. In this way, the network can still maintain good convergence properties in the deep case. The residual structure is shown in Figure 3.

At present, there are many classical residual network models, such as ResNet18, ResNet34, and ResNet50 [16]. In practice, the selection of the model is mainly based on the requirements of input image size, quality, and segmentation accuracy. Generally, the deeper the network, the stronger the feature extraction ability. In this paper, ResNet50 is selected as the basic model according to the problem of small-target region detection in complex image multitarget segmentation. The structure of segmentation model based on ResNet50 is shown in Figure 4.

In Figure 4, the segmentation model is divided into six parts. (1) 7×7 convolution with step size of 2 and the number of output image channels is 64. (2) The pooling operation with step size of 2 and the 3×3 sliding window cascades three residual blocks. Each residual block contains three convolution layers. The first and third are 1×1 convolution to adjust the number of channels of the feature maps. The second is 3×3 convolution with a step size of 1. Equations (3) to (5) are stacked residual blocks. The last one includes average pooling with a step size of 1 and 7×7 sliding window, as well as fully connection and Softmax classification. Extract the output from the second to the fifth part of the model, and then, upsample the four output feature maps for prediction segmentation.

In practice, only relying on the residual network to perform image multitarget segmentation will result in the blurring of the segmentation boundary and the loss of small-target information, that is, the residual network cannot make full use of the image feature information to segment the image. In this paper, based on the residual network, a nonlocal attention mechanism is introduced to construct a dual-attention mechanism model that can integrate spatial and channel attention. The structure is shown in Figure 5.

In Figure 5, the spatial-attention module is used on a nonlocal mechanism to strengthen the dependency between all pixels on the feature maps, which is represented by a similar weight matrix. In this module, the output feature map of the residual network first undergoes 1×1 convolution to obtain three feature maps $F1$, $F2$, and $F3$, which have the same width and height as the input image, but the number of channels is reduced to 1/4 of the original. On this basis, $F1$ is transposed and multiplied by $F2$ and then processed by Softmax normalization to obtain the interpixel similarity matrix $F(HW \times HW)$. F is multiplied by $F3$, and

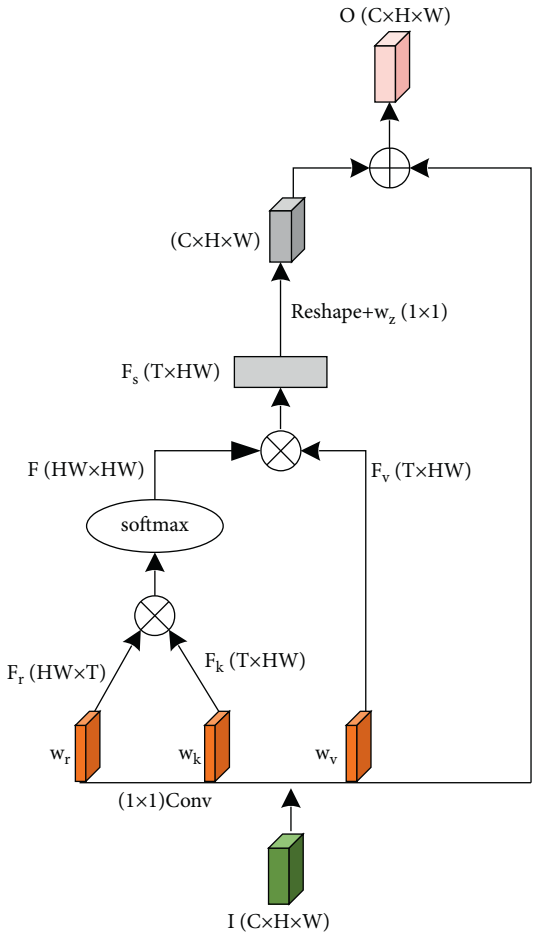

Figure 2: The structure of nonlocal block.

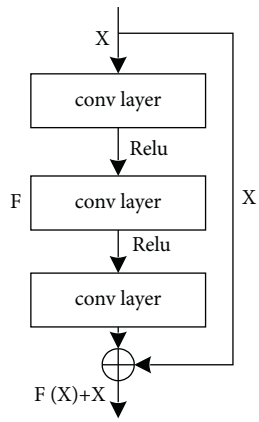

Figure 3: Residual structure.

the original input is added to obtain the feature map after spatial feature optimization.

The channel-attention mechanism is used to capture the interdependence between any two channels and update the value in one channel by using the weighted average of all channels. Its implementation steps are similar to the spatial-attention module. The feature maps obtained by the spatial-attention module and the channel-attention module are added and fused to generate a segmented image strengthened by the attention mechanism.

In Figure 5, the mathematical expression of Softmax is

$$\text{softmax}\left(z_i^l\right) = \frac{e^{z_i^l}}{\sum_{p=0}^{(H\times W)-1} e^{z_p}}, \tag{9}$$

where z_i^l represents the i^{th} pixel value in the l^{th} column of the feature map F. This formula normalizes the similarity measure matrix F by column so that the weight value of each column is within the interval $[0, 1]$.

The nonlocal attention mechanism is combined with the residual network model to construct the generator module in the GAN. The structure is shown in Figure 6.

In Figure 6, the generator module receives image data input, and the input image is extracted features by the ResNet50 network based on the dual-attention mechanism; four output feature maps at different levels are obtained, denoted as $F0$, $F1$, $F2$, and $F3$. The $F1$, $F2$, and $F3$ feature maps are upsampled to the same size as $F0$, and then, the four output feature map channels are spliced and 3×3 convolution is performed to obtain the fused feature map F. The fusion feature map F is spliced with $F0$, $F1$, $F2$, and $F3$ at channel level, respectively, and then, input into the self-attention mechanism module after 3×3 convolution to obtain four prediction segmentation maps. They are added and fused and averaged, and finally, the image multitarget prediction segmentation map is obtained.

2.4. The Adversarial Network Model Based on Convolutional Network. In GAN, the adversarial network module penalizes the lost details and small-size target information of the network through game learning, which makes the multi-target images segmented by the generative network more accurate. In this paper, the adversarial network module in GAN is constructed based on CNN, and its structure is shown in Figure 7.

In Figure 7, the adversarial network model includes two inputs: one is dot-product image of segmentation label and the original image, and the other is dot-product image of generative network and the original image. Dot-product operation is a mask operation on the original image to obtain the area of label and prediction segmentation on the original image. The adversarial network performs feature extraction on the obtained detection area, distinguishes whether the input is a real or predicted segmentation area, and optimizes the segmentation result through the adversarial learning with the generative network. In this paper, the CNN in the adversarial network model contains a total of 5 convolutional layers, 5 pooling layers, and two fully connected layers. The size of the convolution kernel is 3×3, and the step size is 1×1, that is, the convolution operation does not change the size of the input. The pooling layer is the maximum pooling with a size of 2×2, and the step size is 2×2, which reduces the input resolution by half. The ReLu function is selected as the activation function, and the Sigmoid function is used for classification operations.

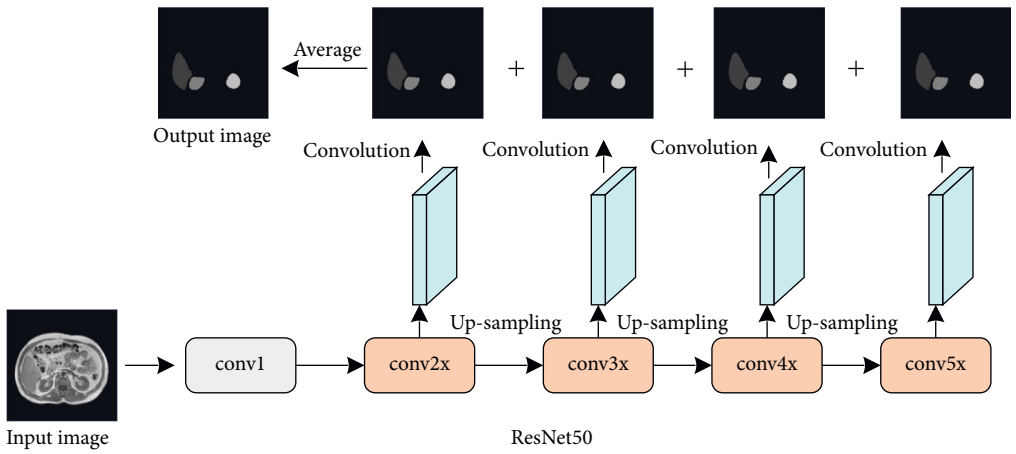

Figure 4: The segmentation model based on residual network.

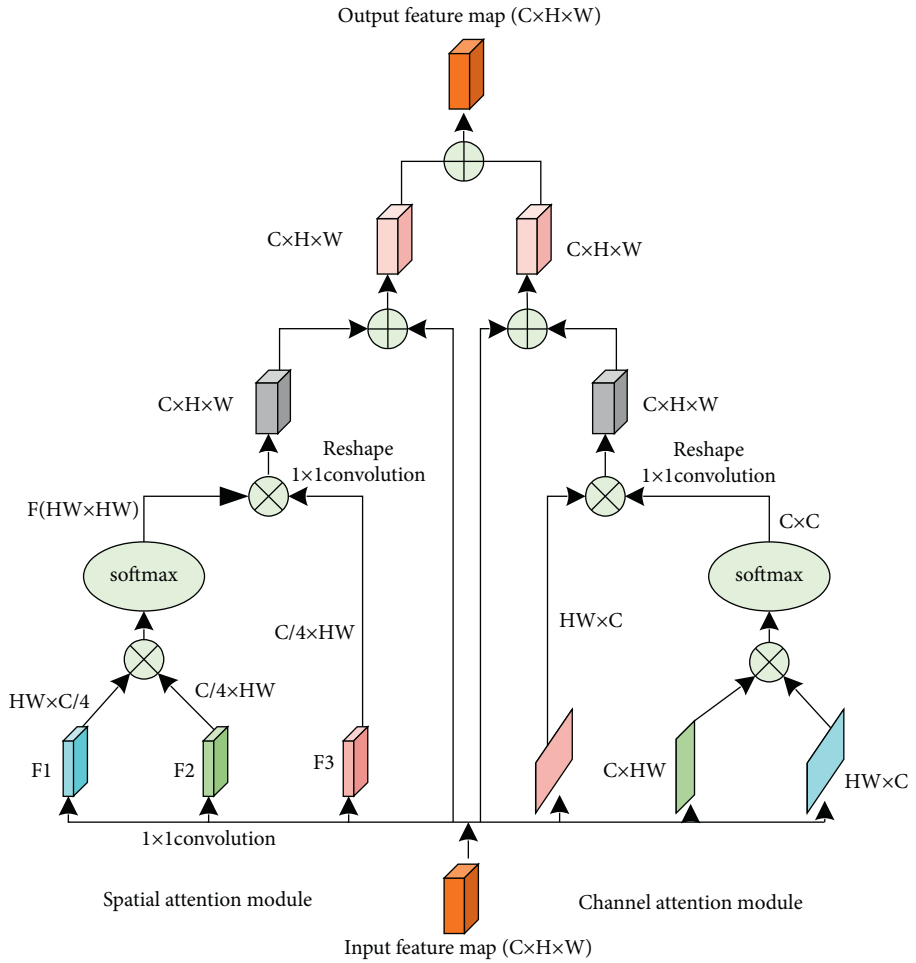

Figure 5: Dual-attention mechanism.

2.5. The Generative Adversarial Network-Fused Attention Mechanism. In this paper, a generative adversarial network segmentation model fused with attention mechanism (AM-GAN) is proposed and the overall structure is shown in Figure 8.

In Figure 8, the gold standard is an accurate result of manual segmentation by experts. In AM-GAN, the generative network is composed of the ResNet50 model and the nonlocal dual-attention mechanism module. It takes the original image as input and the predicted segmentation as

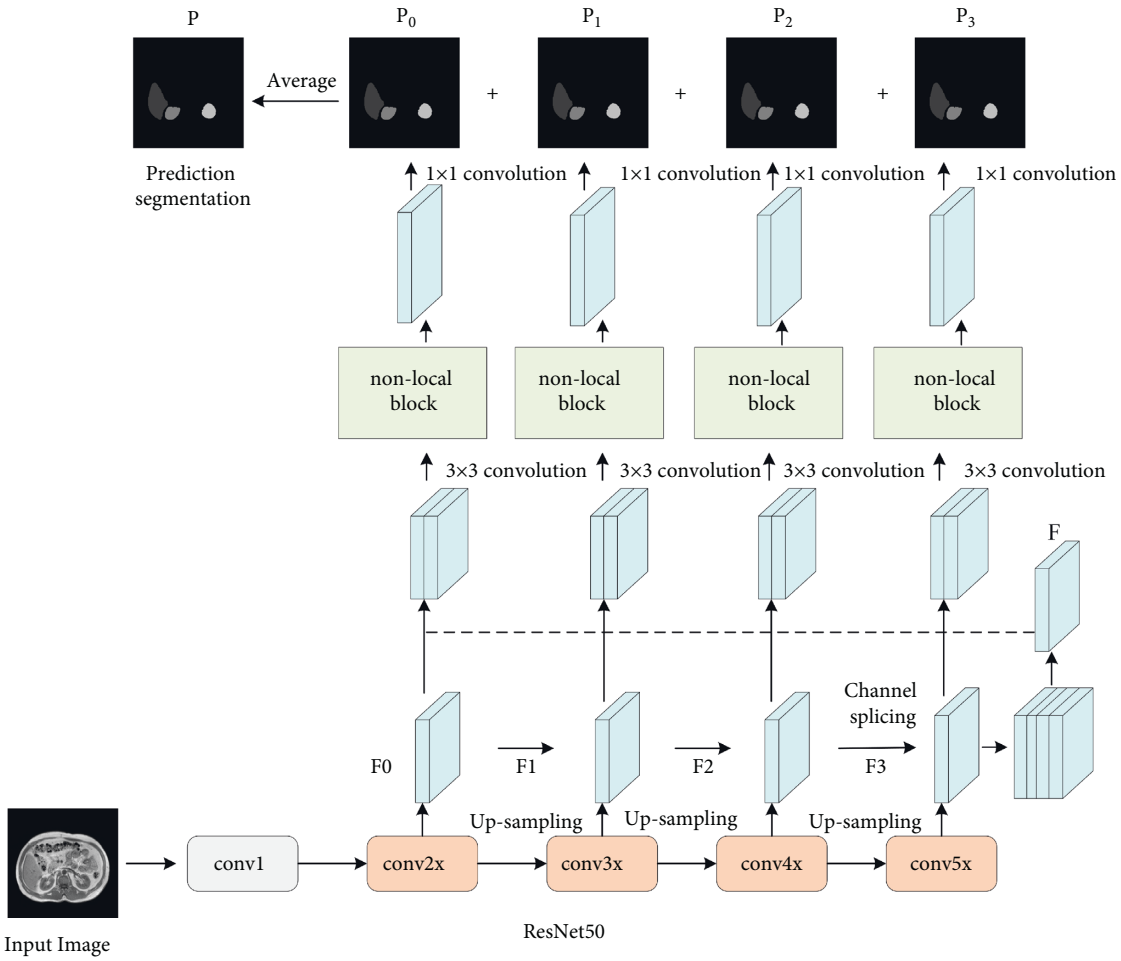

Figure 6: The generate network model.

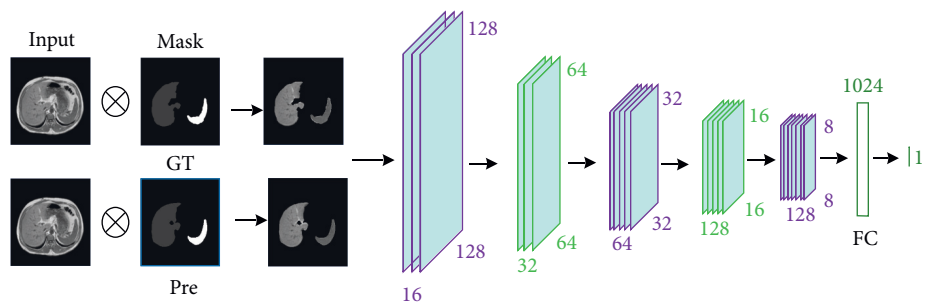

Figure 7: The adversarial network model.

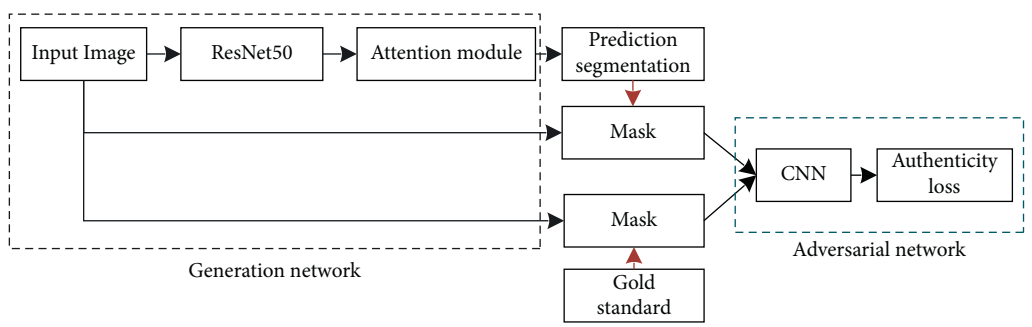

Figure 8: The overall structure of the generative adversarial network segmentation model.

output. The adversarial network is composed of CNNs. Its inputs include the mask of the predicted segmentation and the original image and the mask of the gold standard and the original image. The mask of the predicted segmentation is judged as false (marked as 0), and the mask of the gold standard is judged to be true (marked as 1).

Combining the generative network based on the residual network and the nonlocal dual-attention mechanism with the adversarial network based on the CNNs to build a generative adversarial network model for the multitarget image segmentation. After AM-GAN is trained to reach the optimum, the model used for image segmentation is a generator network. The structure of the AM-GAN segmentation model is shown in Figure 9.

As shown in Figure 9, in the generator network, the *conv*2*x*, *conv*3*x*, *conv*4*x*, and *conv*5*x* parts of the ResNet50 based on the dual-attention mechanism all have a prediction segmentation output, and the final prediction segmentation of the generator network is the average value of the prediction segmentation of these four parts. The input of the generator is subjected to feature extraction through the extended ResNet50 network, and the rough feature map is obtained by upsampling. The calculation formula of the upsampling output F_1 of the *conv*3*x* part in the extended ResNet50 network is

$$F_1 = \text{Upsample}\left(\text{Relu}\left(BN\left(w_1\left(conv3x\left(I\right)+b_1\right)\right)\right)\right), \quad (10)$$

where I represents the input image, $\text{conv3}x\left(I\right)$ represents the convolution output of the *conv*3*x* part, w_1 represents the 1×1 convolution, whose purpose is to reduce the number of channels of the convolution output, b_1 represents the bias, $BN\left(\cdot\right)$ represents the batch normalization, $Relu\left(\cdot\right)$ is the activation function, and $\text{Umsample}\left(\cdot\right)$ is the interpolation upsampling. In this paper, the bilinear interpolation algorithm is used to upsample the feature map, and the output is F_1. Using the same algorithm, the upsampled output F_2 and F_3 of the *conv*4*x* and *conv*5*x* parts can be obtained. The output F_0 of the *conv*2*x* part does not need to be upsampled, and the sizes of F_1, F_2, and F_3 are the same as F_0. After feature extraction and upsampling of the input image, the output information of each part is fused. The calculation formula of the fusion feature map F is

$$F = \text{Relu}\left(w_3\left(\text{Relu}\left(w_2\left(\text{cat}\left(F_0,F_1,F_2,F_3\right)\right)+b_2\right)\right)+b_3\right), \quad (11)$$

where $\text{cat}\left(\cdot\right)$ represents the splicing operation of feature map channel, w_2 is a 3×3 convolution, which is used to fuse the information between the four feature maps, w_3 is a 1×1 convolution, which is used to reduce the number of channels of the fusion feature map, b_2 and b_3 are biases, and $\text{Relu}\left(\cdot\right)$ is the activation function. After the information of each part is fused, the feature map is input into the dual-attention mechanism module for feature enhancement. The calculation formula of the predicted output P_1 of the *conv*3*x* module is

$$\begin{aligned} P_1 = \text{Upsample}\left(w_4\left(p\text{Attention}\left(\text{cat}\left(F,F_1\right)\right)\right.\right. \\ \left.\left.+c\text{Attention}\left(\text{cat}\left(F,F_1\right)\right)\right)+b_3\right), \end{aligned} \quad (12)$$

where $p\text{Attention}\left(\cdot\right)$ represents the spatial-attention mechanism module, $c\text{Attention}\left(\cdot\right)$ represents the channel-attention mechanism module, w_4 represents 1×1 convolution, which is used to convert the number of channels into the number of classification categories, $\text{Upsample}\left(\cdot\right)$ represents a bilinear interpolation up-sampling operation, which is used to restore the size of output feature map to the size of the input image to obtain a predicted segmented image. Using the same method, the predicted output P_0, P_1, and P_3 can be obtained. Finally, P_0, P_1, P_2, and P_3 are added and averaged to obtain the predicted segmented image. The calculation formula is

$$P = \text{avg}\left(P_0+P_1+P_2+P_3\right). \quad (13)$$

Equation (13) is the calculation expression for predicting segmentation P, where $\text{avg}\left(\cdot\right)$ represents the average operation.

In the GAN segmentation model established in this paper, the data processing capability of the ResNet50 extended based on the nonlocal dual-attention mechanism can mechanically ensure that the image multitarget features can be accurately extracted, while nonlocal attention mechanism also strengthens the output feature map of the ResNet50, which can further improve the accuracy of segmentation. The authenticity judgment of the adversarial network can guide the segmentation network to avoid the loss of detailed information as much as possible and optimize boundary segmentation and small-target segmentation.

3. The Learning Algorithm

3.1. The Loss Function. In the complex images' multitarget segmentation, it is necessary to calculate classification errors of multiclass pixel. In this paper, the multiclassification cross-entropy loss [40] is used as the loss function of the generative network, which is defined as follows:

$$L_{\text{mec}}\left(x,y\right) = -\sum_{i=1}^{(H\times W)}\sum_{c=1}^{C}x_{ic}\ln y_{ic}. \quad (14)$$

Equation (14) represents the multiclassification cross-entropy loss of the label image x and the predicted segmentation y, where (H,W,C) represents the length, width, and number of channels of the image.

In the AM-GAN segmentation model, the adversarial network uses the authenticity loss to perform game learning with the generative network, which is essentially a binary classification problem. Therefore, the two-classification cross-entropy loss [41] is used as the loss function of the adversarial network, which is defined as follows:

$$L_{\text{bec}}\left(x,y\right) = -[x\ \ln\ y+(1-x)\ln\left(1-y\right)]. \quad (15)$$

Equation (15) represents the two-classification cross-entropy loss, where x represents the classification label and y represents the classification probability output of the adversarial network.

The loss function of the generative network is defined as follows:

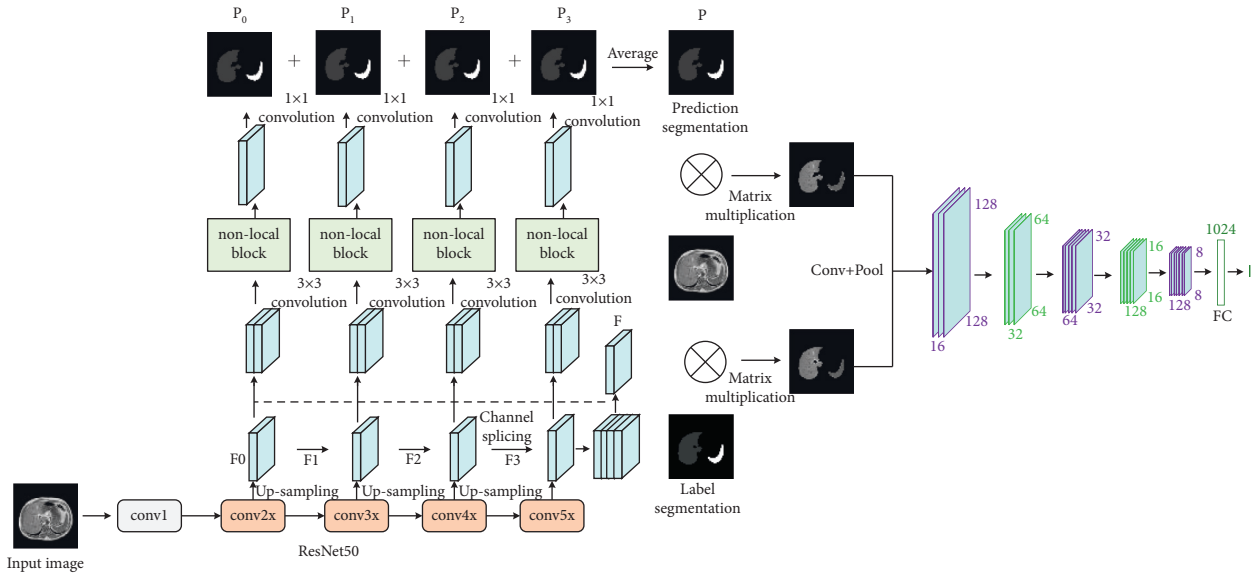

Figure 9: The generative adversarial network segmentation model fused with attention mechanism.

$$G(x_n, y_n, \theta_g) = L_{\text{mec}}\left(g(x_n, \theta_g), y_n\right) + \lambda L_{\text{bec}}\left(d\left(g(x_n, \theta_g), x_n\right), 1\right), \tag{16}$$

where x_n is the input image, y_n is the label segmentation, and θ_g is the training parameter set of the generative network. According to equation (16), the loss of generative network is composed of two parts, that is, the multiclassification cross-entropy loss of prediction segmentation and label segmentation and the loss of adversarial network which predict segmentation. The hyperparameter λ is used to balance these two losses. When optimizing the generative network, we minimize the objective function.

The loss function of the adversarial network is defined as

$$D(x_n, y_n, \theta_d) = L_{\text{bec}}(d(x_n, y_n, \theta_d), 1) + L_{\text{bec}}(d(x_n, g(x_n), \theta_d), 0), \tag{17}$$

where θ_d represents the training parameter set of the adversarial network. From (17), the loss function of the adversarial network consists of two parts. The first is the two-classification cross-entropy loss of the mask map of input image x_n and the label segmentation y_n, which is judged to be true (that is, the value is 1). The second is the two-classification cross-entropy loss of the mask map of input image x_n and the label segmentation $g(x_n)$, which is judged to be false (that is, the value is 0). When optimizing the adversarial network, we minimize the loss function.

The loss function of GAN is composed of the loss of the generative network and the loss of the adversarial network. The calculation expression is

$$V(G, D) = G\left(x_n, y_n, \theta_g\right) + D(x_n, y_n, \theta_d). \tag{18}$$

3.2. The Training Process. AM-GAN adopts the method of alternate training of generative network and adversarial network. In order to ensure the stability of training, the training times of the discriminating module are generally more than that of the generative module. In this paper, the minibatch gradient descent (MBGD) algorithm is used for AM-GAN training. The specific process is as follows:

(1) Randomly sample n samples z_n from noise samples, and n samples x_n from real samples.

(2) Gradient ascent algorithm is used to update the adversarial network:

$$\nabla_{\theta_d} \frac{1}{n} \sum_{i=1}^{n} \left(\log\left(D\left(x^i\right)\right) + \log\left(1 - D\left(G\left(z^i\right)\right)\right)\right). \tag{19}$$

(3) Repeat steps (1) and (2) k times to update the adversarial network k times.

(4) Sampling n generated samples from the noise samples, and we update the generative network once using gradient descent algorithm:

$$\nabla_{\theta_s} \frac{1}{n} \sum_{i=1}^{n} \log\left(1 - D\left(G\left(z^i\right)\right)\right). \tag{20}$$

(5) Sequentially, we repeat the above steps until the model training is stable and optimal.

The specific algorithm implementation is as follows:

Step 1: determine and initialize the GAN training parameter set. The training parameter set of the generated network is $\theta_g = (w_0, w_1, \ldots, w_m, b_0, b_1, \ldots, b_m, g, \theta, \varphi, w_z, w_c)$, where $(w_0, w_1, \ldots, w_m)$ is the set of convolution kernel weight of extended ResNet50 network, $(b_0, b_1, \ldots, b_m)$ is the set of biases, g, θ, φ, and w_z are 1×1 convolution kernel weights of spatial-attention mechanism, and w_c is 1×1 convolution kernel weights of channel-attention mechanism. The training parameter set

of the adversarial network is $\theta_d = (w_0, w_1, \ldots, w_n, b_0, b_1, \ldots, b_n)$, where $(w_0, w_1, \ldots, w_n)$ is the set of convolution kernel weights of CNN, and $(b_0, b_1, \ldots, b_n)$ is the set of biases.

Step 2: randomly select N image samples x_n from the sample set, and select corresponding N label segmentation samples y_n.

Step 3: calculate the loss of the adversarial network: $\nabla L(\theta_d) = 1/N \sum_{n=1}^{N} [L_{bec}(d(x_n, y_n, \theta_d), 1) + L_{bec}(d((x_n, g(x_n), \theta_d)), 0)]$.

Step 4: calculate the training parameter gradient according to the chain rule, $a_i = \sigma(z_i) = \sigma(w_i a_{i-1} + b_i)$. σ is the activation function, a_i is the input feature map of the i^{th} layer, z_i is the feature map of the i^{th} layer after convolution, w_i is the weight of the i^{th} layer, and b_i is the bias of the i^{th} layer.

The formula for calculating the weight gradient using the chain rule is

$$\frac{\partial L(\theta_d)}{\partial w_i} = \frac{\partial L(\theta_d)}{\partial z_i} \cdot a_{i-1} = \sigma'(z_i) \cdot (w_{i+1})^T \cdot \left(\frac{\partial L(\theta_d)}{\partial z_{i+1}}\right) \cdot a_{i-1}, \tag{21}$$

where $\sigma'(\cdot)$ is the derivative of the activation function and $(\cdot)^T$ represents the matrix transpose.

The formula for calculating the bias gradient using the chain rule is

$$\frac{\partial L(\theta_d)}{\partial b_i} = \frac{\partial L(\theta_d)}{\partial z_i} = \sigma(z_i) \cdot (w_{i+1})^T \cdot \left(\frac{\partial L(\theta_d)}{\partial z_{i+1}}\right). \tag{22}$$

Step 5: use the MBGD optimizer to update the convolution weights and biases of the adversarial network:

$$w_i^t = w_i^{t-1} - \eta \frac{1}{N} \sum_{n=1}^{N} \frac{\partial L\left(\theta_d^{t-1}\right)}{\partial w_i^{t-1}}, \tag{23}$$

$$b_i^t = b_i^{t-1} - \eta \frac{1}{N} \sum_{n=1}^{N} \frac{\partial L\left(\theta_d^{t-1}\right)}{\partial b_i^{t-1}}. \tag{24}$$

Step 6: repeat Step 2 Step 5 k times.

Step 7: randomly select N image samples x_n, and select corresponding N label segmentation samples y_n.

Step 8: calculate the loss $\nabla L(\theta_g)$ of the generative network:

$$\nabla L\left(\theta_g\right) = \frac{1}{N} \sum_{n=1}^{N} \left[L_{\text{mec}}\left(g\left(x_n, \theta_g\right), y_n\right) + \lambda L_{\text{bec}}\left(d\left(g\left(x_n, \theta_g\right), x_n\right), 1\right)\right]. \tag{25}$$

Step 9: use the MBGD optimizer to update the parameters θ_g of the generation network.

Step 10: if the network converges to the optimal, then the training ends; otherwise, it returns to Step 2.

The epoch of training is set to 200 and the batch size is 10. According to the above algorithm process, the pseudocode of the training algorithm is shown in Algorithm 1.

4. The Experiment and Result Analysis

4.1. The Experiment Dataset. The dataset comes from the Combined Healthy Abdominal Organ Segmentation (CHAOS) competition dataset [42]. The CHAOS dataset contains the abdomen MRI images. In the experiment, the abdominal MRI abdominal images containing the label information of the liver, left/right kidney, and spleen were selected as the sample set. The typical abdomen image and label segmentation image are shown in Figure 10.

The experimental dataset contains 38 groups of abdominal MRI images, and each group has 26 slices with a size of 256 × 256, totaling 988 slices. The dataset is divided into a training set, a validation set, and a test set. The training set includes 30 groups of slices, the validation set includes 2 groups of slices, and the test set includes 6 groups of slices. Due to the small number of samples in the dataset, random rotation, mirroring, and other operations are used to enhance the data in the experiment. The final number of slices in the training set is expanded to 3120, the verification set is expanded to 104, and the test set is expanded to 312. The sample distribution of dataset is shown in Table 1.

4.2. The Model Structure and Parameter Settings. The training strategy of the AM-GAN is to alternate training between the generative network and the adversarial network. In order to ensure the stability of training, the ratio of the training times of the generative network and the adversarial network is set to 1 : 6, that is, the adversarial network training is performed 6 times first, and then, the generative network is trained once.

The size of the convolution kernel of the extended ResNet50 in the generative network is all 3 × 3, the step size of the downsampling is 2 × 2, and the other step size is 1 × 1. The activation function adopts the ReLu function, and the balance coefficient λ of the loss function of generative network is set to 0.2. The adversarial network includes 5 convolutional layers, 5 pooling layers, and 2 fully connected layers. The convolution kernel size of the convolutional layer is set to 3 × 3, and the step size is 1 × 1. The pooling window size of the pooling layer is set to 2 × 2, and the step size is 2 × 2. The Sigmoid function is used as the classification function. The learning rate of the MBGD optimizer is set to 0.01, and the weight parameters are initialized with truncated normal distribution. The number of training iterations is 200, and the batch of one training is 10. The experimental environment is Linux system with NVIDIA GeForce RTX 2080Ti GPU.

4.3. The Experiment Result and Analysis. The AM-GAN training is carried out according to the algorithm strategy in

Input: abdomen MRI images
Output: abdominal multiorgan prediction segmentation image
(1) Hyperparameters' setting: epochs = 200, hyperparameter $k = 6$, batch size $N = 10$, learning rate $\eta = 0.01$, balance coefficient = 0.2.
(2) **for** epochs **do**
(3) **for** k **do**
(4) N original abdominal MRI images were randomly selected, and N corresponding label segmentation images were selected;
(5) Calculate the loss of the adversarial network by formula (17);
(6) Update the training parameters of the adversarial network by formulas (21) and (22);
(7) N original abdominal MRI images were randomly selected, and N corresponding label segmentation images were selected;
(8) Calculate the loss of the generative network by formula (25);
(9) Update the training parameters of the generative network by formulas (23) and (24);

Algorithm 1: The AM-GAN training algorithm.

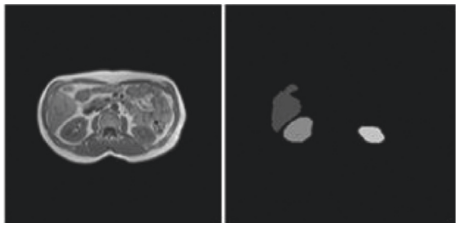

Figure 10: Data sample from CHAOS.

Table 1: Experimental dataset.

Slice type	Training set	Validation set	Test set	Total
Original slice	780	52	156	988
Extended slice	2340	52	156	2548
Total	3120	104	312	3536

Section 3 and the model structure and parameter setting in Section 4.2. Multitarget image segmentation is performed by generative network. Recall rate, precision, F1-score, accuracy, and Dice similarity coefficient are used to evaluate the properties of the segmentation model. The image segmentation experiment results of the test set are shown in Figure 11, and the index evaluation results are shown in Table 2.

In Figure 11, from left to right, there is the original abdominal image, the gold standard, that is, the image manually segmented by the expert, and the predicted segmented image by network. From Figure 11, the segmentation results of the GAN are clear and smooth. Whether it is a large-sized organ such as the liver or a small-sized organ such as the kidney and pancreas, the GAN can distinguish them more accurately, which shows that the proposed method has good applicability.

For analyzing the properties of the learning algorithm, the curves of the Dice coefficient indicators of four organs and tissues, including liver, left kidney, right kidney, and spleen changing with iteration, and the confusion matrix of the validation set are recorded. The results are shown in Figure 12 and Table 3.

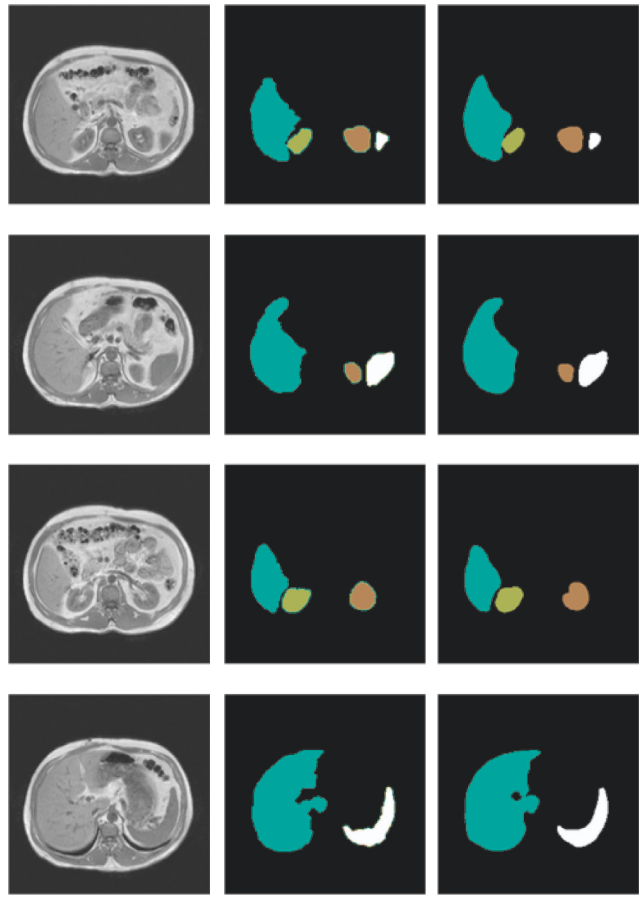

Figure 11: Comparison of model prediction segmentation and gold standard.

Table 2: The evaluation results of AM-GAN.

Recall (%)	Precision (%)	F1-score (%)	Accuracy (%)
90.3	92.1	91.2	92.3

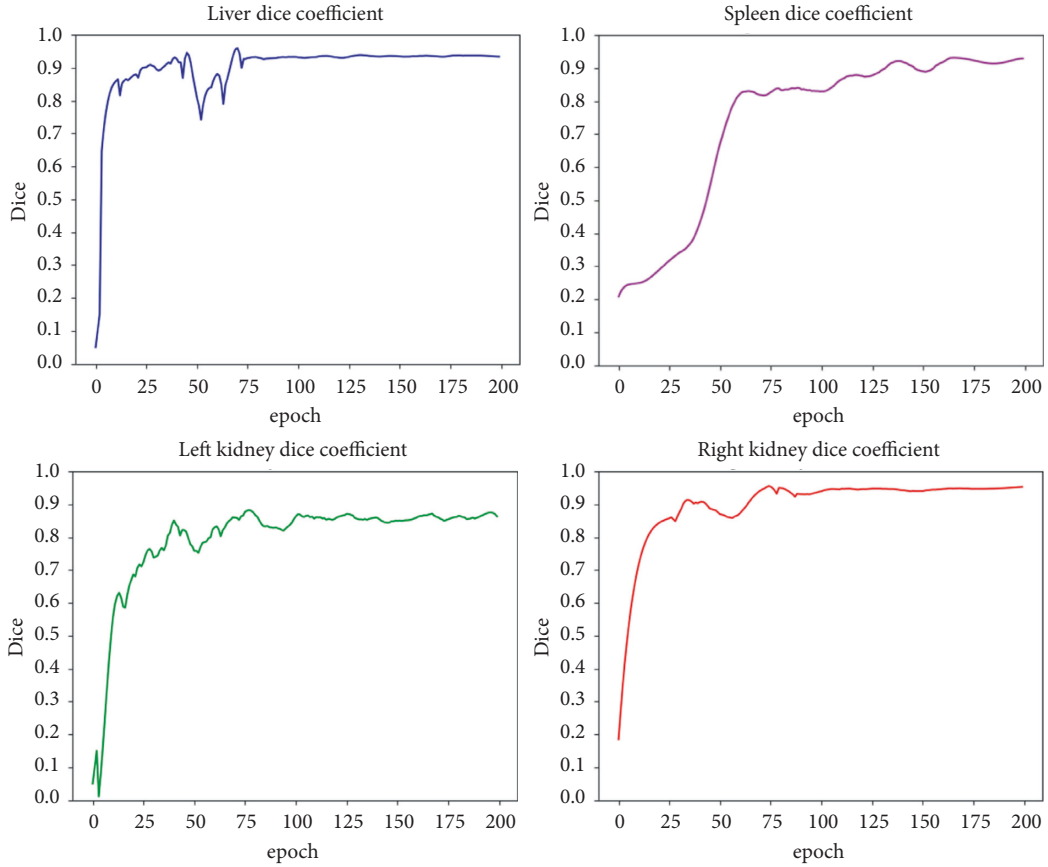

Figure 12: Variation curve of Dice coefficient of the segmented organ.

Table 3: Model numerical confusion matrix.

Type	Background	Liver	Left kidney	Right kidney	Spleen
Background	99.32%	0.59%	0.02%	0.02%	0.05%
Liver	1.89%	97.55%	0.24%	0.14%	0.18%
Left kidney	9.28%	7.73%	82.39%	0.29%	0.31%
Right kidney	4.72%	0%	0%	92.68%	2.6%
Spleen	4.71%	0%	0%	1.17%	94.12%

In Figure 12, the Dice similarity coefficient training curves of the liver, left kidney, right kidney, and spleen are shown. The Dice curve of the liver converges quickly and is relatively stable in the later iterations. This is because its size is relatively large compared to other organs, which causes the segmentation network to be more inclined to learn the classification of the liver area pixels in the early stage. The left kidney, right kidney, and spleen organs are smaller in size than the liver, and the segmentation network has a deeper level, which leads to easy loss of information and large fluctuations in the training. In the later stage of training, the average Dice coefficient of liver, left kidney, right kidney, and spleen on the training set is 0.92, which shows that the overall effect of model segmentation is good, and it has good applicability to small-target segmentation.

From Table 3, the segmentation accuracy of liver organs is the highest, 97.55%. This is because the size of liver organs is relatively large, and the network tends to learn their pixel weights of liver organs. The spatial location of the left kidney and the liver area is relatively close, and the boundary tissues overlap, leading to the segmentation error of the left kidney mainly from the liver and background. The right kidney and spleen organs are close in space, leading to mutual influence. The right kidney and spleen are small-target organs, and there is no boundary overlap, so the segmentation is less affected than the left kidney. From the confusion matrix, the accuracy of left kidney segmentation was 82.39%, while that of right kidney segmentation was 92.68% and that of spleen segmentation was 94.12%. It shows that the method in this paper has a good ability to identify and distinguish the features of complex image pixel categories.

4.4. Comparative Experiment and Analysis. In order to verify the performance improvement of AM-GAN brought by the nonlocal attention mechanism, a ResNet-GAN model was constructed in the comparative experiment. This model removes the nonlocal attention mechanism in AM-GAN, and other structures are the same as the AM-GAN. In addition, current mainstream, AM-FCN [43] embedded based on attention, Attention U-Net [44] fused on attention in both encoder and decoder, DANet [45] embedded based on dual attention, and SEVNet [46] fused on SE module are

Table 4: Comparison of results of different models.

Model	Recall (%)	Precision (%)	F1-score (%)	Accuracy (%)
FCN	85.4	86.2	83.5	85.6
U-net	87.1	87.4	87.2	86.9
DANet	91.7	90.3	91.0	91.2
ResNet-GAN	89.2	91.5	90.1	91.2
SEVNet	86.5	87.6	86.8	87.1
Proposed	90.3	92.1	91.2	92.3

selected as comparison models. The four index values of accuracy, recall, precision, and F1-score are used for comparative evaluation. In order to avoid misleading the performance evaluation by high background indicators, all evaluation indicators do not include the value of background indicators. The specific results are shown in Table 4.

From Table 4, the four indexes of AM-FCN and Attention U-Net are lower than the proposed model. This is because it organically integrates GAN's inherent ability to process and repair low-resolution images, and the global to local content association and feature enhancement mechanism of nonlocal attention improves the accuracy of multitarget segmentation results. Except for the recall rate, other indexes of DANet are lower than that of the proposed model, which shows that the nonlocal attention mechanism used in this model can gather the receptive field from global to local target area to realize information context correlation. In addition, GAN can optimize the segmentation results in mechanism and improve the segmentation properties of the model, and the model and algorithm are relatively robust. Compared with ResNet-GAN, the accuracy of proposed model is 1.1% higher, which shows that the nonlocal attention can effectively realize the information association from global to local. Combined with GAN's probability exploration mechanism, the segmentation accuracy of the model is comprehensively improved. Compared with the SEVNet, due to the inherent high-definition processing and repairing capabilities of the GAN for noise images and the use of a game strategy, the optimization of the segmentation results is realized, and the accuracy, continuity, and smoothness of the segmentation results are comprehensively improved. Based on the above analysis, it shows that the method in this paper has comprehensive advantages when performing image target segmentation with insignificant structural features and achieves a better segmentation effect.

5. Conclusion

In this paper, aiming at the fine segmentation of multitarget complex images, a generative adversarial network model fused with attention mechanism is proposed. In mechanism, the AM-GAN can use the ability to process and restore high-definition images of GAN to effectively reduce the effects of noise, offset distortion, and gray value distortion. Nonlocal spatial-channel dual attention is introduced to realize the information association and constraint of large receptive field content on local targets and maintain the continuity of segmentation results. At the same time, the Nash game strategy of the generative network and the adversarial network is adopted, which reduces the algorithm's loss of detailed features and effectively improves the accuracy of small-target segmentation. In terms of information processing mechanism, this method comprehensively realizes the context correlation of image information, the feature fusion of different levels and scales, the high-definition processing and repair of high-noise images, and the optimization of segmentation results. The experimental results show that the multitarget segmentation method proposed in this paper has good applicability for both small-size and large-size targets. Compared with the other methods, each evaluation index has been greatly improved. AM-GAN comprehensively utilizes the advantages of nonlocal attention mechanism and generative adversarial network, which can finely segment multi-instance targets in complex images. It has good applicability in mechanism to solve the image segmentation problem of insignificant morphological features and weak spatial information relevance. It improves the limitations and deficiencies of the comparison method in solving the above problems, provides a novel deep learning method for image segmentation, and has great application value and a good prospect for promotion. However, the information processing mechanism and algorithm process of the method in this paper are more complicated, and the image semantic knowledge and structural features are less used. From the experimental results, the proportion of image background pixels compared with target area pixels is too large, and the number of samples in the dataset is unbalanced, which still has a certain impact on the segmentation accuracy of this method. How to optimize the model and algorithm, improve the discrimination ability of the local target image and the background image, and embed the scene semantic feature knowledge into the segmentation model, which will be an important work in the next stage of research.

Data Availability

The data that support the findings of this study are available upon request from the corresponding author.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

This work was supported by the Shandong University of Science and Technology Research Fund, under Grant 2019TDJH102.

Supplementary Materials

The AM-GAN model combines the generative network based on the residual network and the nonlocal dual-attention mechanism with the adversarial network based on the CNNs to build a generative adversarial network model for the multitarget image segmentation. After AM-GAN is trained to reach the optimum, the model used for image segmentation is a generator network. (*Supplementary Materials*)

References

[1] S. Minaee, Y. Y. Boykov, F. Porikli, A. J. Plaza, N. Kehtarnavaz, and D. Terzopoulos, "Image segmentation using deep learning: a survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 15, 2020.

[2] C. Chen, C. Qin, H. Qiu et al., "Deep learning for cardiac image segmentation: a review," *Frontiers in Cardiovascular Medicine*, vol. 7, p. 25, 2020.

[3] Z. Zhou, M. M. Rahman Siddiquee, N. Tajbakhsh, and J. Liang, "Unet++: a nested u-net architecture for medical image segmentation," *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, vol. 28, pp. 3–11, 2018.

[4] Z. Gu, J. Cheng, H. Fu et al., "CE-net: context encoder network for 2D medical image segmentation," *IEEE Transactions on Medical Imaging*, vol. 38, no. 10, pp. 2281–2292, 2019.

[5] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-net: Fully convolutional neural networks for volumetric medical image segmentation," in *Proceedings of the 2016 Fourth International Conference on 3D Vision (3DV)*, pp. 565–571, IEEE, Stanford, CA, USA, October 2016.

[6] M. H. Hesamian, W. Jia, X. He, and P. Kennedy, "Deep learning techniques for medical image segmentation: achievements and challenges," *Journal of Digital Imaging*, vol. 32, no. 4, pp. 582–596, 2019.

[7] M. I. Razzak, S. Naz, and A. Zaib, "Deep learning for medical image processing: overview, challenges and the future," *Lecture Notes in Computational Vision and Biomechanics*, vol. 24, pp. 323–350, 2018.

[8] D. Liu, S. Wang, D. Huang, G. Deng, F. Zeng, and H. Chen, "Medical image classification using spatial adjacent histogram based on adaptive local binary patterns," *Computers in Biology and Medicine*, vol. 72, pp. 185–200, 2016.

[9] Y. Zhang, Y. Liu, H. Cheng, Z. Li, and C. Liu, "Fully multitarget segmentation for breast ultrasound image based on fully convolutional network," *Medical, & Biological Engineering & Computing*, vol. 58, no. 9, pp. 2049–2061, 2020.

[10] Y. Tian, A. Dehghan, and M. Shah, "On detection, data association and segmentation for multi-target tracking," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 9, pp. 2146–2160, 2018.

[11] W. Dinghan, F. Guilan, W. Xiong, W. Yufeng, and D. Maohua, "Research on image segmentation algorithm based on features of venous gray value," *Opto-Electronic Engineering*, vol. 45, no. 12, pp. 180066–180071, 2018.

[12] A. Makandar and B. Halalli, "Threshold based segmentation technique for mass detection in mammography," *Journal of Computers*, vol. 11, no. 6, pp. 472–478, 2016.

[13] D. Gupta and R. S. Anand, "A hybrid edge-based segmentation approach for ultrasound medical images," *Biomedical Signal Processing and Control*, vol. 31, pp. 116–126, 2017.

[14] R. Kashyap and P. Gautam, "Modified region based segmentation of medical images," in *Proceedings of the 2015 International Conference on Communication Networks (ICCN)*, pp. 209–216, IEEE, Gwalior, India, November 2015.

[15] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in Neural Information Processing Systems*, vol. 25, pp. 1097–1105, 2012.

[16] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, Las Vegas, NV, USA, June 2016.

[17] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: towards real-time object detection with region proposal networks," *Advances in Neural Information Processing Systems*, vol. 28, pp. 91–99, 2015.

[18] J. Wang, H. Zhu, S.-H. Wang, and Y.-D. Zhang, "A review of deep learning on medical image analysis," *Mobile Networks and Applications*, vol. 26, no. 1, pp. 351–380, 2021.

[19] I. Despotović, B. Goossens, and W. Philips, "Mri segmentation of the human brain: challenges, methods, and applications, computational and mathematical methods in medicine," 2015.

[20] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3431–3440, Boston, MA, USA, June 2015.

[21] O. Ronneberger, P. Fischer, T. Brox, and U-net, "U-net: convolutional networks for biomedical image segmentation," in *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 234–241, Munich, Germany, October 2015.

[22] Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger, "3d u-net: learning dense volumetric segmentation from sparse annotation," *Medical Image Computing and Computer-Assisted Intervention - MICCAI 2016*, vol. 32, pp. 424–432, 2016.

[23] Q. Zhang, Z. Cui, X. Niu, S. Geng, and Y. Qiao, "Image segmentation with pyramid dilated convolution based on resnet and u-net," *Neural Information Processing*, vol. 44, pp. 364–372, 2017.

[24] D. Nguyen, X. Jia, D. Sher et al., "3d radiotherapy dose prediction on head and neck cancer patients with a hierarchically densely connected u-net deep learning architecture," *Physics in Medicine and Biology*, vol. 64, no. 6, Article ID 065020, 2019.

[25] O. Oktay, J. Schlemper, L. L. Folgoc et al., "Attention U-net: learning where to look for the pancreas," 2013, https://arxiv.org/abs/1804.03999.

[26] V. Badrinarayanan, A. Kendall, and R. Cipolla, "Segnet: a deep convolutional encoder-decoder architecture for image segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 12, pp. 2481–2495, 2017.

[27] Y. Chen, J. Tao, L. Liu et al., "Research of improving semantic image segmentation based on a feature fusion model," *Journal of Ambient Intelligence and Humanized Computing*, vol. 7, pp. 1–13, 2020.

[28] J. Zhang, Y. Liu, H. Liu, J. Wang, and Y. Zhang, "Distractor-aware visual tracking using hierarchical correlation filters adaptive selection," *Applied Intelligence*, vol. 16, pp. 1–19, 2021.

[29] Y. Chen, L. Liu, V. Phonevilay et al., "Image super-resolution reconstruction based on feature map attention mechanism," *Applied Intelligence*, vol. 14, pp. 1–14, 2021.

[30] Y. Chen, H. Zhang, L. Liu et al., "Research on image inpainting algorithm of improved total variation minimization method," *Journal of Ambient Intelligence and Humanized Computing*, vol. 22, pp. 1–10, 2021.

[31] Y. Chen, L. Liu, J. Tao et al., "The improved image inpainting algorithm via encoder and similarity constraint," *The Visual Computer*, vol. 37, no. 7, pp. 1691–1705, 2021.

[32] I. Goodfellow, J. Pouget-Abadie, M. Mirza et al., "Generative adversarial nets," *Advances in Neural Information Processing Systems*, vol. 27, 2013.

[33] U. Upadhyay and S. P. Awate, "A mixed-supervision multilevel gan framework for image quality enhancement," in *Proceeedigs of the International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 556–564, Athens, Greece, October 2019.

[34] Y. Chen, H. Zhang, L. Liu et al., "Research on image inpainting algorithm of improved gan based on two-discriminations networks," *Applied Intelligence*, vol. 51, no. 6, pp. 3460–3474, 2021.

[35] Z. Niu, G. Zhong, and H. Yu, "A review on the attention mechanism of deep learning," *Neurocomputing*, vol. 452, pp. 48–62, 2021.

[36] C. Yan, Y. Tu, X. Wang et al., "Stat: spatial-temporal attention mechanism for video captioning," *IEEE Transactions on Multimedia*, vol. 22, no. 1, pp. 229–241, 2019.

[37] R. Xu, Y. Tao, Z. Lu, and Y. Zhong, "Attention-mechanism-containing neural networks for high-resolution remote sensing image classification," *Remote Sensing*, vol. 10, no. 10, p. 1602, 2018.

[38] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7794–7803, Salt Lake City, UT, USA, June 2018.

[39] A. Santoro, R. Faulkner, D. Raposo et al., "Relational recurrent neural networks," 2016, https://arxiv.org/abs/1806.01822.

[40] J.-W. Liu, Y.-F. Wang, R.-K. Lu, and X.-L. Luo, "Multi-view non-negative matrix factorization discriminant learning via cross entropy loss," in *Proceedings of the 2020 Chinese Control and Decision Conference (CCDC)*, pp. 3964–3971, Kunming, China, June 2020.

[41] D. Ramos, J. Franco-Pedroso, A. Lozano-Diez, and J. Gonzalez-Rodriguez, "Deconstructing cross-entropy for probabilistic binary classifiers," *Entropy*, vol. 20, no. 3, p. 208, 2018.

[42] A. E. Kavur, N. S. Gezer, M. Barış et al., "CHAOS Challenge - combined (CT-MR) healthy abdominal organ segmentation," *Medical Image Analysis*, vol. 69, Article ID 101950, 2021.

[43] Z. Yue, F. Gao, Q. Xiong, J. Wang, A. Hussain, and H. Zhou, "A novel attention fully convolutional network method for synthetic aperture radar image segmentation," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 13, pp. 4585–4598, 2020.

[44] M. Islam, V. Vibashan, V. J. M. Jose, N. Wijethilake, U. Utkarsh, and H. Ren, "Brain Tumour segmentation and survival prediction using 3d attention unet," in *Proceedings of the International MICCAI Brainlesion Workshop*, pp. 262–272, Strasbourg, France, September 2019.

[45] J. Fu, J. Liu, H. Tian et al., "Dual attention network for scene segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3146–3154, Long Beach, CA, USA, 2019.

[46] Y. Zhao, J. Chen, X. Xu, J. Lei, and W. Zhou, "Sev-net: residual network embedded with attention mechanism for plant disease severity detection," *Concurrency and Computation: Practice and Experience*, vol. 33, no. 10, Article ID e6161, 2021.