Improved WGAN for Image Generation Methods

Jionghui Wang[1(✉)], Jiale Wu[2], Xueyu Huang[2,3], and Zhilin Xiong[2]

[1] Minmetals Exploration and Development Co. Ltd., Beijing 100010, China
wangjh@minmetals.com
[2] School of Software Engineering, Jiangxi University of Science and Technology, Nanchang 330013, People's Republic of China
[3] Nanchang Key Laboratory of Virtual Digital Factory and Cultural Communications, Nanchang 330013, People's Republic of China

Abstract. For the problem of generating high-quality and diverse images, an image generation method combining residual module, spectral parametric normalization, and self-attention mechanism is proposed to be applied in WGAN networks. The specific improvement of the method is to introduce the residual module into the generator and discriminator networks to better capture the deep image information. The spectral parametric normalization technique is also applied to each convolutional layer of the residual block to improve the stability of the image generation process. The self-attention mechanism is introduced into the generator to enable the network to learn in a targeted manner and generate higher-quality images. The experimental results demonstrate that the combined application of these techniques can effectively solve the challenge of generating image samples, obtain stable and diverse data samples, generate better results than the original WGAN method and DCGAN method, and use the generated data samples as the dataset for expanding the classification experiments, which improves the recognition accuracy of the image classification network to a certain extent.

Keywords: Image generation · Generative adversarial network · Residual network · Self-attention mechanism · Spectral parametric normalization

1 Introduction

At present, deep learning has made great progress in the field of image recognition, including real-time object detection, semantic segmentation, face recognition, and image generation, but the implementation of these tasks requires a large dataset for training support, and the dataset collection workload is large, time-consuming, and requires professional equipment for operation, so it has limitations.

In response to the difficulty of collecting sample data in the field of image classification, we want to expand the data set by data augmentation with small samples. Currently, traditional data enhancement and adversarial methods are commonly used to expand data. Traditional data augmentation mainly consists of panning, rotating,

C. Wu et al. (Eds.): MONAMI 2023, LNICST 559, pp. 199–211, 2024.
https://doi.org/10.1007/978-3-031-55471-1_15

and scaling of images, but the diversity of images obtained in this way is insufficient. The generative adversarial approach, which generates images by training a generative adversarial network, was used by Wang et al. [1] used WGAN method for data enhancement to solve the problem of insufficient and unbalanced sample data for tomato leaf disease identification, and Guo et al. [2] proposed a conditional Wasserstein generative adversarial network model for image generation, which not only can effectively improve the accuracy of image generation, but also can improve the convergence speed of the network, although some experiments show that the quality of images generated by generative adversarial networks is better than traditional enhancement methods to some extent, but the training process will be unstable and pattern collapse during the training of generative adversarial networks [3]. However, during the training process of generative adversarial networks, problems such as unstable training process and pattern collapse can occur, which can also lead to poor quality of the generated images or failure to generate normal images [4]. However, during the training process of generative adversarial networks, problems such as stability of the training process and collapse of the model may occur, which may also lead to low quality or failure to generate normal images. Qiu et al. [5] proposed a DCGAN method combining spectral normalization and self-attention mechanism to address the problems of instability and poor generation effect in the training process, and Liu et al. [6] propose an algorithmic model of feature graph connectivity generation adversarial network to alleviate the problem of gradient disappearance.

An improved WGAN image generation method is proposed in response to the above analysis. The method replaces the network structure of the generator and discriminator with the residual structure to avoid problems such as pattern collapse during training; introduces spectral parametric normalization in the residual structure to improve the stability of overall network training; and introduces a self-attention mechanism in the generator to enhance the extraction of deep-level image features. The quality of the images generated by the improved WGAN method is experimentally demonstrated to be improved, and the accuracy in classification recognition is also improved to a certain extent.

2 Related Content

2.1 Generative Adversarial Networks

Generative Adversarial Networks [7] is a framework consisting of a deep learning model for generating realistic synthetic data. Its structure consists of a generator, which receives a random noise vector as input and generates as realistic a sample of synthetic data as possible, and a discriminator, which is a binary classifier whose goal is to distinguish the samples generated by the generator from the real samples. Through adversarial training, the generator strives to deceive the discriminator so that it cannot accurately discriminate between synthetic and real data, while the discriminator improves its accuracy by learning features of both real and synthetic data.

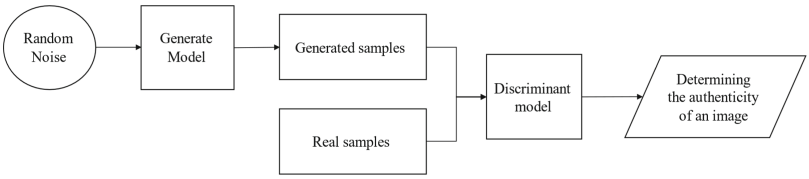

Fig.1. Generating an adversarial network model

The principle is shown in Fig. 1. The generator generates samples by mapping the true samples to a new data space and generating them. The discriminator takes the real sample and the generated sample as input and outputs a probability value indicating the probability of the real sample. The generator gradually improves the quality of the generated samples based on the feedback from the discriminator and eventually reaches a "Nash equilibrium" state, i.e., the generated samples are indistinguishable from the discriminator. This adversarial process motivates the generator to generate more realistic samples.

Generate the optimization objective function for the adversarial network:

$$\min_{G}\max_{D} V(D,G) = E_{x\sim p_{data}(x)}\big[logD(X)\big] + E_{z\sim p_z(z)}\big[log(1-D(G(z)))\big] \quad (1)$$

In the expression (1), the D represents the discriminator, the G represents the generator, the x represents the real sample, and z represents the random noise, and $P_{data}(x)$ represents the distribution of the real sample, and $P_Z(z)$ denotes the distribution of the generated samples. The discriminator is trained to make its prediction value of the real data $lg\ D(x)$. The discriminator is trained to make its prediction of the real data as large as possible in order to improve the discriminator's ability to determine true and false. The generator is trained so that the generated samples are discriminated as true, even if they are $1 - D[G(z)]$ as small as possible, in order to achieve the effect of falsehood.

2.2 WGAN

WGAN7 is an improved method based on GAN proposed in 2017, with some improvements mainly in the original GAN to generate stable and high-quality images:

1) remove the sigmoid from the last layer of the discriminator, which is a regression problem rather than a classification problem since the discriminator is to fit the Wasserstein distance, so the values are not limited to 0–1;
2) the loss of generators and discriminators does not take log;
3) truncate the absolute values of the discriminator parameters to no more than a fixed constant c after each update of them;
4) Use non-momentum-based optimization algorithms.

WGAN introduces Wasserstein distance [9] instead of JS distance, it avoids the problem of gradient disappearance to some extent. Wasserstein distance has good continuity and convex property, it can quantify the distance between the real data distribution and the generator output distribution, and when the two distributions overlap, Wasserstein distance can still give a meaningful value. Compared with traditional metrics, Wasserstein distance can provide smoother gradients, making the training more stable. The loss functions of the generator and discriminator in WGAN are defined as follows:

Loss functions of generators:

$$L_G = -\frac{1}{N}\sum_{i=1}^{N} D(G(z^{(i)})) \tag{2}$$

Loss function of the discriminator:

$$L_D = \frac{1}{N}\sum_{i=1}^{N} D(G(z^{(i)})) - \frac{1}{N}\sum_{i=1}^{N} D(x^{(i)}) \tag{3}$$

where $D(G(z^{(i)}))$ denotes the degree to which the generated samples are discriminated as true samples by the discriminator. $D(x^{(i)})$ that is, the degree to which the true samples are discriminated as true samples by the discriminator. The loss function of the generator is the opposite of the mean of the discriminator's output of the generated samples, while the loss function of the discriminator is the difference between the mean of the output of the true samples and the generated samples.

3 Methods

3.1 Residual Network

Traditional deep neural networks suffer from the problem of gradient disappearance when the number of network layers increases, i.e., the gradient gradually becomes smaller with backpropagation leading to training difficulties [10]. This problem is solved by introducing residual connections. Residual connections build a "shortcut" by directly skipping one or more layers of network output and adding the skipped portion to the subsequent network output, allowing the gradient to propagate more easily. This structure allows the network to learn the difference between the input and the output, rather than learning the entire mapping directly.

An important variant of residual networks is the residual block, which is a module consisting of multiple residual connections with the same dimensionality. Each residual block is usually composed of two convolutional layers and one residual connection inside. By stacking multiple residual blocks, a deeper network structure can be constructed. The basic structure of a residual network is shown in Fig. 2.

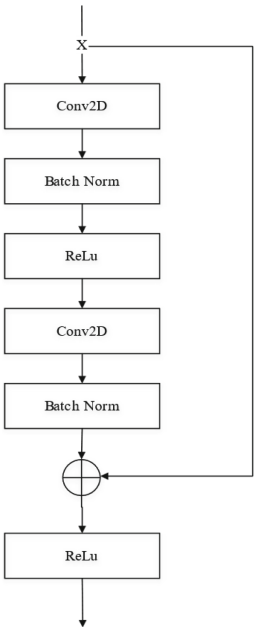

Fig. 2. Residual structure diagram

3.2 Self-attention Mechanism

Traditional GAN models for generating high-resolution images are usually generated by learning texture features from low-resolution images with fixed spatially localized points. This approach makes it relatively easy to learn texture features, such as subtle texture and color variations. However, traditional GAN models have difficulties in capturing specific structural and geometric features in images. This is because traditional GAN models do not explicitly model or constrain the structural and geometric features in the image during generation, but only focus on how to generate the whole image by learning the relationship between local pixels [11]. Therefore, when generating high-resolution images, these models may produce some results that do not match the structure and geometry of the real image.

To solve this problem, the structural and geometric features in the images are better captured by introducing a self-attention mechanism. Such a model is able to better generate high-resolution images and more accurately represent the structure and geometry of the image while maintaining the texture features.

Self-attention mechanism [12] is a mechanism to calculate the relationship between each element in a sequence and other elements. In the self-attention mechanism, each element in the input sequence calculates a relationship score with all other elements in the sequence, and then a weighted average is applied to all elements based on the score to obtain a contextual representation of each element. The specific calculation process is:

1) Suppose the input feature map is $X \in R^{C \times N}$, where C denotes the number of channels and N denotes the length of the sequence. After two 1×1 convolutions for linear transformation and channel compression, the two tensors are converted to matrix form, transposed, and then do the multiplication operation to get S_{ij} The attention map $\beta_{j,i}$ is normalized by Softmax**,** as shown in Eq. (4):

$$\beta_{j,i} = \frac{exp(S_{ij})}{\sum_{i=1}^{N} exp(S_{ij})} \tag{4}$$

2) Combine the attentional map $\beta_{j,i}$ with the linearly transformed original feature map $h(X_i)$ point by point, and obtain the self-attention feature map o_j, as shown in Eq. (5):

$$o_j = \sum_{i=1}^{N} \beta_{j,i} h(X_i) \tag{5}$$

3) Finally, the self-attention feature mapping and the original feature mapping are weighted and summed up as the final output, as shown in Eq. (6):

$$y_i = ro_i + X_i \tag{6}$$

r is a transition parameter to control the assignment of weights, which can be interpreted as a scaling factor with an initial value of 0.

3.3 Spectral Parametric Normalization

Traditional weight normalization methods, such as weight decay or batch normalization, can prevent problems such as overfitting or gradient disappearance to some extent, but they do not directly limit the spectral parametric number of the weight matrix. And the spectral parametric normalization [13]. The core idea is to ensure that the parameters of each layer of the network satisfy the condition that Lipschitz is equal to 1 by imposing a Lipschitz constraint on the discriminator parameters. This constraint is achieved by dividing the weight matrix of each layer by the spectral norm of the weight matrix of that layer, i.e., by applying Eq. (7) to the weights W

$$W_{normalized} = \frac{W}{\sigma} \tag{7}$$

where σ denotes the two-parametric number of weights W. By spectrally normalizing the weights of each layer of the discriminator, the discriminator can be considered as a function mapping and its Lipschitz constraint to be less than 1 to improve the stability of model training.

4 Network Structure

In this section, a residual module with embedded spectral parametric normalization is designed to replace some network layers of the discriminator and generator, and a self-attention mechanism is introduced in the intermediate layer of the generator to alleviate the problems of pattern collapse and poor quality and diversity of the generated images that occur during the training process. The specific flow of the generative adversarial network model is shown in Fig. 3.

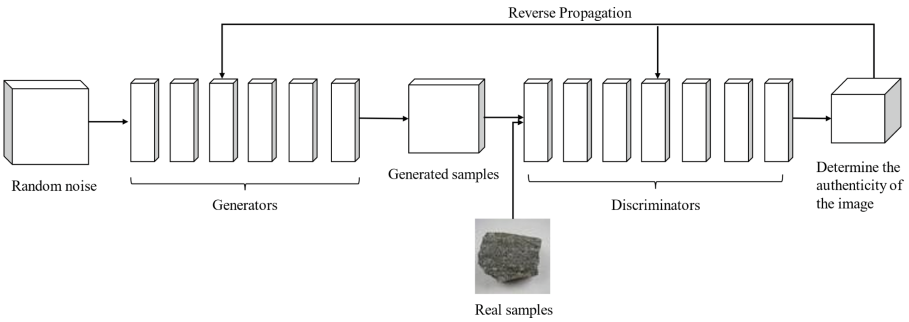

Fig. 3. Overall structure diagram

4.1 Generator Network

First, the input 100-dimensional noise is converted into a tensor with a size of 64 × 2 × 2 by a linear layer. Then, this tensor is rearranged into a 64-channel, 2 × 2-sized feature map. Next, the feature map is gradually increased in size through a series of layers and blocks, including BasicBlock, UpsamplingNearest2d, and self-attention, while the features are processed and corrected. These layers and blocks are divided into 4 stages (layer1–layer4), where layer2–laye4 have the same network layer structure. Finally, the feature map is converted into the final generated image through a series of layers, including Conv2d, Tanh, etc., and the generated image is made flatter in terms of tonal distribution, as shown in Fig. 4.

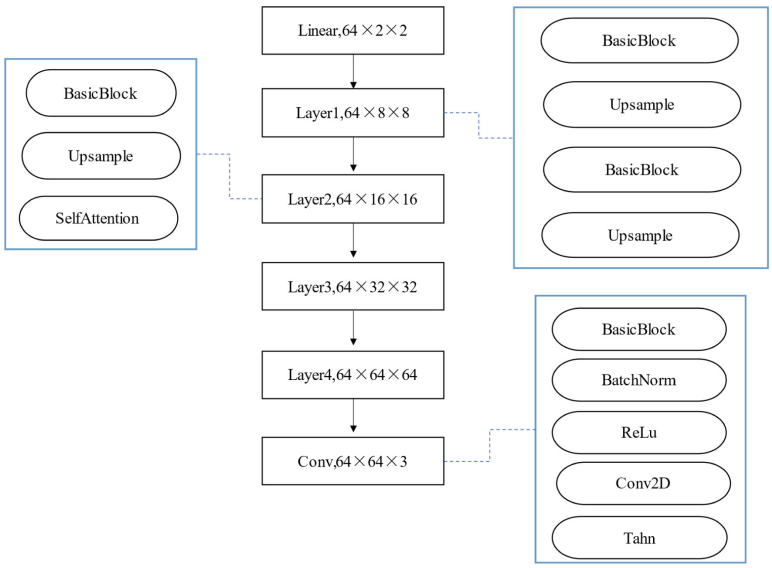

Fig. 4. Generator network structure diagram

4.2 Discriminator Network

The discriminator network has seven network layers. First, a 64 × 64 RGB image is used as the input of the discriminator network, and a convolution operation is performed by a Conv2d layer with a convolution kernel of 3 × 3 size to transform the number of channels of the input image from 3 to 64 and keep the size of the input image constant. Next, a series of layers and blocks, including BasicBlock, AvgPool2d, BatchNorm2d, and ReLU, are passed to gradually extract the features of the image. These layers and blocks are divided into 5 stages (layer1–layer5), where layer2–layer4 have the same network layer structure, and the feature map size of each stage is halved by AvgPool2d. Finally, a single value is an output after a fully connected layer, which represents the probability that the input image is a real image, and finally, the output value is converted to a probability value between 0 and 1 by a Sigmoid function. The specific structure is shown in Fig. 5.

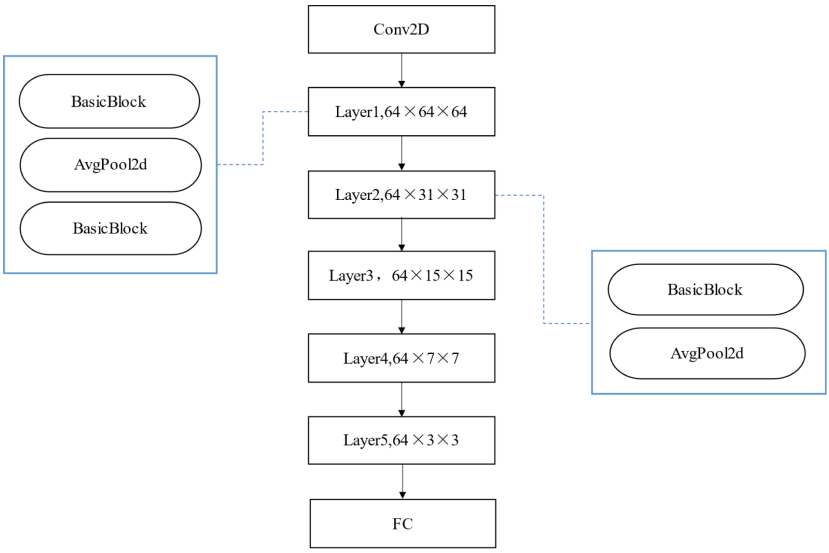

Fig. 5. Discriminator network structure diagram

5 Experiment and Analysis

5.1 Experimental Environment and Data Set

In order to test the feasibility of the proposed method, the experiments were run under the Windows operating system based on the PyTorch deep learning framework, with Python version 3.7, and the GPU used was RTX3060. The dataset used for the experiment is the ore images obtained by this project group. 1207 concentrate images and 664 waste images are selected from the dataset, and the size of the original data sample is 6000 × 4000, which is scaled equally to 64 × 64 in order to save computational resources, as shown in Fig. 6.

Fig. 6. Concentrate and waste ore images

5.2 Experimental Setup

The learning rate of both the generator and discriminator is 0.00005, the batch size is 32, the size of the generated image is $64 \times 64 \times 3$, and the training is stopped after 10,000 epochs of iterations. The FID metric is used to evaluate the quality of the generated images, and the recognition accuracy is used as a further evaluation metric on MobileNet V2. Each class is expanded with 2000 data sets as the input data of MobileNet V2 network, and the recognition accuracy is obtained after 500 epochs of training.

5.3 Experimental Results

Generating Images for Display. Figure 7 shows the generated images of the original WGAN, DCGAN, and the improved WGAN methods, (a) is the real image, (b) is the image generated by DCGAN, (c) is the image generated by the original WGAN, and (d) is the image generated by the improved WGAN. From the figure, it can be seen that the images generated by DCGAN under the same training environment are still blurred and the features are not obvious, and the quality of the images generated by the original WGAN is improved compared with DCGAN, but the edge lines are not very clear yet, while the images generated by the improved WGAN are closer to the real images in terms of surface features and outperform DCGAN and the original WGAN in terms of quality and diversity method.

Fig. 7. Generating images

Generating Image Evaluation Metrics. In this paper, FID (Fréchet Inception Distance [14] FID is a widely used metric for generative adversarial networks to quantify the quality of generated images by comparing the statistical information of the features of the real image distribution and the generated image distribution.

The computational process of FID consists of two key steps: feature extraction and feature statistics. First, the features are extracted from the real and generated images using a pre-trained convolutional neural network. These feature vectors capture the abstract features of the image and map the image to a high-dimensional feature space.

FID is the distribution distance between the real image and the generated image in the feature space [15]. Specifically, it calculates the Fréchet distance between the mean and covariance matrices of the feature vectors of the real and generated images. This distance measures the similarity between the two distributions, and a smaller value indicates that the generated image is closer to the real image. However, the FID value can only be used as an objective criterion to judge the quality problem of the generated images, because when encountering the pattern collapse problem, the generator tends to generate very similar samples, ignoring the diversity in the real data, which may lead to a lower FID value, thus making it easier for the classifier to find patterns for correct classification.

In the experiments of this paper, the two types of images generated by the original WGAN, DCGAN and the improved WGAN are compared. From the table of experimental results, it can be seen that the FID values of the improved WGAN are much lower than those of the original WGAN and DCGAN, which indicates that the improved WGAN outperforms the other methods in terms of quality and diversity of the generated images (Table 1).

Table 1. FID values

Generating method	Concentrate sample FID values	FID value of waste ore sample
DCGAN	148	183
WGAN	144	202
Improved WGAN	86	102

5.4 Comparison Experiments

To demonstrate the effectiveness of the improved WGAN method in generating ore images, the improved WGAN was experimentally compared with the original WGAN, traditional enhancement methods, and DCGAN, and the images generated by these four different data enhancement methods were expanded to the original dataset as the data input for the classification experiments, and the experiments and tests were conducted on the MobileNet V2 classification network, and the experimental results are shown in Table. As shown in the table, where the original data set is 1207 concentrate images and 665 waste images without any data enhancement methods, and this is used as the training set of each method for image enhancement experiments to obtain the recognition accuracy of different methods, as shown in Fig. 8.

The specific values are shown in Table 2. The traditional data enhancement method can improve the accuracy of ore binary classification to a certain extent, but this improvement is limited because using the traditional data enhancement method only changes the position of the ore in the image by rotation, translation, and other operations, and the generated image is relatively single, while the DCGAN method improves the accuracy by about 4% compared with the traditional enhancement method, but in the training process The effect of the original WGAN is better than the above methods, but there is still room for improvement in generating high-quality images.

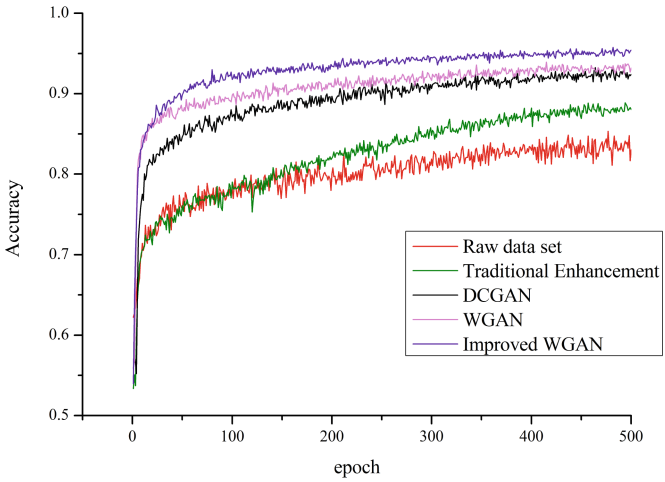

Fig.8. Recognition accuracy of different methods

In contrast, the improved WGAN method is stronger than the above methods. Replacing the original network structure with the residual structure in the original WGAN can improve the network performance and generate images with distinct features, so the recognition rate of ore images generated using the improved WGAN is higher than that of the traditional enhancement method, the original WGAN method and the DCGAN method on the classification network.

Table 2. Comparison of experimental results

Experimental method	MobileNet V2
Raw data set	83.01%
Traditional Enhancement	88.27%
DCGAN	92.38%
WGAN	93.19%
Improved WGAN	95.39%

5.5 Ablation Experiments

To demonstrate that the various improvements of the network are effective in improving the performance of the model, ablation experiments are conducted. In the ablation experiments, the accuracy of the classification network Mobilenet V2 is used as the main indicator, and the ablation experiments include the introduction of residual blocks, the introduction of spectral parametric normalization, and the introduction of a self-attentive mechanism.

To demonstrate that the introduction of residual blocks in the original WGAN can improve the model accuracy, comparative experiments are designed to replace the original network layers of the generator and discriminator with residual blocks; to demonstrate the contribution of spectral parametric normalization to the model, it is embedded on each convolutional layer of the residual blocks; and to demonstrate that the self-attentive mechanism is effective for model improvement, it is embedded in the middle layer of the generator.

The experimental results are shown in the Table 3. It can be seen that after replacing the original network layers of the generator and discriminator with residual blocks, the recognition accuracy of the model is improved by 2% and the phenomenon of gradient disappearance during training is moderated; after introducing spectral parametric normalization on each convolutional layer of the residual blocks, the stability of the model training is improved and the recognition accuracy of the model is also improved by 0.5%; after embedding the self-attention mechanism in the second, third, fourth and fifth layers of the generator After embedding the self-attention mechanism, the recognition accuracy of the model is improved by 1.9%, and the generated images are of higher quality and richer in features.

Overall, by introducing residual blocks, spectral parametric normalization, and self-attentiveness mechanisms, the algorithm proposed in this study is applied to ore classification recognition with 2.2% higher accuracy than the pre-improvement model and improves the quality and diversity of the generated images.

Table 3. Results of ablation experiments

Experimental method	MobileNet V2
WGAN	93.19%
WGAN + spectral parametric normalization	93.67%
WGAN + Self-Attention Mechanism	94.94%
WGAN + Residuals Module	95.08%
Improved WGAN	95.39%

6 Conclusion

In this paper, we propose a generative adversarial network model incorporating residual blocks, spectral parametric normalization, and self-attentive mechanism. Introducing residual blocks into the network can extract deeper features and generate better images, applying the weight normalization technique of spectral parametric normalization to each convolutional layer of the residual blocks can slow down the convergence of the discriminator more effectively, stabilize the training process, and ensure that the generator is fully trained training process and ensure that the generator is fully trained, and introducing the self-attention mechanism into the generator is more attentive to the structural and geometric features in the image and can generate images with richer diversity.

The images generated by the improved WGAN method were used to expand the sample dataset, and the newly generated images were tested by FID to show that the stability of the training process and the generation quality of the method were improved. Compared with the original WGAN method, the accuracy of the images generated by the improved WGAN method for classification and recognition is also improved. However, the method still faces some challenges in generating some other high-resolution images, which will be improved and optimized in the future.

References

1. Wang, Z.Q., Yu, X., Yang, X.J., et al.: Tomato leaf disease identification based on WGAN and MCA-MobileNet. J. Agric. Mach. **54**(05), 244–252 (2023)
2. Guo, M., Yang, Q., Zhao, L.: Image generation based on conditional Wassertein generative adversarial network. Comput. Appl. **41**(05), 1432–1437 (2021)
3. Heusel, M., Ramsauer, H., Unterthiner, T., et al.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. In: Proceedings of the Advances in Neural Information Processing Systems, pp. 6626–6637. Curran Associates, Long Beach (2017)
4. Wu, S., Li, X.: A review of research progress in generative adversarial networks. Comput. Sci. Explor. **14**(03), 377–388 (2020)
5. Li, Q.-L., Ma, L.: A study of DCGAN combining spectrum normalization and self-attentive mechanism. Comput. Appl. Softw. **38**(02), 227–232+290 (2021)
6. Liu, J., Ma, S.H.: Data enhancement of colon cancer glandular cells based on feature map gradient connection generative adversarial network. Comput. Digital Eng. **50**(11), 2557–2561+2573 (2022)
7. Liang, J., Wei, S.J., Jiang, C.F.: A review on generative adversarial networks GAN. Comput. Sci. Explor. **14**(01), 1–17 (2020)
8. Gulrajani, I., Ahmed, F., Arjovsky, M., et al.: Improved training of wasserstein GANs. arXiv: 1704.00028 (2017)
9. Wang, Y., Han, J., Fan, L.: Research on speech enhancement algorithm based on WGAN[J]. J. Chongqing Univ. Posts Telecommun. (Nat. Sci. Ed.) **31**(01), 136–142 (2019)
10. Guo, Y., et al.: A review of residual network research. Comput. Appl. Res. **37**(05) (2020)
11. Wang, X.L., Girshick, R., Gupta, A., et al.: Non-local neural networks. Comput. Vision Pattern Recogn. **2**, 7794–7780 (2017)
12. Zhu, Z., Rao, Y., Wu, Y., et al.: Research progress of attentional mechanism in deep learning. Chin. J. Inf. **33**(06), 1–11 (2019)
13. Miyato, T., Kataoka, T., Koyama, M., et al.: Spectral normalization for generative adversarial networks. In: International Conference on Learning Representations (2018)
14. Hu, L.-M., Zhang, Y.: Short-wave infrared-visible face image translation based on generative adversarial networks. J. Opt. **40**(5), 75–84 (2020)
15. Chen, X., Huang, X., Xie, L.: A multiscale conditional generation adversarial network-based approach for blood cell image classification and detection. J. Zhejiang Univ. (Eng. Ed.) **55**(09), 1772–1781 (2021)