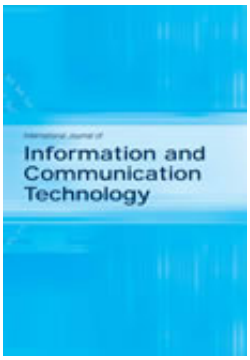
INDERSCIENCE PUBLISHERS
Linking academia, business and industry through research



International Journal of Information and Communication Technology

ISSN online: 1741-8070 - ISSN print: 1466-6642
https://www.inderscience.com/ijict

Adversarial learning-based image generation algorithm for AI art creation

Zhou Chen, Zaixi Xia

Article History:

Received:	08 December 2024
Last revised:	17 December 2024
Accepted:	18 December 2024
Published online:	25 February 2025

Adversarial learning-based image generation algorithm for AI art creation

Zhou Chen*

Chengdu Academy of Fine Arts,
Sichuan Conservatory of Music,
Chengdu, 610021, China
Email: chenzhou197511@sina.com
*Corresponding author

Zaixi Xia

Digital Media Arts College,
Chengdu Vocational University of the Arts,
Chengdu, 611433, China
Email: xiazaixi123@gmail.com

Abstract: Generative adversarial network (GAN) is widely used in AI art creation image generation tasks. Focusing on the issue of poor quality of images generated by existing algorithms, this paper firstly optimises the GAN by minimising mutual information (MIGAN), captures the features of the image by adopting multiscale null convolution, and employs the spatial feature transformation to ensure the completeness of semantic information of the image. Then self-attention and channel attention are introduced in MIGAN to focus on the important features of the image from both spatial and channel dimensions to finally get the generated image. Adversarial learning between the generated image and the real image through the discriminator, and designing the adversarial loss function to optimise the network model, which effectively improves the image generation effect. Finally, the algorithm is experimentally verified to produce high-quality AI art creation images with at least 33% reduction.

Keywords: image generation; generative adversarial network; GAN; multiscale null convolution; self-attention mechanism; channel attention.

Reference to this paper should be made as follows: Chen, Z. and Xia, Z. (2025) 'Adversarial learning-based image generation algorithm for AI art creation', *Int. J. Information and Communication Technology*, Vol. 26, No. 4, pp.57–71.

Biographical notes: Zhou Chen received his Doctorate degree in Fine Arts from the Central Academy of Fine Arts in 2018. Currently, he teaches in the Chengdu Academy of Fine Arts, Sichuan Conservatory of Music. His research interests include experimental art, practice of contemporary transformation of traditional painting language.

Zaixi Xia received her Master's degree in Art from Southwest Minzu University in 2015. She works in Digital Media Arts College, Chengdu Vocational University of the Arts. Her research interests include digital media art, innovative transformation of ethnic minority visual arts in Southwest China.

1 Introduction

As the artificial intelligence (AI) with modern society deeply integrating, more and more artists are using AI as a tool and vehicle for creation. AI art creation has a wide range of applications in a number of fields, such as computer vision (CV), Natural Language Processing (NLP), and so on (Santos et al., 2021). In the process of traditional painting, a good painter will copy a large number of excellent masters in the process of learning, and incorporate the ideas and styles of different writers into his works, so as to complete a very personal artistic painting. Drawing on this idea, image-generation techniques have been used to fuse the characteristics of paintings from different styles and authors to produce images of art that are as authentic as they are real (Enjellina and Rossy, 2023). On the one hand, it can enrich the image data of art paintings and generate more art images; on the other hand, it can inspire painters to create more works. Therefore, generating high-quality images can not only help humans and computers interact better, but also promote the development of art creation in China (Zhao et al., 2023).

Early models of image generation for art creation were autoregressive models, variational autocoders (VAE), and stream-based models (Cetinic and She, 2022). Dong et al. (2013) predicted the values of the pixel points in a sequential manner, and the distribution of the generated image can be obtained by multiplying the probabilities of all the pixel points by using the previous generated pixel points as a reference. Xu et al. (2019) achieved image generation by reconstructing the input image to achieve hidden variable distribution to real data distribution transformation and based on maximising the minimum likelihood of the data. Suryadevara (2020) achieved input and latent space interconversion by finding reversible bijections with tractable sampling and latent variable reasoning, which has an advantage in the generation of images for artistic creations with advantages on it.

Due to the ambiguity of images generated by traditional generative algorithms, Generative Adversarial Networks (GANs), as an adversarial learning-based image generation model, enable the generator to continuously optimise its generation strategy through adversarial learning of the generator and the discriminator until it is able to produce realistic data sufficient to deceive the discriminator. This mechanism ensures the superior performance of GANs in image generation. Based on these advantages, scholars have utilised GAN for image generation. Huang and Jafari (2023) used Wasserstein distance to replace the original JS dispersion, but due to the Lipschitz continuity constraints, the effect of the generated image is general. Qi et al. (2021) proposed a stacked network-based GAN that uses multiple stages for initial image generation and refinement of latent variables, but does not focus on the fine-grained content of the image, leaving the details of the image still blurred. Alruily et al. (2023) proposed an image generation method based on U-Net and GAN that uses swish layers and swish-gated residual blocks to filter the information transmitted by the layers to generate more realistic images. Andreini et al. (2020) implemented discriminators and generators using convolutional neural networks (CNNs) and proposed an unsupervised deep convolutional generative adversarial network (DCGAN), which improves the effectiveness of the generated images through the powerful feature extraction capability of the convolutional layers. Wang and Ma (2023) introduced residual structure in GAN algorithm to extract both rich image features and other redundant information.

Attentional mechanism (AM) can filter important information and capture long-distance dependencies in the input sequence, which is important for improving the

quality of images. Mi et al. (2020) proposed attention GAN, which introduces AM to improve the fine-grained details of different sub-regions in the generated image to produce higher quality images. Zhang et al. (2022) proposed an adaptive attention GAN to adaptively align the layout structure of the generated image with the visual features of its corresponding real image through semantic similarity loss, which improves the realism of the generated image. Shen et al. (2023) proposed a new spatial attention GAN (SPA-GAN) that computes attention in its discriminator and uses it to help the generator focus more on the most discriminative region between the original and target domains, thus improving image generation.

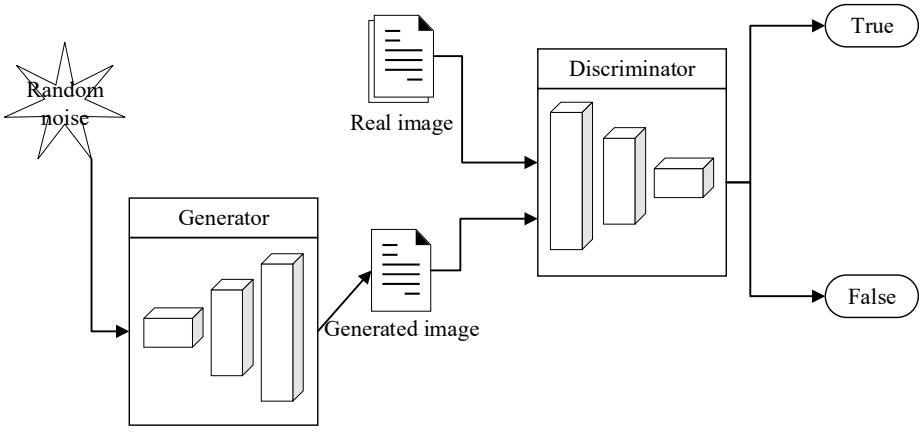
Overall, GAN has a good application prospect in the field of art image generation, but it does not consider multi-scale feature extraction when generating images, which leads to the quality of the generated images to be improved. In order to solve the above problems, this paper optimises GAN using minimised mutual information (MMI) and applies the improved GAN to AI art creation image generation. Firstly, the traditional GAN is optimised by minimising mutual information (MIGAN), and the multiscale structural similarity index (MS-SSIM) is introduced to improve the effect of feature learning. Secondly, the multi-scale information of the image is extracted using multi-scale cavity convolution to generate images with richer texture details. Based on this, spatial feature transformation (SFT) is used to ensure the integrity of the semantic information of the image. Then self-attention (SA) and channel attention are introduced in MIGAN to adjust the output features from both spatial and channel dimensions, so that the generator focuses on the important features of the image, and finally the generated image is obtained. The discriminator is adopted to learn the generated image against the real image, and the network model is optimised by designing the loss function, which effectively improves the quality of the generated image. The experimental outcome implies that the offered algorithm is superior to the comparison algorithm in Fréchet inception distance (FID), inception score (IS) and other indicators, which proves that the algorithm has excellent image generation effect.

2 Relevant theoretical foundations

2.1 Generating adversarial network

GAN is a deep learning model that generates new, similar data to the training data by training two neural networks against each other. GANs are more suitable for image generation than CNNs, RNNs, and BPs. CNNs are good at image classification and feature extraction, but their generative ability is relatively weak. RNNs may need to be converted to sequence form when processing image data, which may increase the processing difficulty and computational cost. BP neural networks are usually used for classification and regression tasks, rather than generative tasks. Although it is theoretically possible to generate images by training a BP neural network, it is not as direct and effective as a GAN. GAN is a generation algorithm based on adversarial learning (Wang et al., 2017), which is composed of generator (G) and discriminator (D). Its basic idea is derived from zero-sum game in game theory. In the form of minimax game, G minimises D's ability to generate images correctly. D, on the other hand, maximises this capability and ultimately produces a clearer and more realistic image, as shown in Figure 1.

Figure 1 Generating adversarial network



Taking image generation as an example, the goal of G is to transform the input random noise z, which obeys the normal distribution $N(0, 1)$, into a new data image $G(z)$, which is as close as possible to the distribution of the real image in the feature space, fooling D so that the complex D cannot distinguish whether the image is generated from G or from the real image in the training set. D takes samples as input and predicts the truth of these samples to classify them as accurately as possible and to determine whether they are real images or generated images. In mathematical terms, the goal of GAN generation is as follows.

$$\min_{G}\max_{D} V(G,D) = E_{x\sim P_r}\left[\log D(x)\right] + E_{z\sim P_z}\left[\log\left(1-D\left(G(z)\right)\right)\right] \tag{1}$$

where x represents the real image, z represents the random noise that follows a normal distribution, P_r represents the data distribution of the real image, and $G(z)$ represents the image generated after G.

In terms of D, if the input image comes from is a real dataset then the network output will be maximised and conversely if the input image comes from a fake image generated by the generative network then the network output value will be minimised. Thus the loss function for D is as follows.

$$L_D = -E_{x\sim P_r}\left[\log D(x)\right] - E_{z\sim P_z}\left[\log\left(1-D\left(G(z)\right)\right)\right] \tag{2}$$

2.2 Attention mechanism

AM is inspired by the human attentional way of thinking and is able to automatically learn and selectively focus on certain parts of the input data while ignoring the unimportant parts in order to improve the efficiency and accuracy of processing the input data. Its essence is to analyse and weight the importance of different parts of the input data to determine which information is more useful for downstream tasks (Lu and Dosher). The main role of the attention mechanism in the field of image generation is to help the model to pay better attention to the critical parts of the input information, thus generating higher quality images. This mechanism reduces the computational overhead of the model in processing irrelevant information. This helps to increase the speed of image

generation and allows the model to adapt faster to new datasets and tasks. AM in CV distinguishes the importance of different parts of the input image in order to improve the model's representational and predictive performance for a given task.

AM consists of two computational processes: one is to compute the distribution of attention on the input features, i.e., the features compute the weights; the other is to weight the input features to combine them according to the obtained attention weights. In the image generation task, the input feature vector $X = [x_1,\ldots,x_N]$ is adopted to find its correlation information and q is used to form the representation vector for the task. Given q and X, the probability α of selecting the i^{th} feature vector is computed and used as weights in the weighted sum. Calculate the probability α_n of selecting the nth vector as follows.

$$\alpha_n = \frac{\exp\left(s\left(x_n, q\right)\right)}{\sum_{j=1}^{N} \exp\left(s\left(x_j, q\right)\right)} \tag{3}$$

where $s(x, q)$ is the correlation calculation function.

The attentional weight distribution is computed for all feature vectors after having a representation vector q for the task in question. They are then summarised using an information aggregation selection mechanism as follows.

$$attn(X, q) = \sum_{n=1}^{N} a_n x_n \tag{4}$$

3 Optimisation of GAN based on minimising mutual information

Traditional GAN models generate a realistic image by inputting noise vectors into G. However, G is usually highly coupled with input noise vectors z, which makes the relationship between features very complex and difficult to interpret. Therefore, there is a need to better control the generated results by modifying a particular dimension in z. To this end, this paper improves the objective function of GAN (MIGAN) to learn interpretable feature representations. MIGAN decomposes z into latent variables and noise distributions and learns these interpretable latent variables by MMI (Fan et al., 2023) and introduces multi-scale structural similarity index (MS-SSIM) (Chen and Bovik, 2011) as a mutual information factor to retain the features of the images to improve the feature learning.

The original noise input is divided into two parts, one is the noise vector z', and the other is the hidden variables c. These hidden variables c will have a prior probability distribution indicating the different feature dimensions of the generated data, generally denoted by c_1, $c_2,\ldots,c_L$ for the hidden variables c. Assuming that the dimensions are independent of each other, then the following equation exists.

$$p\left(c_1, c_2, \cdots, c_L\right) = \prod_{i=1}^{L} p\left(c_i\right) \tag{5}$$

Input z' and c in G and output $G(z', c)$. But in a GAN, G will find a solution that makes $P_G(x \mid c) = P_G(x)$ such that the hidden variable c does not work at all. Therefore, inspired from information theory, the regularisation method of MMI is used for GAN optimisation. There is a high degree of mutual information between the outputs of the hidden variable c and G. Mutual information $I(X; Y)$ is the amount of information about X that can be obtained if Y is known, and is given by the following formula.

$$I(X;Y) = H(X) - H(X \mid Y) = H(Y) - H(Y \mid X) \tag{6}$$

where H is the computational entropy, $H(X \mid Y)$ is used to measure the uncertainty of X when Y is known, and $I(X; Y)$ is maximised when X and Y are explicitly related. Therefore, it is necessary to make $I(c, G(z, c))$ as large as possible. At this point, the objective function of the GAN can be rewritten as follows, where λ is the mutual information factor.

$$\min_G \max_D V_I(D,G) = V(D,G) - \lambda I\left(c; G\left(z', c\right)\right) \tag{7}$$

To improve the feature representation of the image, MS-SSIM is introduced to calculate the pixel-level similarity between any image $x \in X$ and its paired image $y \in Y$ in a single learning process, which reflects the overall features of the picture as much as possible, and MS-SSIM is adopted as the mutual information factor with the following formula, where C is a constant term.

$$\lambda = -\left(MS - SSIM\left(F_{X \to Y}(x), y\right) + C\right) \tag{8}$$

The MS-SSIM loss matches the luminance (l), contrast (c), and structural (s) information of the generated image with that of the input image, which is helpful in improving the quality of the generated image. S-SSIM Loss Consider the SSIM loss over the entire scale as follows.

$$MS - SSIM\left(F_{X \to Y}(x), y\right) = \left[l_M(x,y)\right]^{\alpha_M} \prod_{j=1}^{M} \left[c_j(x,y)\right]^{\beta_j} \left[s_j(x,y)\right]^{\gamma_j} \tag{9}$$

where $l(x,y) = (2\mu_x\mu_y + c_1)/(\mu_x^2 + \mu_y^2 + c_1)$, $c(x,y) = (2\sigma_x\sigma_y + c_1)/(\sigma_x^2 + \sigma_y^2 + c_1)$, $s(x, y) =$, $(\sigma_{xy} + c_3)/(\sigma_x\sigma_y + c_3)$, μ_x, σ_x^2, and σ_{xy} are the mean of x, the variance of x, and the covariance of xy, respectively.

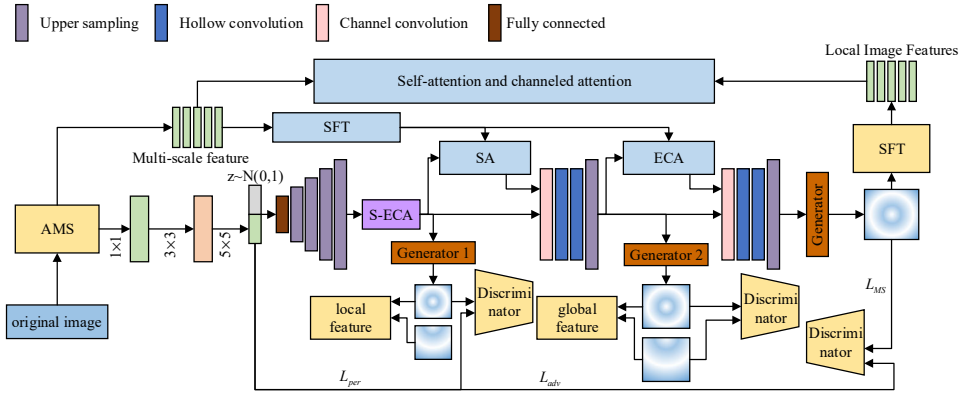
4 Adversarial learning-based image generation algorithm for AI art creation

4.1 Feature extraction based on multi-scale cavity convolution

For the goal of solve the existing image generation algorithms ignoring the problem of semantic integrity and visual authenticity, AI art creation image generation algorithm is designed based on MIGAN, as shown in Figure 2. The multiscale null convolution (AMS) is first utilised to extract the multiscale special this part of the image, while SFT is employed to ensure the integrity of the semantic information. Then S-ECA is proposed

by combining SA and efficient channel attention (ECA) mechanism to adjust the output features from both spatial and channel dimensions so that the generator focuses on the important features of the image. The quality of image generation is effectively improved by discriminator to discriminate the generated image from the real image and designing the adversarial loss function to optimise the network model.

Figure 2 MIGAN-based image generation for AI art creation (see online version for colours)



Aiming at the issue of relatively few texture features in AI art creation images, relying on AMS extracts the texture features of images. Different from the traditional pooling operation, AMS can extract multi-scale features without reducing the image resolution, provide more complete feature information for subsequent image generation, and improve the generation quality of AI art creation images.

AMS first extracts the feature maps at different scales, and then stitches them together to obtain a feature map with multiple different receptive fields. The first part is the original characteristic dimension F_{in} with receptive field 1 × 1. The second part is the characteristic map $F_{3\times3}$ obtained by ordinary convolution with convolution kernel size 3 × 3 with receptive field 3 × 3. The third part is the characteristic map $F_{5\times5}$ obtained by combining two convolution kernels with ordinary convolution of size 5 × 5 with receptive field 5 × 5. After inputting the characteristic map to AMS, three combinations of characteristic maps with different receptive field sizes are obtained to obtain complete background information and improve the quality of image generation. Let the output characteristic map be F_{out} as follows.

$$F_{out} = Concat\left(F_{in}, F_{3\times3}, F_{5\times5}\right) \tag{10}$$

4.2 SFT for AI art creation images

After obtaining the output feature map F_{out}, the SFT is introduced to learn the semantic features of F_{out}. The normalised modulation parameters are obtained to achieve the spatial transformation of each layer of features to better learn the semantic layout features of the image in the target viewpoint and to better guide G to generate realistic images. The SFT structure is shown in Figure 3. The inputs to the module include F_{out} and input feature f.

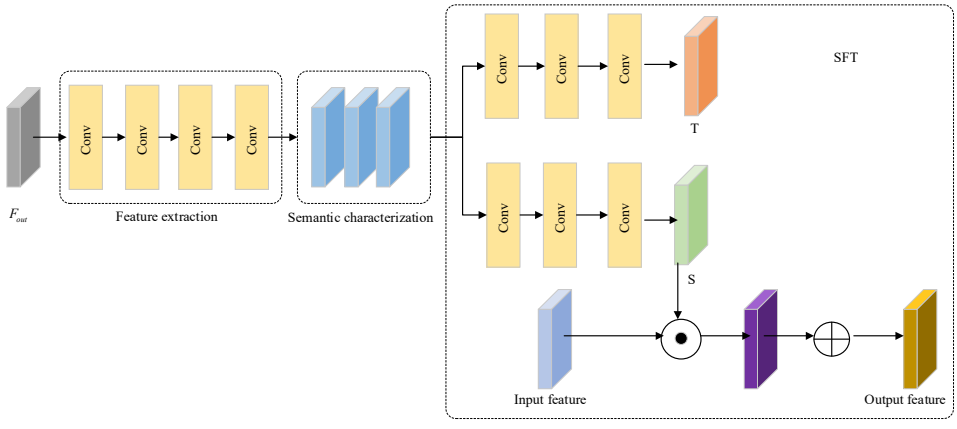
In the SFT module, two-way convolutional learning is first performed on the public semantic features to obtain the scaling feature matrix S and the translation feature matrix

T. Next, the affine transformation of f is performed, and the dot product operation is performed with S. The result of the dot product is then added with T to obtain the final representation F with semantic features at the same scale. Let the whole public semantic feature be K. The operation w, on K, yields S and T, with input feature F_{out} and output feature F. The corresponding transformation relation is as follows.

$$\begin{cases} (S,T) = \omega(K) \\ F = S \odot F_{out} + T \end{cases} \tag{11}$$

After the image is transformed by SFT, the scene object information in the original viewpoint and the deep semantic information in the target viewpoint are extracted, which lays the foundation for the subsequent generation of high-quality AI art creation images.

Figure 3 The SFT structure (see online version for colours)



4.3 GAN-based image generation for AI art creation

To improve the image generation effect, this paper designs S-ECA module in G of GAN to enhance the details of texture, colour and other features of the AI art creation image to improve the generalisation and robustness of the algorithm. The dual-scale D design is used to synthesise the global and local features of the image, and the G parameters are adjusted to obtain a higher quality image.

1 Generator design. A multi-scale convolutional kernel is used in the generator, and LRelu is used as the activation function, removing the original BN layer to reduce the amount of model computation. To address the issue that the traditional G is limited by the local receptive fields and cannot extract the global information well, the S-ECA module is designed to make the G focus on the important features from the space and channel.

   The SFT-processed feature graph F is partitioned into three spatial regions, each of which corresponds to the query, key and value in SA, respectively. For each query vector, a set of attention scores $s_{i,j}$ is obtained by dot-multiplication with all other key vectors as follows.

$$s_{i,j} = f\left(x_i\right)^T g\left(x_j\right) \tag{12}$$

where $f(x)$ and $g(x)$ are the feature spaces of keys and values respectively. $s_{i,j}$ is normalised by softmax function as follows.

$$\eta_{i,j} = \frac{\exp\left(s_{i,j}\right)}{\sum_{i=1}^{N} \exp\left(s_{i,j}\right)} \tag{13}$$

The normalised attention score is multiplied with the corresponding value vector to obtain the weighted value vector o_j.

$$o_j = \sum_{k=1}^{N} \eta_{j,i} h\left(x_k\right) \tag{14}$$

where $h(x)$ is the vector corresponding to the keys and values, and a new self-attentive characteristic map $F_{SA} = \mu o_i + x_i$ is obtained by hyperparameter tuning, where μ is the learnable parameter.

F_{SA} is input to the ECA module to obtain the weights of the neighbouring channel information by employing a local cross-channel approach.

$$w_i = \sigma\left(\sum_{j=1}^{n} w_i^j F_{SAi}^j\right) \tag{15}$$

where $F_{SAi}^j \in \Omega_i^n$, F_{SAi}^j are the features of channel i and the nearest neighbour channel j, w_i^j is the weight of the inter-channel features, Ω is the set of channels adjacent to F_{SA}, and n is the number of channels.

Subsequently, a global average pooling operation is performed to convert the features of each channel into a scalar value, a convolution operation is performed on the pooled result, and a weight is generated for each channel. Information sharing between channels is realised.

$$w = \sigma\left(CID_n\left(F_{SA}\right)\right) \tag{16}$$

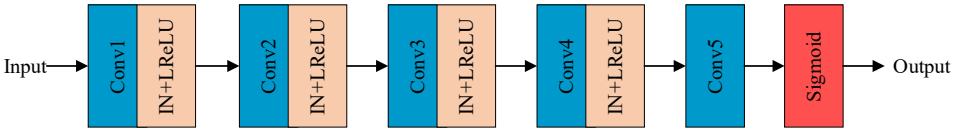
where σ is the activation function and $CID_n(\ )$ is the channel convolution operation.

By using SA to establish global dependencies and then using ECA to optimise long-distance dependencies, we combine dependency modelling in both spatial and channel dimensions. Not only can we understand the dependencies between different regions in the image, but we can also utilise the dependencies between channels to enhance the representation of detailed features, thus improving the effect of image generation.

2 Discriminator design: D enhances its discrimination ability by comparing the difference between the generated image and the real image. To improve the image generation effect, this paper considers the global and local features of the image, adopts multiscale D design to realise the global discrimination of the image or the

local discrimination of pixels, which greatly improves the generalisation ability of the model, and pays attention to the texture details of the image while taking into account the overall content of the image. The structure is shown in Figure 4.

Figure 4 The structure of the discriminator (see online version for colours)



3 Loss function improvement. Due to the large spatial span and wide coverage of AI art creation images, the model training process needs to take into account the quality of the image and the generation of complex backgrounds, and needs to improve the composition of the loss function, in this paper, we take the MS-SSIM loss L_{MS}, the antagonistic loss L_{adv}, and the perceptual loss L_{per} as the G loss function.

$$L_G = \lambda L_{MS} + L_{adv} + \eta L_{per} \tag{17}$$

where λ is the mutual information factor and η is the equilibrium factor.

As can be seen from Section 3, used to measure the content consistency between the generated image and the original image, in order to limit the effect of high frequency noise and to ensure the integrity of the image details, L_{MS} is quantised and calculated to obtain equation (18).

$$L_{MS} = \frac{1}{N}\sum_{n=1}^{N}\left[\frac{1}{HWC}\left\|G_A\left(G_B(X)\right)-X_1+\frac{1}{HWC}G_A\left(G_B(X)\right)-X_1\right.\right] \tag{18}$$

where H, W, C are the image height, width and number of channels respectively; G_A and G_B are the generated image and the real image respectively.

The gradient penalty optimisation L_{adv} is used to penalise D of the input gradient to improve the quality of the generated samples and the convergence speed. The antidumping loss is as follows.

$$L_{adv} = -E_{x\sim P_{data}}\left[D(x)\right] \tag{19}$$

Perceptual loss is calculated between convolutional layers to capture high-level feature differences between images in order to make the high-resolution effect of the generated images more compatible with human vision.

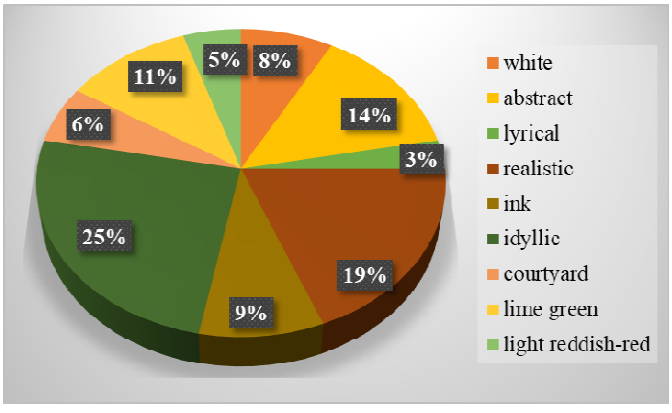
$$L_{per} = \frac{1}{N}\sum_{n=1}^{N}\left[\frac{1}{HWC}\left\|\varphi\left(G_{\mathrm{A}}\left(G_{\mathrm{B}}(X)\right)\right)-\varphi(X)\right\|_2^2+\frac{1}{HWC}\left\|\varphi\left(G_{\mathrm{B}}\left(G_{\mathrm{A}}(X)\right)\right)-X\right\|_2^2\right] \tag{20}$$

where $\varphi(\ )$ is a feature at different scales.

5 Experimental results and analyses

To evaluate the performance of the proposed algorithm, this paper uses the dataset of landscape paintings created by Castellano et al. (2021), which is a total of 35,128 paintings, including nine categories of landscape paintings, namely, white, abstract, lyrical, realistic, ink, idyllic, courtyards, lime green, and light reddish-red. The percentage of landscape paintings in different categories is shown in Figure 5. The dataset is divided into training set and test set according to 6:4. The experiments were conducted using Linux operating system with system version Ubuntu 18.04.6 LTS 64-bit, processor Intel(R) Xeon(R) Gold 5218 CPU @ 2.30GHz 24-core processor 48, RAM 176GB, and GPU used was NVIDIA A100. Pytorch deep learning framework and Cuda 11.6 were used to run the experimental code. To maintain the consistency of the experimental results, the same hyperparameter settings were used for the comparison algorithms. The model is optimised with the Adam optimiser, with the learning rates of G and D set to 0.0001 and 0.0004, respectively, and the batch size set to 32.

Figure 5 Percentage of landscape paintings in different categories (see online version for colours)



For performance evaluation, in this paper, FID, IS, SSIM, learned perceptual image patch similarity (LPIPS), sharpness difference (SD) are used as evaluation metrics to measure the quality of image generation. Where smaller value of FID indicates higher quality of generation and larger value of IS, SSIM, LPIPS, SD indicates higher quality of generated image. To facilitate the analysis, the CNN-GAN algorithm (Andreini et al., 2020), the SPA-GAN algorithm (Shen et al., 2023) and the algorithm OURs proposed in this paper are selected for the comparison experiments. In this paper, 200,000 iterations of the three algorithms are performed, with 2,000 images generated every 500 iterations, and then the FID values for the current iteration are calculated. As can be seen in Figure 6, when the number of iterations is 50,000, the FID values of CNN-GAN, SPA-GAN, and OURs are 108.32, 72.84, and 48.25, respectively, and OURs reduces by 55% and 32% compared to CNN-GAN and SPA-GAN, respectively, and the speed of decreasing and fluctuating values of FID value of OURs are better than the other two algorithms. The quality of the generated image is better in OURs.

Comparison of IS, SSIM, LPIPS, and SD values of the three algorithms is shown in Figure 7. In terms of IS metrics, OURs performs the best, extracting the multi-scale features of the image through AMS, while using SA and ECA to highlight the key features from different dimensions, and adopting the idea of adversarial learning to refine the details of the image and generate a finer image. Comparing the SSIM and LPIPS metrics, the SSIM and LPIPS of OURs are 0.6917 and 0.5834, respectively, which are improved by 17.98% and 5.47% compared to CNN-GAN, and 22.87% and 6.23% compared to SPA-GAN, respectively. CNN-GAN, although it incorporates CNN into traditional GAN to enhance the algorithm's feature extraction capability, it does not enhance the important features of the image, resulting in a loss of detail in the generated image. SPA-GAN highlights the detailed features of the image by incorporating AM in GAN, but does not improve the traditional GAN, so the image generation is not as effective as OURs. SD is a quantitative measure of the sharpness of the edges of an object in an image, the SD of CNN-GAN, SPA-GAN and OURs are 3.6241, 4.5892 and 4.9366 respectively, OURs has the largest SD value, OURs produces high-quality images, and its generated images are clearer and more varied as compared to other methods.

Figure 6 Comparison of FID values for the three algorithms (see online version for colours)

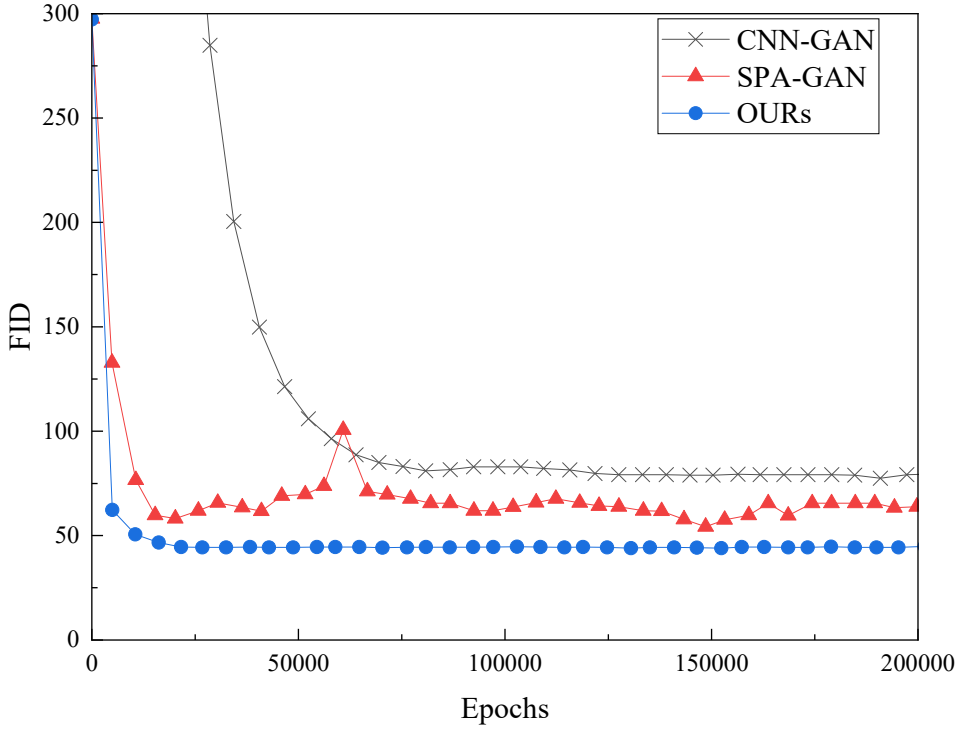


Figure 7 Comparison of image generation quality metrics for different algorithms (see online version for colours)

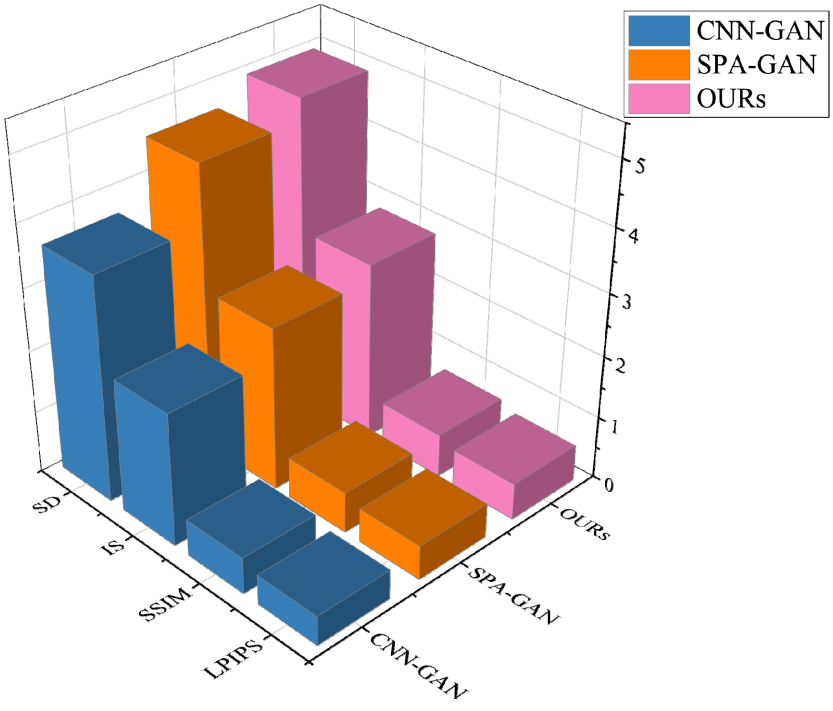


Table 1 Experimental outcome of ablation of different modules in the OURs algorithm

Algorithm	*FID*	*IS*	*SSIM*	*LPIPS*	*SD*
OURs/AMS	69.53	1.89	0.5934	0.4962	4.0569
OURs/SFT	72.14	1.65	0.5765	0.4814	3.8624
OURs/SA	56.96	2.28	0.6491	0.5103	4.5891
OURs/ECA	51.84	2.61	0.6724	0.5538	4.8142
OURs	48.25	2.86	0.6917	0.5834	4.9366

To verify the contribution of different modules in the OURs algorithm, ablation experiments are carried out on the AMS, SFT, SA, and ECA modules in the algorithm, i.e., experiments are carried out by removing the corresponding modules, which are denoted as OURs/AMS, OURs/SFT, OURs/ SA, and OURs/ECA, respectively, and the results of the ablation experiments for different modules are shown in Table 1. The performance indexes of OURs/AMS are slightly better than those of OURs/ SFT, but the gap with OURs is relatively large in all the indexes, indicating that OURs can significantly improve the generation of the algorithm by utilising AMS for feature extraction and SFT for SFT. The performance indexes of OURs/ SA and OURs/ECA are better than OURs/AMS and OURs/SFT, but they are not as good as OURs, and SA and ECA are sufficient to understand the dependency relationship between different regions in the image, which can significantly improve the effect of image generation. In summary, OURs, which incorporates all modules, performs the best, extracting multi-

scale key features of the image, with richer detail generation and clearer contour information.

6 Conclusions

Existing AI art creation image generation algorithms suffer from the problem of poor quality of generated images. For this reason, this paper proposes an AI art creation image generation algorithm based on contrast learning. Firstly, the GAN is improved by MMI (MIGAN) to decompose the noise into latent variables and noise distribution, and MS-SSIM is introduced as a mutual information factor to retain the features of the image and improve the feature learning. Secondly, AMS feature extraction module is utilised to obtain multi-scale features to avoid irreversible information loss to the image, and SFT is employed to ensure the integrity of the semantic information of the image. Then SA and ECA are introduced in MIGAN to make the generator focus on the important features of the image and finally get the generated image. Adversarial learning between the generated image and the real image through the discriminator, and designing the adversarial loss function to optimise the network model, which effectively improves the quality of image generation. The experimental results show that the FID and IS of the proposed algorithm are 48.25 and 2.86, respectively, which are better than the comparison algorithm, and can maximise the preservation of the detail information on the original image in order to enhance the realism of the generated image and improve the quality of the generated image.

The algorithm suggested in this paper can improve the quality of AI art creation image generation, but there is still room for improvement in the proposed algorithm, and the training of GAN usually has a large number of parameters and computational cost, which largely limits the efficiency of image generation. Reducing the number of network parameters and computing cost can significantly improve the running efficiency of the algorithm and reduce the training and inference cost of the model. Lightweight GAN will be further studied in the future to improve the efficiency of AI art creation image generation.

Acknowledgements

This work is supported by Sichuan Conservatory of Music 2024-2026 Higher Education Talent Cultivation Quality and Teaching Reform School-Level Project “Reform of Online-Offline Mixed Teaching of Chinese Painting Ink Composition Class under the Perspective of New Liberal Arts”.

References

Alruily, M., Said, W., Mostafa, A.M. et al. (2023) ‘Breast ultrasound images augmentation and segmentation using GAN with identity block and modified U-Net 3+’, *Sensors*, Vol. 23, No. 20, p.8599.

Andreini, P., Bonechi, S., Bianchini, M. et al. (2020) ‘Image generation by GAN and style transfer for agar plate image segmentation’, *Computer Methods and Programs in Biomedicine*, Vol. 184, p.105268.

Castellano, G., Lella, E. and Vessio, G. (2021) 'Visual link retrieval and knowledge discovery in painting datasets', *Multimedia Tools and Applications*, Vol. 80, pp.6599–6616.

Cetinic, E. and She, J. (2022) 'Understanding and creating art with AI: review and outlook', *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, Vol. 18, No. 2, pp.1–22.

Chen, M-J. and Bovik, A.C. (2011) 'Fast structural similarity index algorithm', *Journal of Real-Time Image Processing*, Vol. 6, pp.281–287.

Dong, W., Zhang, L., Lukac, R. et al. (2013) 'Sparse representation based image interpolation with non-local autoregressive modeling', *IEEE Transactions on Image Processing*, Vol. 22, No. 4, pp.1382–1394.

Enjellina, B. and Rossy, A.G.C. (2023) 'A review of AI image generator: Influences, challenges, and future prospects for architectural field', *Journal of Artificial Intelligence in Architecture*, Vol. 2, No. 1, pp.53–65.

Fan, X., Gong, M., Wu, Y. et al. (2023) 'Maximizing mutual information across feature and topology views for representing graphs', *IEEE Transactions on Knowledge and Data Engineering*, Vol. 35, No. 10, pp.10735–10747.

Huang, G. and Jafari, A.H. (2023) 'Enhanced balancing GAN: minority-class image generation', *Neural Computing and Applications*, Vol. 35, No. 7, pp.5145–5154.

Lu, Z-L. and Dosher, B.A. (1998) 'External noise distinguishes attention mechanisms', *Vision Research*, Vol. 38, No. 9, pp.1183–1198.

Mi, Z., Jiang, X., Sun, T. et al. (2020) 'GAN-generated image detection with self-attention mechanism against GAN generator defect', *IEEE Journal of Selected Topics in Signal Processing*, Vol. 14, No. 5, pp.969–981.

Qi, Z., Sun, J., Qian, J. et al. (2021) 'PCCM-GAN: photographic text-to-image generation with pyramid contrastive consistency model', *Neurocomputing*, Vol. 449, pp.330–341.

Santos, I., Castro, L., Rodriguez-Fernandez, N. et al. (2021) 'Artificial neural networks and deep learning in the visual arts: a review', *Neural Computing and Applications*, Vol. 33, pp.121–157.

Shen, L., Yan, J., Sun, X. et al. (2023) 'Wavelet-based self-attention GAN with collaborative feature fusion for image inpainting', *IEEE Transactions on Emerging Topics in Computational Intelligence*, Vol. 7, No. 6, pp.1651–1664.

Suryadevara, C.K. (2020) 'Generating free images with OpenAI's generative models', *International Journal of Innovations in Engineering Research and Technology*, Vol. 7, No. 3, pp.49–56.

Wang, H. and Ma, L. (2023) 'Image generation and recognition technology based on attention residual GAN', *IEEE Access*, Vol. 11, pp.61855–61865.

Wang, K., Gou, C., Duan, Y. et al. (2017) 'Generative adversarial networks: introduction and outlook', *IEEE/CAA Journal of Automatica Sinica*, Vol. 4, No. 4, pp.588–598.

Xu, W., Keshmiri, S. and Wang, G. (2019) 'Adversarially approximated autoencoder for image generation and manipulation', *IEEE Transactions on Multimedia*, Vol. 21, No. 9, pp.2387–2396.

Zhang, M., Wang, H., He, P. et al. (2022) 'Exposing unseen GAN-generated image using unsupervised domain adaptation', *Knowledge-Based Systems*, Vol. 257, p.109905.

Zhao, B., Zhan, D., Zhang, C. et al. (2023) 'Computer-aided digital media art creation based on artificial intelligence', *Neural Computing and Applications*, Vol. 35, No. 35, pp.24565–24574.