

VISUAL EFFECTS OF URBAN ENVIRONMENT ART DESIGN BASED ON COMPOSITE CYCLE GENERATIVE ADVERSARIAL NETWORK

Changning WANG¹, Na LI¹, Jiyu WANG²✉

¹Henan Technical College of Construction, 450000 Zhengzhou, China

²Faculty of Engineering, SIAS University, 451150 Zhengzhou, China

Highlights:

- this study proposes a Cycle Generative Adversarial Network, which incorporates a multi-head self-attention mechanism for enhanced design;
- the study compares the proposed network with publicly available datasets and commonly used optimization networks;
- these results demonstrate that the proposed model analyzes images in various artistic styles and optimizes the visual effects of urban environments.

Article History:

- received 9 April 2025
- accepted 4 December 2025

Abstract. In recent years, many regions have invested significant efforts in the in-depth study of artistic visual effects in urban environments to enhance the aesthetic appeal of urban districts. However, traditional visual optimization models are unable to refine image edges and textures, leading to distorted results. To optimize the artistic effects of urban district images, this study proposes a Cycle Generative Adversarial Network that incorporates a multi-head self-attention mechanism to enhance design. The network architecture is modified by adding multi-head self-attention modules. Spectral normalization is also introduced to improve the optimization process of both the generator and the discriminator. Finally, the images generated at different layers are concatenated, and the average of the extracted features is used as the output. The study compares the proposed network with publicly available datasets and commonly used optimization networks. The generated images achieve a pixel accuracy of 96.8% and a classification accuracy of 97.4%. In validation tests with a self-built dataset, the Peak Signal-to-Noise Ratio reaches 34.2 dB, outperforming the comparison models. In a comprehensive evaluation across various metrics, the Structural Similarity Index is 98.23% and the Visual Semantic Similarity is 96.78%. These results demonstrate that the proposed model analyzes images in various artistic styles and optimizes the visual effects of urban environments. It is beneficial for creating urban districts with distinctive artistic styles, enhancing the local cultural atmosphere, and attracting more domestic and international tourists.

Keywords: Cycle Generative Adversarial Network, multi-head self-attention mechanism, visual effects, urban art, style transfer, landscape management.

✉Corresponding author. E-mail: wangjiyuzz@163.com

1. Introduction

Urban district environments symbolize the historical development of city culture and simultaneously convey local culture to visitors (Kumar et al., 2023). With the increasing popularity of intelligent digital networks, traditional visual artistic effects in urban districts can no longer meet tourists' needs (Elpus, 2022). Effectively conveying a city's unique cultural characteristics and enhancing the visual features of urban district art have become key challenges (Shen et al., 2023). Research shows that to optimize the visual effects of urban district environments, non-photorealistic rendering techniques based on image analogies are proposed. This technique adjusts transformation parameters and weights to transfer the style of target images by analogizing multiple image styles. However, the method requires large amounts of image data and is inefficient

(Rosin, et al., 2022). Meanwhile, scholars have found that convolutional neural networks can effectively capture the artistic effects of style transfer, but they suffer from issues such as the loss of content features and inconsistent styles. With the ongoing development of intelligent technology, many researchers have found that Generative Adversarial Networks (GANs) can generate high-quality images while analyzing diverse artistic styles (Chai et al., 2022). Convolutional neural networks rely on a fixed receptive field for local feature extraction, making it difficult to capture long-range spatial correlations and leading to distorted representations of complex textures and cross-regional structures in block images. However, these networks still lack refinement in handling image details. The compound neural structure of the Cycle Generative Adversarial Network (CycleGAN) can effectively enrich image edges and texture details, presenting artistic styles in a more complete man-

ner, with visual effects that are more unified and distinctive. Compared with the existing architecture, CycleGAN demonstrates greater stability and better detail restoration when handling style transfer tasks, especially excelling at texture retention and color coordination in complex street scene images. Its unique cyclic consistency constraint mechanism effectively mitigates information loss during image conversion. The multi-head self-attention mechanism can extract feature information at multiple levels via multiple attention heads, enrich the model's data, and improve its analytical performance. The existing art designs mostly rely on manual craftsmanship, making it difficult to achieve unification and innovation in large-scale street style. With the development of Artificial Intelligence (AI), automated, intelligent solutions for street art design are now possible. However, current intelligent methods still have limitations when it comes to integrating multi-scale cultural elements, especially in precisely mapping cross-modal cultural semantics. Based on this, the research proposes a style transfer model integrating CycleGAN and a multi-head self-attention mechanism, aiming to enrich the visual effect of the block environment and present distinctive artistic designs to tourists. The innovation of the research lies in constructing an intelligent style transfer framework that combines capabilities for detail preservation and cultural semantic understanding. By introducing a multi-head self-attention mechanism, the model's ability to capture the features of traditional patterns, architectural colors, and regional symbols is enhanced. Combined with the bidirectional mapping structure of CycleGAN, the artistic reconstruction of block images is achieved. Enhance the consistency and cultural fit of visual expression while reducing human intervention. This model achieves the collaborative optimization of texture details and cultural semantics through a dual-path feature fusion module.

2. Related works

With the rapid advancements in deep learning and neural network technologies, CycleGAN has been widely applied across various fields. For instance, Liu et al. (2023) proposed a CycleGAN combined with a heterogeneous discriminator to enhance the dehazing capability of computer vision images. By incorporating an auxiliary detail-generation module, they improved image quality. Experimental results showed that the model improved the visual quality of images and outperformed the comparison models across multiple quantitative metrics (Liu et al., 2023; Lakshmi et al., 2023). The Bargshady team constructed a novel network architecture by combining a GAN and a CycleGAN to analyze the extent of coronavirus-related damage to the human body. Using standard data augmentation techniques, they improved the model's generalization. The model achieved an average absolute error of only 0.16 and was applied to the analysis of X-ray images of the coronavirus (Bargshady et al., 2022). Sun and his team built a magnetic resonance tomography diagnostic model us-

ing CycleGAN to enhance medical image synthesis. The model perceives and discriminates images from multiple angles, successfully performing image transformation between unordered data, reducing computational load, and improving synthesis efficiency (Sun et al., 2023). Fan et al. (2024) proposed a multi-feature fusion CycleGAN to reduce the interference of background and clothing factors on pedestrian recognition. By ranking the sample's computational similarity scores, the model was proven to effectively identify pedestrians while avoiding the influence of factors such as background, veils, and clothing (Fan et al., 2024).

Additionally, as living standards have improved, the demand for high-quality lifestyles has grown, making individuals more inclined to seek out visually distinctive imagery. To cater to the public's preference for unique artistic images, scholars worldwide have conducted extensive research. For example, Yang et al. (2022) proposed an improved GAN image optimization model to enhance the visual effects of artistic portrait videos. They introduced an open-source deep learning framework to transfer video styles. Experimental results demonstrated that the model could generate high-quality videos while maintaining stylistic consistency (Yang et al., 2022). Logeshwaran et al. (2022) introduced a visual analysis algorithm that uses segmentation to refine visual processing methodologies. By improving pixel resolution and clarity, the model aims to enhance image quality. It is adaptable to various image frame rates, enhancing the overall visual effect (Lu & Hu, 2022). Mao et al. (2022) proposed a multi-scale feature GAN model for the style transfer of text images, using an unsupervised learning algorithm to achieve real-time monitoring of image styles. The model can synthesize features of various text styles and improve the visual quality of text-to-image transformation (Mao et al., 2022). Chen et al. (2023) constructed an independent adversarial network based on a GAN to improve image generation quality. He incorporated repair and optimization network modules, whose synergy improved both the visual fidelity and overall image quality. Experimental results demonstrated that this model surpassed comparable models in image rendering capability (Chen et al., 2023). Image marking is also a key link in improving the efficiency of image understanding and retrieval. In recent years, researchers have significantly improved the accuracy and semantic richness of labeling by introducing deep generative models to achieve automatic labeling (Iraji, 2025a). Iraji et al. (2024) explored the synergistic effects of wavelet transform, visual transform, dropout layer, interpolation regularization, and ensemble strategy in semi-supervised generative adversarial networks. The discovery indicates that extracting multiscale texture features via wavelet transforms and capturing long-range dependencies, combined with visual transformations, can effectively enhance the model's discrimination ability for unlabeled data. Therefore, a new image labeling framework is proposed. By introducing a consistency regularization mechanism, the fluctuations

in pseudo-labels during training are further suppressed. The results show that the error rate of this method has been significantly reduced (Iraji et al., 2024). To address the degradation in model performance caused by pseudo-label noise in semi-supervised learning for the diagnosis of skin cancer lesions, Iraji (2025b) proposed a co-training strategy that integrates wavelet feature enhancement and a visual attention mechanism. The model's prediction consistency on unlabeled samples is constrained by interpolation regularization, and ensemble learning reduces variance. The results show that this method achieves a balanced accuracy of 80.35% on the ISIC-2019 dataset with only 10% of the data labeled (Iraji, 2025b).

To sum up, there are issues such as the inability to guarantee high-quality style transfer and image output quality in visual presentations. It is urgent to refine image texture features to avoid distortion. Existing research has not yet solved the problem of cross-scale feature fusion. However, the combination of cycle and the multi-head self-attention mechanism can effectively integrate feature information of different scales and enhance the ability to restore image details. By introducing a multi-head attention mechanism, the model can dynamically focus on key areas during generation, enhance texture consistency, and suppress distortion. Therefore, this study selects CycleGAN and combines it with the multi-head self-attention mechanism to enhance the extraction of different style features in the target image, thereby achieving accurate analysis of real images. The research aims to create realistic artistic effects, enhance the richness of the pictures, and contribute to the development of the cultural tourism industry.

3. Design of block art visual effect model based on CycleGAN

3.1. Optimization of CycleGAN design by integrating a multi-head self-attention mechanism

GAN is an artificial neural network framework consisting of a generator and a discriminator, both of which are trained via adversarial optimization (Kammoun et al., 2022). This structure enables the generator to produce results that

resemble real data, while the discriminator more accurately distinguishes between real and fake data. Therefore, GANs are often used for artistic style transfer, ensuring that the visual effects produced are consistent with the target image. The relationship between the generator and discriminator in this network is shown in Eq. (1).

$$\min_B \max_A V(B, A) = E_{X \sim P_d(x)} [\log A(x)] + E_{z \sim P_z(x)} [\log (1 - A(B(z)))]. \quad (1)$$

In Eq. (1), z represents the noise of the input generator, B is the generator, X is the image feature of the source domain, A is the discriminator, $B(z)$ is the image generated by the generator, x is the sample feature of the data domain, $P_d(x)$ is the feature distribution of the input image, $P_z(x)$ is the noise distribution generated by the noise passing through the generator, and E is the expectation function. The generator and discriminator can improve their ability to analyze and classify image features during adversarial training. However, GAN has limitations in analyzing complex artistic images. To address this, research has found that CycleGAN can be trained directly without paired data in advance and can effectively analyze key image features (Wen et al., 2022). The CycleGAN structure is shown in Figure 1.

As shown in Figure 1, CycleGAN is a bidirectional cyclic structure composed of two generators and two discriminators. One generator converts the source-domain image to the target-domain image, while the other performs the reverse conversion. The two discriminators determine whether the generated image is from the target or source domain. By introducing cyclic consistency loss, it is ensured that the input image is restored as closely as possible to its original content after bidirectional conversion, thereby avoiding distortion in the generated result (Wang et al., 2023). To enhance the precision and quality of image transformation, the network structure integrates various loss functions into its objective function. The objective function within the target domain is presented in Eq. (2).

$$O_c(B_1, A_Y, X, Y) = E_{y \sim P_d(y)} [\log A_Y(y)] + E_{x \sim P_d(x)} [\log (1 - A_Y(B_1(x)))]. \quad (2)$$

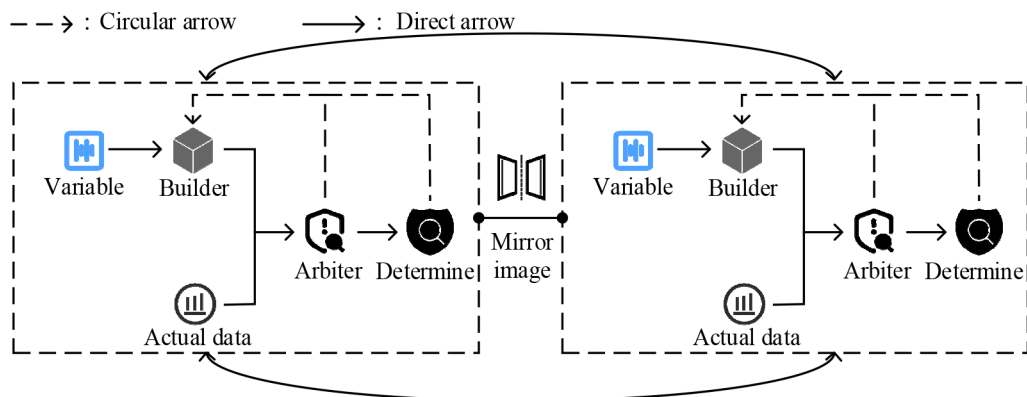


Figure 1. Schematic diagram of CycleGAN structure

In Eq. (2), B_1 represents a generator in CycleGAN, Y represents the image features of the target domain, and Y represents the sample features in the data domain. Meanwhile, CycleGAN ensures the accuracy of the cross-mapping generated by the generator by adding a cycle-consistency loss, thereby enhancing the fidelity and overall quality of the generated images. The cycle-consistency loss function of the CycleGAN is given by Eq. (3).

$$O_g(B_1, B_2) = E_{x \sim p_d(x)} [F(B_1(X)) - 1] + E_{y \sim p_d(y)} [B_1(F(Y)) - 1]. \quad (3)$$

In Eq. (3), B_2 represents another generator of CycleGAN. The cycle-consistency loss function is part of the objective function of the CycleGAN. The complete objective function is shown in Eq. (4).

$$O(B_1, B_2, A_X, A_Y) = O_c(B_1, A_Y, X, Y) + O_c(B_2, A_X, Y, X) + \lambda O_g(B_1, B_2). \quad (4)$$

In Eq. (4), λ represents the weight coefficient. The complete objective function of CycleGAN consists of the target-domain and source-domain losses and the cycle consistency loss. However, CycleGAN is unable to distinguish complex style types. When generating image features, CycleGAN cannot extract detailed local areas of artistic styles, nor can it express multiple semantic features. Therefore, a multi-head self-attention mechanism is introduced. This mechanism comprises multiple independent heads that can learn distinct input information and integrate various spatial features, thereby enhancing the model's ability to analyze features (Jia et al., 2024). The CycleGAN improvement mechanism with multi-head self-attention is shown in Figure 2.

As shown in Figure 2, the study introduces multi-head self-attention modules into the encoding and transformation layers of the CycleGAN to capture long-range dependencies and local detail features in the image. And the research aims to optimize its feature extraction process. The network structure consists of two independent feature extraction modules. They are respectively the content feature extraction module and the style feature extraction module.

The content feature extraction module uses a pre-trained VGG network to extract high-level semantic information from images and retain the content structure of source-domain images. The style feature extraction module captures the texture and color distributions of artistic styles by analyzing the statistical properties of multi-layer convolutional feature maps. The self-attention module consists of multiple parallel attention heads, each of which independently computes queries, keys, and values to capture feature associations in different subspaces. Through the zooming dot-product attention mechanism, the model can focus on key areas of the image, thereby enhancing its ability to express local style details. The outputs of each head are concatenated after linear transformation, and then a nonlinear mapping is performed by the feedforward network to achieve the fusion and enhancement of multi-scale semantic features. The self-attention calculation within the multi-head self-attention module is presented in Eq. (5).

$$head_i = \text{Attention}(Q_j, K_j, V_j) \in R^{L \times D_h}. \quad (5)$$

In Eq. (5), $head_i$ represents the computation of a single head structure, i represents the number of head structures, Q_j represents the query vector, K_j represents the key vector, V_j represents the value vector, L represents the sequence length, and D_h represents the input feature dimension for each head structure. The multi-head self-attention module aggregates all computations, with the output given by Eq. (6).

$$Z = \text{Multi-Head} \triangleq (head_1 \oplus head_2 \oplus \dots \oplus head_H). \quad (6)$$

In Eq. (6), Z represents the output vector, H represents the number of head structures, and $\oplus$ indicates the concatenation process. The multi-head self-attention module combines the calculation results of Eqs. (5)–(6) to achieve global dependency modeling of input features, effectively capturing long-distance semantic associations. The attention module maps the query, key, and value to multiple subspaces via linear projections, and computes attention weights in parallel across different representation subspaces, thereby capturing multi-angle context dependencies in the input sequence. Subsequently, the outputs from each head are concatenated, and the

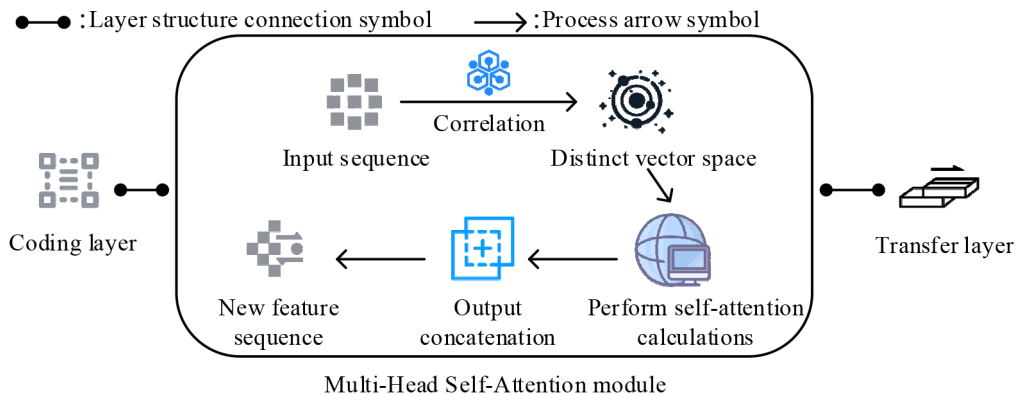


Figure 2. Schematic of the improved CycleGAN structure

information is integrated through global linear transformation to enhance the model's ability to express complex semantic structures (Reddi et al., 2024). The attention mechanism module analyzes the feature distribution via singular value decomposition to determine the optimal projection dimension, thereby optimizing the subspace mapping efficiency for queries, keys, and values, reducing computational redundancy, and ensuring semantic integrity. By controlling the rank of the projection dimension, the main feature directions are retained, noise interference is suppressed, and the attention weights are more focused on the key associated fragments. The weight distribution is based on the similarity between the query and the key. After scaling the dot product, it is normalized to a probability distribution through the Softmax function, enabling the model to dynamically focus on the key positions in the input sequence. This module effectively enhances the compound CycleGAN's ability to analyze and express multiple semantics. It effectively identifies disparities between the source and target domains' features, thereby enhancing the model's capacity to capture these differences. The optimization design of the compound CycleGAN, utilizing the multi-head self-attention module, is illustrated in Figure 3.

As shown in Figure 3, the optimized network structure developed in this research comprises an encoding layer, a transformation layer, a self-attention module layer, and a discrimination layer. After the encoding layer extracts the image features of the source domain, they are passed through the transformation layer to the multi-head self-attention module. Through multiple sets of parallel attention

mechanisms, long-distance semantic dependencies are captured to better express key features. The outputs from each head are concatenated and sent to the feed-forward network for nonlinear transformation. Eventually, the decoding layer reconstructs the target-domain image, thereby achieving collaborative optimization of details and structure in cross-domain mapping. Among them, the encoding layer is composed of three convolutional units, the transformation layer is composed of nine groups of residual module units, and the decoding layer is composed of two transposed convolutional units and one convolutional unit. The discrimination layer, as the final step of CycleGAN, divides the input image into independent units, determines whether each unit is true or false, and finally makes a comprehensive judgment on each unit's results.

3.2. Method for constructing a visual effect model based on CycleGAN

To enhance CycleGAN's deep learning neural network's style recognition ability for complex images, a multi-head self-attention module is added. Meanwhile, the study further verifies the optimized compound neural network and applies it to urban environment art design. However, the compound CycleGAN cannot effectively balance the adversarial relationship between the generator and discriminator, and achieving synchronous optimization is difficult. Therefore, the study introduces spectral normalization to optimize the compound CycleGAN structure, and compares the structure before and after optimization in Figure 4.

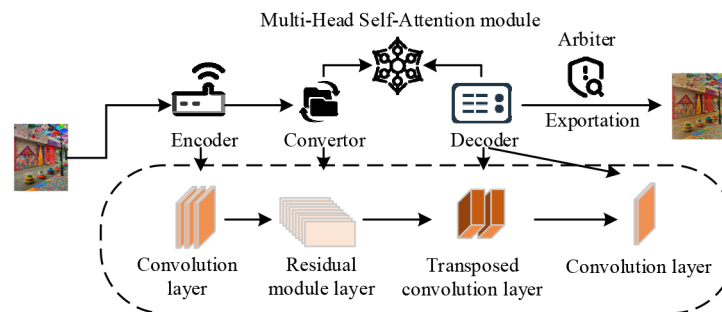


Figure 3. Optimized composite CycleGAN workflow diagram

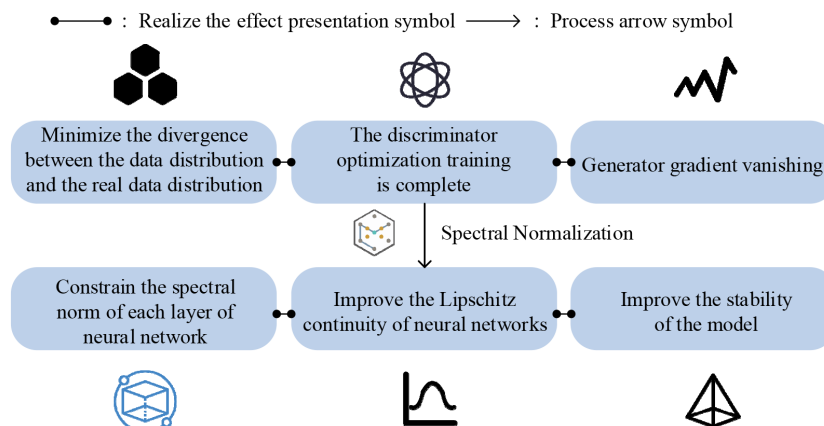


Figure 4. The optimization of complex CycleGAN by spectral normalization

As shown in Figure 4, the compound CycleGAN analyzes the scatter distribution between generated data and real data. After optimizing the discriminator's training, this leads to a vanishing gradient problem in the generator of the network. Spectral normalization imposes weight constraints on each layer of the neural network, enabling it to satisfy the Lipschitz condition and enhancing the model's stability. The spectral norm calculation in spectral normalization is shown in Eq. (7).

$$\sigma(W) = \sup_{x_2 \leq 1} Ax_2. \quad (7)$$

In Eq. (7), x represents the image features input into the convolutional structure, and W is the parameter matrix. The study applies spectral normalization to each convolutional operation of the discriminator, estimates the maximum singular value of the weight matrix via the iterative power method, and imposes dynamic constraints on it, thereby suppressing gradient oscillations during training. The study constrains the spectral norm of the weight matrix to around 1 to ensure that the discriminator's gradient change during model updates remains within a reasonable range, avoiding training collapse from overly strong or weak feedback signals. The research introduces it into the generation process of artistic style transfer, enabling the model to better express detailed textures while maintaining the cultural characteristics of the block. The spectral normalization mechanism is applied after each convolutional layer and before the activation layer. After obtaining feature representations from the convolutional network, spectral norm constraints are first applied to the weights and then passed to the activation function, thereby ensuring the stability of forward propagation. Spectral normalization effectively curbs the discriminator's overresponse to generated samples early in training by constraining the spectral norm of each layer's weight matrix, enabling the discriminator to focus more on subtle differences between real and generated data. The introduced spectral normalization technique effectively ensures that both the generator and discriminator are optimized simultaneously, allowing the composite CycleGAN to transfer the style of block

art and unify the artistic effect of the block. The network structure of the block art style transfer generated by the composite CycleGAN is shown in Figure 5.

From Figure 5, it can be seen that the generative network architecture of CycleGAN's artistic visual effect model comprises three modules: content information extraction, style information extraction, and feature map fusion. The content extraction module includes the recognition of local and global features in block art images, focusing on the block image's contours and spatial structure. After extracting the content features of the source image, the feature map is passed to the multi-head self-attention module, where multiple attention heads further analyze the feature information, summarizing the self-attention feature map. Another part of the CycleGAN visual effect model consists of the style feature extraction module and feature fusion module. By expanding the channels of the source image, the model improves its ability to extract the image's overall style. Finally, the content feature map and style feature map are concatenated to form the final generated image. The loss function in the composite CycleGAN is typically used to measure the difference between generated and real images. The loss function is shown in Eq. (8).

$$\text{loss}(x, y) = \frac{1}{n} \sum_{i=1}^n |y_i - f(x_i)|. \quad (8)$$

In Eq. (8), y_i represents the features of the real image, $f(x_i)$ represents the features of the generated image, and n represents the number of samples. However, the loss function cannot guarantee that the image retains sharp contours, textures, and other fine details. Therefore, the study selects the smooth L1 loss, as expressed in Eq. (9).

$$\text{loss}_s(x, y) = \frac{1}{n} \sum_{i=1}^n \begin{cases} 0.5 * (y_i - f(x_i))^2, & \text{if } |y_i - f(x_i)| < 1 \\ |y_i - f(x_i)| - 0.5, & \text{otherwise.} \end{cases} \quad (9)$$

In Eq. (9), $\text{loss}_s(x, y)$ denotes the value of the smooth L1 loss function. The smooth L1 loss effectively minimizes the discrepancy between the generated and input images, ensuring style consistency and maintaining the same visual

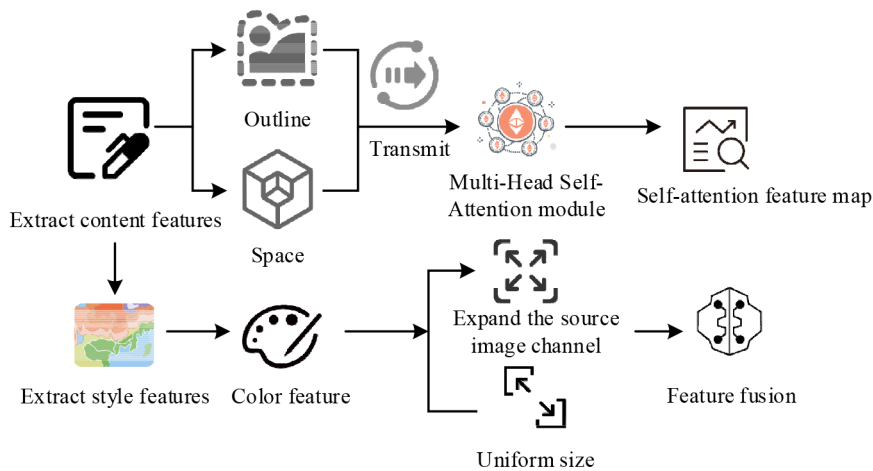


Figure 5. Composite CycleGAN-generated structural map of neighborhood art styles

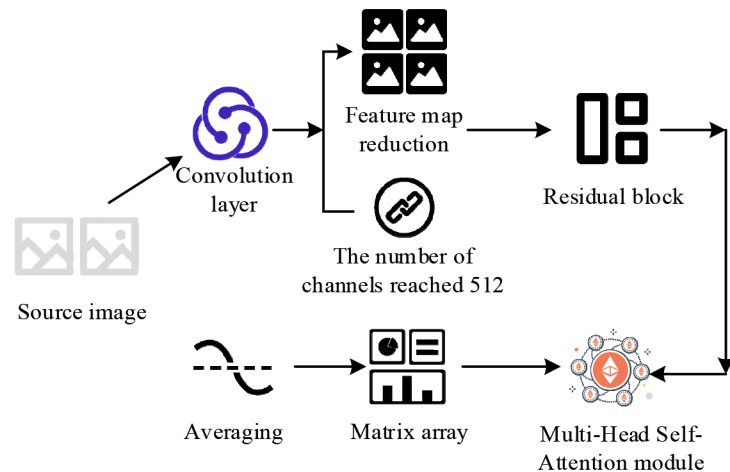


Figure 6. Graph of the discriminant network in the composite model

effect. After the CycleGAN visual model ensures the style consistency of the generated image, the discriminative network component of the visual model is illustrated in Figure 6.

In Figure 6, the structure of the discriminative network in the visual effect model is shown. The source image first passes through a convolutional layer, where the feature map is down sampled and the number of channels is increased to 512. The feature map is then processed by a residual module, which captures both shallow and deep-layer image information. Following this, multiple attention heads within the multi-head self-attention module analyze the source image's features. The self-attention feature map is then concatenated with the output of the residual module, resulting in a mapped array of average values. The research on the composite network introduces cyclic consistency loss to ensure bidirectional mapping closure between the source and target domains, thereby satisfying the closed-loop constraints of the forward and reverse mappings and reducing image distortion. This loss term further enhances the ability to preserve structure by minimizing pixel-level differences between the original and reconstructed images. Meanwhile, by combining the L1 smoothing loss with the adversarial loss, a multi-task optimization objective is formed, effectively suppressing gradient dispersion during gradient updates and improving model convergence. In addition, the research embeds the multi-head self-attention module, spectral normalization layer, and residual connection mechanism into the deep structures of the generator and discriminator to enhance the cross-layer transfer and global dependency modeling capabilities of key features. Spectral normalization constrains the Lipschitz constant of the weight matrix layer by layer to suppress gradient oscillations during training and improve the model's stability. The multi-head self-attention mechanism effectively captures long-range spatial associations, preserves image texture and edge structure, and significantly enhances semantic consistency in style transfer.

Based on the above content, the learning rate of the proposed hybrid model in the research is 0.0002, the batch

size is set to 8, the optimizer adopts AdamW, the number of iterations is set to 200 rounds, and the weight attenuation coefficients of the generator and the discriminator are $1e-4$ and $5e-5$, respectively. The modified network layer includes the 3rd to 8th residual blocks of the generator and the multi-head self-attention module in the discriminator, introducing spectral normalization to enhance training stability. The network architecture is as follows: The input image first passes through a convolutional layer with 7×7 convolutional kernels, 64 output channels, and a stride of 2. After instance normalization and ReLU activation function processing, it is sent to two down sampled convolutional layers, each using 3×3 convolutional kernels with a stride of 2, with the number of channels doubling to 128 and 256, respectively. Subsequently, six residual blocks are connected, each containing two 3×3 convolutional layers and instance normalization, maintaining a 256-channel feature output. In the upsampling stage, transposed convolution is used to restore spatial resolution. Eventually, it is mapped to the three-channel output via a 7×7 convolutional layer, with the Tanh activation function generating the final image. The discriminator adopts the PatchGAN architecture, using a 70×70 receptive field to distinguish the authenticity of local image regions and outputting an $N \times N$ feature matrix. The model still has limitations in real-time rendering, especially in generating high-resolution medical images, where inference speed is constrained by the computational redundancy of deep networks. In response to this, the research introduces knowledge sublimation technology to transfer the deep semantic knowledge of the pre-trained generator to the lightweight student network. Through the dual constraints of feature distillation and output distillation, a balance is achieved between model compression and performance maintenance. The attention map of the middle layer of the teacher network is aligned with the corresponding layer of the student network for L2 loss. Meanwhile, the Softmax output distribution uses KL divergence as the supervisory signal, ensuring the fidelity of style details while improving reasoning efficiency.

4. Performance verification of block environment art visual model based on CycleGAN

4.1. Performance verification of composite CycleGAN

To verify the superior performance of the proposed composite CycleGAN integrated with a multi-head self-attention mechanism, the study used the following software and hardware configurations: Ubuntu 18.04.5, 64-bit operating system, Tesla V100 SXM2 model for the GPU, deep learning frameworks including TensorFlow 2.0, Pytorch 1.10, Python 3.8+, and CUDA 11.2. The study selected the Vangogh2photo dataset with an image size of 256×256 . 50% of the dataset was used as the training set, 30% as the validation set, and 20% as the test set. Data set link to (<https://aistudio.baidu.com/datasetdetail/41779>). GAN, Deep Convolutional GAN (DCGAN), and Conditional Generative Adversarial Networks (CGAN) were chosen as the comparison networks. First, the images generated by the four networks were compared for similarity with the real images in the dataset, and the Mean Squared Error (MSE) of the similarity is shown in Figure 7.

From Figure 7a, it can be seen that when the composite CycleGAN was trained for 100 iterations on the training set, the MSE value dropped to 0.72 px^2 . After 600 iterations, the MSE value further reduced to 0.12 px^2 . CGAN had a slightly lower MSE value than the composite CycleGAN after 100 iterations, with a value of 0.79 px^2 . After 600 iterations on the training set, DCGAN showed an MSE of 0.25 px^2 . The traditional GAN performed the worst, with an MSE value of 0.30 px^2 after 600 iterations. Figure 7b shows the MSE values for the generated images on the validation set across the four networks, with iteration counts ranging from 100 to 300. Composite CycleGAN, DCGAN, and CGAN showed significant error reduction. After 600 iterations, the MSE value for the composite CycleGAN was 0.08 px^2 . In contrast, the traditional GAN had an MSE value of 0.21 px^2 . Overall, the composite CycleGAN outperformed the other three networks in terms of MSE values for style recognition on the Western oil painting dataset. After analyzing the

MSE values of the generated images, the study further compared their pixel and style classification abilities. The comparison results are shown in Figure 8.

As shown in Figure 8a, GAN achieved a pixel accuracy of 70.5% after 50 iterations. When the number of iterations increased to 500, the pixel accuracy rose to 79.8%. The classification accuracy averaged at 85.2% after over 500 iterations. In Figure 8b, CGAN showed an overall upward trend in both pixel and classification accuracy during its iterations. After 500 iterations, the pixel accuracy was 83.6%, and the average classification accuracy was 91.6%. Figure 8c shows that DCGAN achieved higher pixel accuracy than GAN and CGAN after 50 iterations, with 81.3%. By 500 iterations, the classification accuracy improved to 92.1%. In Figure 8d, the composite CycleGAN outperformed the comparison networks in both pixel and classification accuracy. After 500 iterations, the pixel accuracy reached 96.8%, and the average classification accuracy was 97.4%. Overall, the composite CycleGAN achieved superior pixel and style classification accuracy in the style transfer image output compared to the other three networks, effectively recognizing the input images. To further validate the quality of the generated images, the Fréchet Inception Distance (FID) metric was used to analyze the four networks. The comparison results are shown in Figure 9.

As shown in Figure 9a, the GAN achieved an FID of 59.4 after 100 iterations. By 600 iterations, the FID value decreased to 31.2, indicating subpar training performance. In contrast, the composite CycleGAN had an FID of 50.8 after 100 iterations, which was significantly lower than those of the comparison networks. After 600 iterations, the FID value further decreased to 23.3. Figure 9b shows the comparison results of the four networks on the validation set. The composite CycleGAN started with an FID of 38.9 in the early stages, which dropped to 16.9 in later iterations. After 600 iterations, the CGAN and DCGAN had FID values of 23.1 and 21.9, respectively. Overall, the composite CycleGAN demonstrated the closest feature distribution to real images among the generated images in the Western oil painting dataset, producing higher-quality images than the comparison networks.

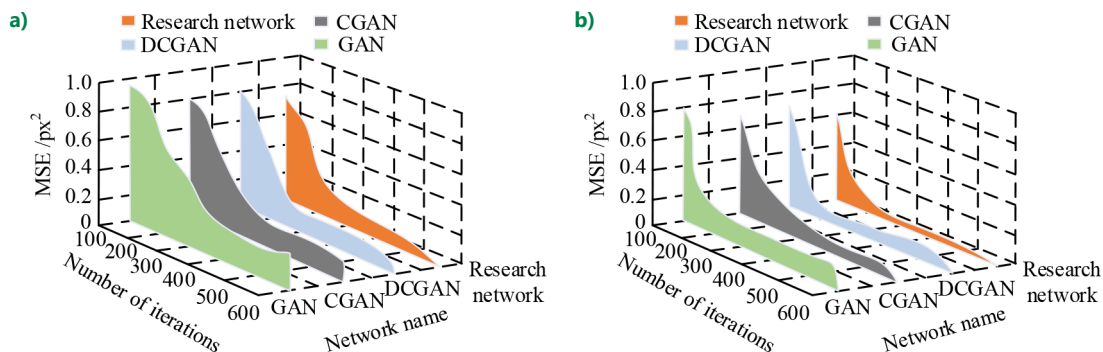


Figure 7. Comparison of four networks in MSE: a) MSE comparison in training set; b) MSE comparison in validation set

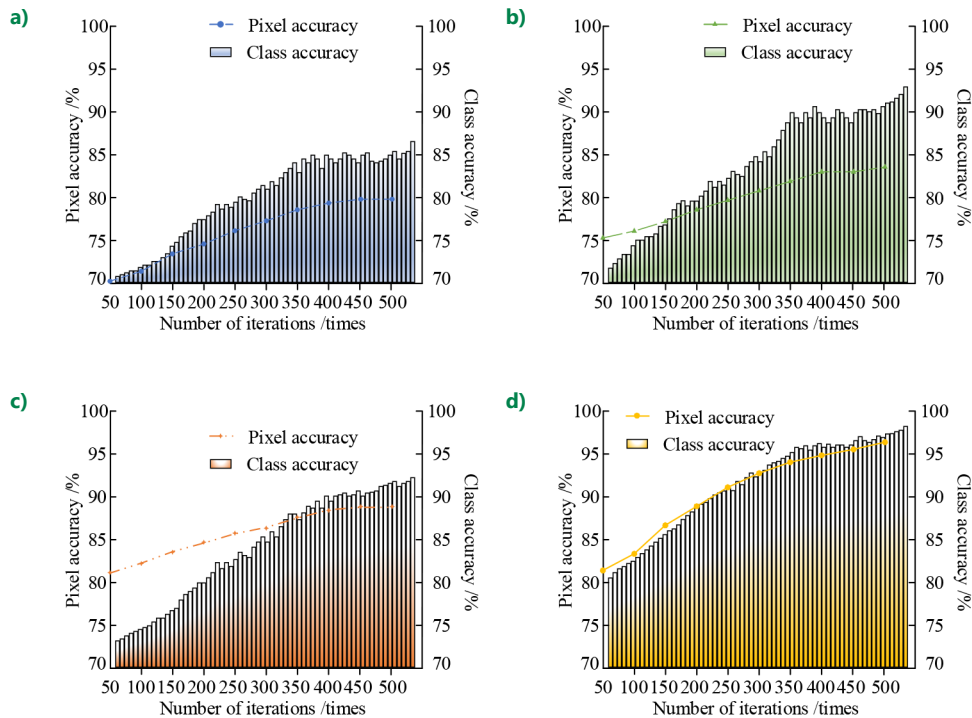


Figure 8. Comparison of the accuracy of pixels and classification: a) Accuracy of GAN in pixels and categories; b) Accuracy of CGAN in pixels and categories; c) Accuracy of DCGAN in pixels and categories; d) Accuracy of research network in pixels and categories

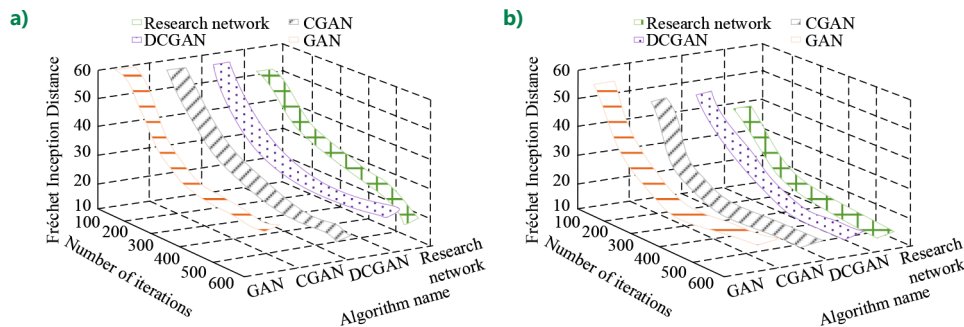


Figure 9. Comparison of FID of four networks: a) Comparison of FID in training set; b) Comparison of FID in verification set

4.2. Performance verification of the art visual effect model based on CycleGAN

After validating the superiority of the composite CycleGAN, the study further applied the CycleGAN-based art visual effect model for performance analysis. The study selected excellent street scene cases from domestic urban district art designs, capturing images from various angles and styles as dataset samples. 60% of the dataset was used for training, 20% for validation, and 20% for testing. All image acquisition is carried out under natural lighting conditions, avoiding direct strong light and shadow interference to ensure true color reproduction. EXIF metadata records information such as shooting time, device model, aperture, and shutter speed, which is used for subsequent illumination consistency correction. All images are uniformly resized to 1024×1024 and preprocessed using a tone-mapping algorithm based on HDRNet to enhance lo-

cal contrast and detail expressiveness. The preprocessed images are normalized and then input into the model for training. The experiment was carried out using Python 3.8, Pytorch 1.10, CUDA 10.1, among other configurations. The experimental setup is shown in Table 1.

Table 1. Experimental parameters and configuration

/	Configuration item	Version model
Hardware	GPU	NVIDIA GeForce GTX 1080Ti
	Video memory	16G
	CPU	Intel-Xeon Gold 6226R
	Internal memory	32G
Software	Operating system	Windows 10
	Programming language	Python 3.8
	Frame	Pytorch1.10
	CUDA	10.1

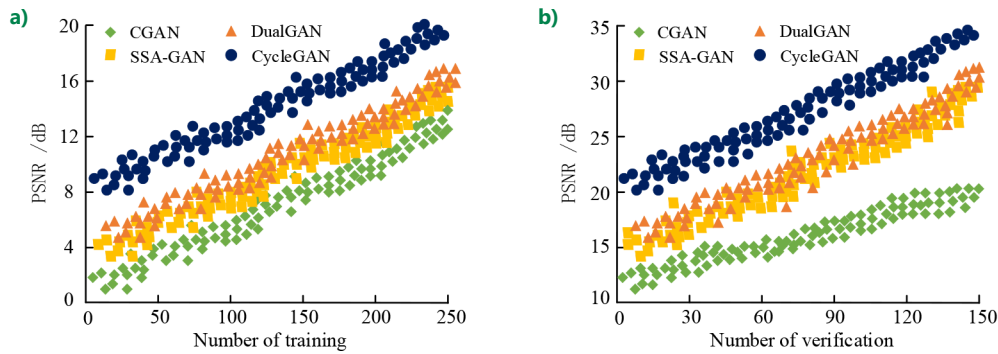


Figure 10. Comparison of four models in PSNR: a) Comparison of PSNR in training set; b) Comparison of PSNR in validation set

Based on the experimental conditions in Table 1, the study selected CGAN, Semantic-Spatial Aware Generative Adversarial Networks (SSA-GAN), and Dual Generative Adversarial Networks (DualGAN) as comparative models for verification. The study first selected the Peak Signal-to-Noise Ratio (PSNR) metric for comparing and analyzing the four models across both the training and validation sets. The specific results are shown in Figure 10.

As shown in Figure 10a, the CGAN produced lower-quality images during the early stages of training, exhibiting distortions. After 250 iterations, the PSNR value improved to 13.2 dB. The DualGAN produced slightly better image quality than the SSA-GAN, with a PSNR of 16.6 dB at later stages of training. The composite CycleGAN model achieved a PSNR value of 8.2 dB in the early stages of training, which improved to 19.8 dB after 250 iterations. In Figure 10b, the CGAN model showed slow improvement on the validation set, with a PSNR of only 18.9 dB after 150 iterations. DualGAN and SSA-GAN showed little difference, with DualGAN achieving a PSNR value of 30.4 dB and SSA-GAN reaching 28.9 dB after 150 iterations. CycleGAN achieved a PSNR of 21.6 dB during early validation training and improved to 34.2 dB in later stages. In summary, CycleGAN generated images with the least difference from the real images and exhibited the best quality in both the training and validation sets. After analyzing the PSNR metric, the study selected three additional metrics – Structural Similarity Index (SSIM), R-Precision (RP) curve, and Visual Semantic Similarity (VS) – to provide a comprehensive evaluation of the four networks. The comparison results are shown in Figure 11.

As shown in Figure 11a, CycleGAN performed best on the training set, with RP, VS, and SSIM values of 77.35%, 77.89%, and 77.62%, respectively. CGAN showed the worst performance across all metrics, with RP of 59.92%, VS of 57.64%, and SSIM of 61.21%. Figure 11b shows the comparison of these metrics in the validation set. SSA-GAN outperformed DualGAN on RP and VS values, achieving 86.71% and 85.98%, respectively. DualGAN achieved RP and VS values of 85.91% and 84.68%, respectively. CycleGAN outperformed all other models across all metrics, with all values exceeding 96%. The SSIM value was 98.23%, indicating minimal difference between the generated and target images and high similarity between the generated images. The VS value was 96.78%, and the RP value was 97.95%. In summary, CycleGAN demonstrated the best ability for style transfer in visual effects models, with the generated images closely matching the expected artistic effects. The results aligned well with the desired outcomes for the image style transfer task.

To further verify the application effect of the improved CycleGAN model proposed in the research in practical applications. The study introduced StyleGAN3, TransGAN, U-GAT-OIT, and CycleGAN+ transformer hybrid models for comparative experiments. The experiments were conducted on the same dataset and hardware environment to evaluate each model's performance. The study adopted MOS testing to subjectively evaluate the quality of the generated images. Fifteen visual experts were invited to score the images generated by each model on a scale of 1 to 5. The higher the score, the higher the quality of the generated image, and the more satisfactory the visual realism and detail retention are.

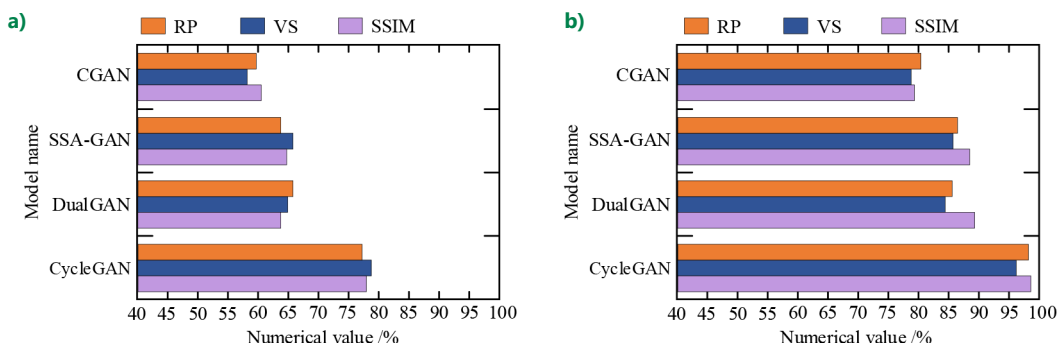


Figure 11. Comparison of four models in three indicators: a) Comparison of four models in training set; b) Comparison of four models in validation set

Table 2. Comparison of the practical application effects of each model

Model	Urban street scene style			Interior design		
	MOS	Training time (s)	GPU memory usage rate (%)	MOS	Training time (s)	GPU memory usage rate (%)
Improve CycleGAN	4.85±0.10*&##	3.47±0.85*&##	62.33±2.35*&##	4.88±0.09*&##	3.40±0.83*&##	62.32±2.83*&##
StyleGAN3	4.06±0.08	4.74±0.78	75.18±2.46	4.03±0.08	4.78±0.81	75.19±2.45
TransGAN	4.30±0.12	4.00±0.90	70.08±2.57	4.34±0.11	4.02±0.82	70.10±2.33
U-GAT-OIT	3.83±0.13	5.06±0.77	78.54±2.49	3.80±0.12	5.13±0.80	78.68±2.46
CycleGAN+ transformer hybrid model	4.11±0.10	4.58±0.84	72.36±2.50	4.14±0.09	4.60±0.79	72.25±2.70

Note: In Table 2, * indicates a significant difference between the results of Improved CycleGAN and StyleGAN3, $p < 0.05$; & indicates that the results of Improved CycleGAN and TransGAN are significantly different, $p < 0.05$; % indicates a significant difference between the Improved CycleGAN result and the U-GAT-OIT result, $p < 0.05$; # indicates a significant difference between the results of Improved CycleGAN and those of CycleGAN+ transformer hybrid model, $p < 0.05$.

In addition, training time and GPU memory usage rate have been introduced as indicators of efficiency. The experiment uses two scenarios: one is the style transfer of urban street scenes, and the other is image conversion for interior design. The results are shown in Table 2.

As shown in Table 2, the improved CycleGAN achieved the highest MOS scores in both scenarios, at 4.85 and 4.88, respectively, significantly outperforming other comparison models. At the same time, it performs best in terms of training efficiency, with the shortest training time and the lowest GPU memory usage rate, at only about 62.3%. This indicates that the proposed improvement strategy effectively enhances the generation quality and resource utilization efficiency of the model. Especially in tasks such as urban street scenes and interior design, where complex textures and structural layouts are equally important, the improved CycleGAN demonstrates a stronger ability to restore details and style coordination.

Ablation experiments were conducted to verify the effectiveness of each component applied in the research. During the experiment, CycleGAN was used as the baseline model, and the multi-head self-attention mechanism (Component A), spectral normalization (Component B), and cascade fusion (Component C) were gradually introduced to evaluate the contribution of each component to model performance. The evaluation indicators of the experimental process are MOS, PSNR, SSIM, and RP. The results are shown in Table 3.

Table 3. Comparison of ablation experiment results

Model	MOS	PSNR (dB)	SSIM (%)	RP (%)
CycleGAN+A	4.32	26.84	82.11	84.20
CycleGAN+B	4.30	26.11	82.56	84.11
CycleGAN+C	4.33	26.72	82.30	84.21
CycleGAN+A+B	4.52	26.30	87.34	89.35
CycleGAN+A+C	4.50	26.18	86.92	89.28
CycleGAN+B+C	4.51	26.20	87.00	89.70
CycleGAN+A+B+C	4.89	33.52	91.25	96.37

As shown in Table 3, the complete model with all components introduced achieved the best performance in all indicators. The MOS score reached 4.89, the PSNR increased to 33.52 dB, the SSIM reached 91.25%, and the RP reached 96.37%. Compared with the baseline, the synergy of each component significantly enhanced the feature capture and image reconstruction capabilities, verifying the effectiveness and necessity of the multi-head self-attention mechanism, spectral normalization, and cascading fusion module.

During the experiment, there were still failures, such as texture blurring and edge artifacts. Even under high-contrast lighting conditions, the model's ability to capture details remained limited. In some indoor scenes, the style transfer is inconsistent in areas with dense decorative elements, leading to visual distortion. These limitations mainly stem from the generator's insufficient trade-off between local textures and global structures, indicating that there is still room for further research to optimize the coverage of the attention mechanism and the fusion strategy.

To prove that the method proposed in the research not only performs well in the transfer of Western oil painting styles, but also can achieve a good transfer effect in the ink-wash painting of Eastern aesthetic styles. The study further selected 100 traditional Chinese ink-wash paintings as the target style set to conduct a cross-cultural style transfer test on the model. The Dataset selected is the Shan Shui and Ink Painting Multi-style Dataset (SUM-CH), and the URL of this dataset is <https://github.com/LingDong-/shan-shui-inf>. The dataset encompasses a variety of subjects, including landscapes, flowers and birds, and figures, and covers ink-wash painting styles from different historical periods and schools, making it highly artistic and diverse. Compare the MOS, inference time, pixel accuracy, and classification accuracy of each model. Pixel accuracy and classification accuracy respectively, measure the model's performance in terms of style texture retention and content structure consistency. The pixel accuracy formula is the ratio of correctly predicted pixels to the total number of pixels. The classification accuracy is

Table 4. The performance of each model in the Chinese ink-wash style dataset

Model	MOS	Reasoning time (s)	Pixel accuracy rate (%)	Classification accuracy rate (%)
Improve CycleGAN	4.88±0.09*&%@	1.01±0.23*&%@	93.45±1.05*&%@	94.84±0.78*&%@
StyleGAN3	4.00±0.10	1.83±0.30	88.45±1.32	89.34±0.74
TransGAN	4.15±0.12	1.56±0.28	90.02±1.45	91.20±0.75
U-GAT-OIT	3.67±0.12	2.00±0.24	88.15±1.08	88.75±0.88
CycleGAN+ transformer hybrid model	4.08±0.10	1.78±0.30	89.73±1.34	91.00±0.89

Note: In Table 4, * indicates a significant difference between the results of Improved CycleGAN and StyleGAN3, $p < 0.05$; & indicates that the results of Improved CycleGAN and TransGAN are significantly different, $p < 0.05$; % indicates a significant difference between the Improved CycleGAN result and the U-GAT-OIT result, $p < 0.05$; @ indicates a significant difference between the results of Improved CycleGAN and those of CycleGAN+ transformer hybrid model, $p < 0.05$.

calculated by extracting features through the pre-trained VGG-19 network and then performing style classification. The results are shown in Table 4.

As shown in Table 4, the improved CycleGAN performs best in both MOS scores and classification accuracy, achieving a pixel-level accuracy of 93.45%, indicating that it outperforms other models in preserving the details of ink-wash texture. Although StyleGAN3 has the shortest reasoning time, its generated results exhibit clear distortions in brushstroke coherence and in the sense of ink color gradation. In contrast, although the CycleGAN+Transformer hybrid model introduces an attention mechanism, it fails to effectively capture long-distance semantic associations in complex landscape paintings, resulting in a relatively low classification accuracy.

The research selected five styles – landscape ink wash, flower-and-bird ink wash, figure ink wash, landscape oil painting, and portrait oil painting – for cross-style transfer experiments to verify the model's generalization across different art forms. The sample size for each style is 100. The confusion matrix and loss of the method proposed in the study are shown in Figure 12.

As shown in Figure 12, the confusion matrix indicates that the misjudgment rate for each style is all below 8%,

	Landscape ink wash painting	Flower-and-bird ink wash painting	Figure oil painting	Landscape oil painting	Figure ink wash
Landscape ink wash painting	99	0	0	1	0
Flower-and-bird ink wash painting	1	98	0	0	1
Figure oil painting	1	1	97	1	0
Landscape oil painting	1	1	1	96	1
Figure ink wash	1	2	1	1	95

Figure 12. Confusion matrix result

suggesting that the model has a good ability to distinguish different artistic styles. Among them, the cross-recognition accuracy rate of landscape ink wash and flower-and-bird ink wash is relatively high, indicating that there are certain commonalities in the structure of brush and ink and composition layout between the two. A 5-level Likert scale was designed for 30 participants to subjectively evaluate the artistry, style consistency, and visual naturalness of the generated images, with a score of 100. The higher the score, the better the evaluation. And supplement the frame rate test on the NVIDIA Jetson TX2 platform. The results show that the improved CycleGAN scored 94.6 points in the artistic score, with style consistency and visual naturalness reaching 95.2 and 93.8 points, respectively, significantly outperforming other comparison models. Achieve an inference speed of 17.3 FPS on NVIDIA Jetson TX2, meeting the real-time generation requirements of edge devices.

The pseudo-code of the proposed method is shown in Table 5.

Table 5. Algorithm of the proposed method

Attentive CycleGAN with Spectral Normalization
<pre>// 1. Network component definition Generator: Input: x (image tensor) Output: y (transformed image tensor) Steps: 1. Encoder downsampling for multi-scale feature extraction 2. Residual block processing characteristics 3. The multi-head self-attention module processes the feature map Calculate the Q, K, and V matrices - Perform scaling dot product attention Multi-head feature splicing and projection 4. The decoder upsamples and concatenates the features of each layer 5. Multi-level feature fusion (stitching + average pooling) 6. Output the generated image Discriminator: Input: x (image tensor) Output: probability (real/fake) Steps: 1. Feature extraction through convolution downsampling 2. Spectral normalization is applied to all layers</pre>

End of Table 5

Attentive CycleGAN with Spectral Normalization
3. Output the discrimination result // 2. Training process for each training iteration: // Forward loop fake_Y = Generator_G(real_X) cycled_X = Generator_F(fake_Y) // Backward loop fake_X = Generator_F(real_Y) cycled_Y = Generator_G(fake_X) // Loss calculation adv_loss = adversarial_loss(fake_Y, real_Y) cycle_loss = cycle_consistency_loss(cycled_X, real_X) identity_loss = identity_loss(Generator(real_X), real_X) // Backpropagation and optimization total_loss = adv_loss + λ_{cycle} * cycle_loss + $\lambda_{\text{identity}}$ * identity_loss optimize (generators, discriminators, total_loss)

5. Conclusions

The current urban street environment design is too simple. However, images generated by traditional neural networks suffer from issues like sharp edge transitions and image distortion, which cannot enrich the artistic visual effects of the street environment. Therefore, the study proposed optimizing CycleGAN by incorporating multi-head self-attention modules into its network structure and outputting generated images via concatenation. Spectral normalization was also introduced to improve the neural network's design. The experimental results showed that after 600 iterations on the training set, the CycleGAN model's MSE was only 0.12 px². At the same time, the pixel accuracy of the generated images reached 96.8%, with an average classification accuracy of 97.4%, outperforming the comparison networks. In the FID metric comparison, after 600 training iterations, the CycleGAN model's value decreased to 23.3. Next, in the self-built dataset, the CycleGAN model's PSNR increased from 21.6 dB to 34.2 dB. In the comprehensive comparison of three metrics, the CycleGAN model also outperformed the comparison models, with SSIM reaching 98.23%, VS at 96.78%, and RP at 97.95%. These results indicated that the CycleGAN model could effectively recognize image styles, generating distinctive image styles and enriching the artistic visual effects of the street environment. However, the visual effect model was built within the English computer language system, which limited its ability to generate complex scene images. This issue can be optimized in future designs.

Conflict of interest

The authors have no relevant financial or non-financial interests to disclose.

Data availability statement

The datasets used and/or analyzed during the current study are available from the corresponding author upon reasonable request.

References

- Bargshady, G., Zhou, X., Barua, P. D., Gururajan, R., Li, Y., & Acharya, U. R. (2022). Application of CycleGAN and transfer learning techniques for automated detection of COVID-19 using X-ray images. *Pattern Recognition Letters*, 153(5), 67–74. <https://doi.org/10.1016/j.patrec.2021.11.020>
- Chai, X., Wang, Y., Chen, X., Gan, Z., & Zhang, Y. (2022). TPE-GAN: Thumbnail preserving encryption based on GAN with key. *IEEE Signal Processing Letters*, 29(4), 972–976. <https://doi.org/10.1109/LSP.2022.3163685>
- Chen, Y., Xia, R., Zou, K., & Yang, K. (2023). RNON: Image inpainting via repair network and optimization network. *International Journal of Machine Learning and Cybernetics*, 14, 2945–2961. <https://doi.org/10.1007/s13042-023-01811-y>
- Elpus, K. (2022). Access to arts education in America: The availability of visual art, music, dance, and theater courses in US high schools. *Arts Education Policy Review*, 123(2), 50–69. <https://doi.org/10.1080/10632913.2020.1773365>
- Fan, Y., Li, H., & Sun, B. (2024). Cycle GAN-MF: A cycle-consistent generative adversarial network based on Multifeature fusion for pedestrian re-recognition. *Transactions on Science and Engineering*, 2(1), 37–44.
- Iraji, M. S., Tanha, J., Balafar, M. A., & Feizi-Derakhshi, M.-R. (2024). Image classification with consistency-regularized bad semi-supervised generative adversarial networks: A visual data analysis and synthesis. *The Visual Computer*, 40(10), 6843–6865. <https://doi.org/10.1007/s00371-024-03360-z>
- Iraji, M. S. (2025a). A novel wavelet-transformer discriminator for semi-supervised GANs with controlled regularization and ensemble techniques. *Multimedia Tools and Applications*, 84, 49355–49395. <https://doi.org/10.1007/s11042-025-21103-7>
- Iraji, M. S. (2025b). Semi-supervised generative adversarial networks for imbalanced skin lesion diagnosis with an unbiased generator and informative images. *Engineering Applications of Artificial Intelligence*, 159(1), Article 111643. <https://doi.org/10.1016/j.engappai.2025.111643>
- Jia, C., He, H., Zhou, J., Li, K., Li, J., & Wei, Z. A. (2024). Performance degradation prediction model for PEMFC based on bi-directional long short-term memory and multi-head self-attention mechanism. *International Journal of Hydrogen Energy*, 60, 133–146. <https://doi.org/10.1016/j.ijhydene.2024.02.181>
- Kammoun, A., Slama, R., Tabia, H., Ouni, T., & Abid, M. (2022). Generative adversarial networks for face generation: A survey. *ACM Computing Surveys*, 55(5), 1–37. <https://doi.org/10.1145/3527850>
- Kumar, A., Fahad, M., Arif, H., & Hussain, H. K. (2023). Navigating the uncharted waters: Exploring challenges and opportunities in block chain-enabled cloud computing for future research. *Bullet: Jurnal Multidisiplin Ilmu*, 2(6), 1297–1305.
- Lakshmi, L., Kalyani, A. N., Madhuri, D. K., Potluri, S., Pandey, G. S., Ali, S., & Ismail, E. A. (2023). Performance analysis of cycle GAN in photo to portrait transfiguration using deep learning optimizers. *IEEE Access*, 11, 136541–136551. <https://doi.org/10.1109/ACCESS.2023.3337430>
- Liu, X., Zhang, T., Zhang, J. (2023). Toward visual quality enhancement of dehazing effect with improved Cycle-GAN. *Neural Computing and Applications*, 35(7), 5277–5290. <https://doi.org/10.1007/s00521-022-07964-1>
- Logeswaran, J., Kiruthiga, T., Aravindarajan, V., & Ravi, S. P. (2022). SVPA – the segmentation based visual processing algorithm (SVPA) for illustration enhancements in digital video processing (DVP). *ICTACT Journal on Image and Video Processing*, 12(3), 2669–2673. <https://doi.org/10.21917/ijivp.2022.0379>

- Lu, E., & Hu, X. (2022). Image super-resolution via channel attention and spatial attention. *Applied Intelligence*, 52, 2260–2268. <https://doi.org/10.1007/s10489-021-02464-6>
- Mao, W., Yang, S., Shi, H., Liu, J., & Wang, Z. (2022). Intelligent typography: Artistic text style transfer for complex texture and structure. *IEEE Transactions on Multimedia*, 25, 6485–6498. <https://doi.org/10.1109/TMM.2022.3209870>
- Reddi, P., Srinivas, G., Reddy, P. P., & Krihsna, D. S. (2024). A multi-head self-attention mechanism for improved brain tumor classification using deep learning approaches. *Engineering, Technology & Applied Science Research*, 14(5), 17324–17329. <https://doi.org/10.48084/etasr.8484>
- Rosin, P. L., Lai, Y.-K., Mould, D., Yi, R., Berger, I., Doyle, L., & Winnemöller, H. (2022). NPRportrait 1.0: A three-level benchmark for non-photorealistic rendering of portraits. *Computational Visual Media*, 8(3), 445–465. <https://doi.org/10.1007/s41095-021-0255-3>
- Shen, X., Han, D., Guo, Z., Chen, C., Hua, J., & Luo, G. (2023). Local self-attention in transformer for visual question answering. *Applied Intelligence*, 53, 16706–16723. <https://doi.org/10.1007/s10489-022-04355-w>
- Sun, B., Jia, S., Jiang, X., & Jia, F. (2023). Double U-Net CycleGAN for 3D MR to CT image synthesis. *International Journal of Computer Assisted Radiology and Surgery*, 18(1), 149–156. <https://doi.org/10.1007/s11548-022-02732-x>
- Wang, Y., Hu, W., Wen, L., & Gao, L. (2023). A new foreground-perception cycle-consistent adversarial network for surface defect detection with limited high-noise samples. *IEEE Transactions on Industrial Informatics*, 19(12), 11742–11751. <https://doi.org/10.1109/TII.2023.3252410>
- Wen, L., Wang, Y., & Li, X. (2022). A new cycle-consistent adversarial networks with attention mechanism for surface defect classification with small samples. *IEEE Transactions on Industrial Informatics*, 18(12), 8988–8998. <https://doi.org/10.1109/TII.2022.3168432>
- Yang, S., Jiang, L., Liu, Z., & Loy, C. C. (2022). Vtoonify: Controllable high-resolution portrait video style transfer. *ACM Transactions on Graphics (TOG)*, 41(6), 1–15. <https://doi.org/10.1145/3550454.3555437>