



International Journal of Information and Communication Technology

ISSN online: 1741-8070 - ISSN print: 1466-6642

<https://www.inderscience.com/ijict>

Artistic image restoration and semantic reconstruction driven by multimodal AIGC

Huiling Huang

DOI: [10.1504/IJICT.2026.10076562](https://doi.org/10.1504/IJICT.2026.10076562)

Article History:

Received:	09 September 2025
Last revised:	04 December 2025
Accepted:	05 December 2025
Published online:	10 March 2026

Artistic image restoration and semantic reconstruction driven by multimodal AIGC

Huiling Huang

College of Art Engineering,
Guangxi Vocational College of Performing Arts,
Nanning, 530226, China
Email: H18007800408@163.com

Abstract: Generative artificial intelligence technology provides new insights for art image inpainting and semantic reconstruction. To address the problem of semantically incorrect restoration content in existing research, this paper first optimises generative adversarial network by combining gated convolution with spectral normalisation. Based on this, an image inpainting and semantic reconstruction method is built by integrating text and art image features. The text disentanglement module of the suggested method can obtain key textual features that help restoration. A cross-modal attention module is designed to ensure restored results are as consistent as possible with text semantics. A dual-channel reconstruction module is also designed to enhance the network's ability to predict image structure and text semantics. Experimental results show that the fr chet inception distance (FID) of the proposed method is 3.01, which can restore realistic art images satisfying textual semantics.

Keywords: image restoration; semantic reconstruction; generative adversarial networks; multimodal features; cross-modal attention.

Reference to this paper should be made as follows: Huang, H. (2026) 'Artistic image restoration and semantic reconstruction driven by multimodal AIGC', *Int. J. Information and Communication Technology*, Vol. 27, No. 17, pp.18–34.

Biographical notes: Huiling Huang received her Bachelor's degree from Guangxi Arts University in 2005 and a Master's degree in Software Engineering from Xiamen University in 2016. She is currently an Associate Professor at Guangxi Vocational College of Performing Arts. Her research fields include art design, vocational education and image restoration.

1 Introduction

Art images carry historical, cultural, and aesthetic values. However, due to factors such as time erosion and human damage, many precious art images have issues like fading, missing parts, mould deterioration, posing a risk of permanent loss of their original appearance and deep semantic information (Wang et al., 2020). Traditional art image restoration methods mainly rely on the experience and skills of manual restorers; they are not only time-consuming and labour-intensive with low repair efficiency but also difficult to accurately restore the historical context and artistic characteristics of images (Chen et al., 2023). Early computer-aided restoration technologies, limited by single-modal data

processing capabilities, often had limitations in meeting high semantic consistency restoration needs (Zhang et al., 2018). In recent years, intelligent technologies centred on generative artificial intelligence. Artificial intelligence generated content (AIGC) have developed rapidly, especially making breakthrough progress in the field of multimodal fusion. Multimodal AIGC technology can integrate multi-dimensional information such as text descriptions and image features through cross-modal semantic understanding and generation models to achieve accurate modelling and creative generation of complex content (Kumar and Gupta, 2023). Through methods like text semantics guidance, it is possible to deeply mine the underlying characteristics of images so that the restored images are not only visually complete but also accurately convey the cultural connotations of the original artworks.

Art image restoration and semantic reconstruction technology is a technique for predicting and filling in areas where images and their semantics are missing. Early art image restoration methods were mainly divided into diffusion-based approaches (Barbu and Favini, 2014) and block-matching-based approaches (Liang et al., 2021). Yue et al. (2024) completed the image by diffusing pixels along the direction field of isocontours, estimating the isophote line directions at the boundaries between intact and damaged regions of the image information, then propagating effective information in the inferred directions until the repair region was fully restored. Bhele et al. (2023) proposed the PatchMatch algorithm, which uses probabilistic thinking to quickly match and fill the most similar image regions in the context for completion, thereby achieving image restoration and semantic reconstruction. Lyu et al. (2024) proposed a repair method that combines anisotropic diffusion processes with enhanced shock filling of edges and corners, improving the quality of image restoration in missing regions. Zha et al. (2020) reduced the error filling rate of block matching by introducing a dispersion term. However, block-matching-based methods, which fill damaged areas of art images only through block copying and pasting, are prone to causing misalignment of structural information in the restored art images.

Although the above methods can quickly and effectively solve some image restoration problems, they perform very poorly in terms of visual semantics (Zhai et al., 2023). This is because traditional algorithms are based on statistical information from visible parts of images. They only search at the pixel level context for filling, failing to delve into feature-level aspects of the image, thus naturally struggling to understand its visual semantics. Deep learning-based image restoration methods use deep neural network models to automatically learn characteristics of art images in order to fill missing parts, thereby achieving more realistic restorations in terms of visual semantics. Artistic image restoration not only requires consideration of the restoration of local textures, but also needs to take into account the surrounding structures, the shape of the entire object, and the spatial relationship of the scene. Deep learning methods work simultaneously at different scales. At small scales, they focus on the fineness of textures and the sharpness of edges; at large scales, they grasp the position, posture, and compatibility with the background of objects. Main deep learning restoration approaches include encoder-decoder based restoration methods (Wang and Tao, 2016) and AIGC-based restoration methods (Ma et al., 2022). Zhang et al. (2021) proposed an art image restoration and semantic reconstruction method based on a hybrid autoencoder model, which first divides the art images into blocks of regular size and then uses a convolutional neural network (CNN) and vision transformer (ViT) integrated network to restore them, thereby enhancing the ability to reconstruct visual semantics.

Amaranageswarao et al. (2022) proposed a two-stage restoration model based on hypergraph convolution that employs hypergraph convolution to enhance feature capturing capacity; however, this approach does not consider constraints for global semantic feature consistency, leading to results where boundary effects and blurred semantic content often appear. Chen et al. (2022) proposed an image-assisted restoration model based on Transformer; however, this method only uses local information as constraints and lacks global semantic consistency constraints, leading to the problem of missing local details in the restored results. Generative adversarial networks, as a typical algorithm of AIGC (Zhan et al., 2023), learn the prior distribution of original mural images through generators, then complete restoration by adversarial training between discriminators and generators. Ren et al. (2024) first regarded image restoration tasks as constrained conditional image generation problems and proposed co-modulated generative adversarial networks. Due to a lack of spatial adaptability, co-modulation methods are sensitive to invalid encoded features within holes. Kumar et al. (2024) proposed a generative adversarial network (GAN)-based art image restoration and semantic reconstruction model that restores images at both the image level and feature level, ensuring consistency in texture and semantics.

The above studies only consider single-modal image features, leading to poor quality in image restoration and semantic reconstruction. Single-modal methods only rely on the pixel information of RGB images and lack an understanding of deep features such as image depth, structure, and semantics. For instance, when restoring artistic images with complex textures, it is difficult to accurately infer the details of the missing parts merely based on pixel information. Multimodal methods can extract richer and more comprehensive image features by fusing data from different modalities. These features not only include pixel-level information, but also deep-level features such as depth, structure and semantics. This helps the model to understand the content and structure of the image more accurately, thereby generating more reasonable restoration results. Art images also contain substantial text information; the core issue in achieving text-guided image restoration is how to properly handle fusion between text and images. The purpose of fusion is to fill missing semantics into damaged images with semantically relevant content from texts, and the challenge lies in accurately filling these semantics (Wang et al., 2023). Marivani et al. (2022) proposed a cross-modal restoration network that gradually optimises generated results from coarse to fine and uses text-to-image and image-to-text training modes to generate clear and natural images. Wang et al. (2024) proposed conditional GAN (CGAN) to generate high-quality images, where CGAN introduced a dynamic memory module to process the initial low-resolution images and designed a memory write gate that can dynamically select words related to the image to ensure that the generated images match the text descriptions. Chen et al. (2025) proposed a multimodal image restoration and semantic reconstruction network model, which takes textual descriptions as guidance and consists of a repair module, a visual-text fusion module, and a multimodal semantic alignment module, focusing on effective fusion between modalities and semantic consistency during the text-guided image restoration process.

In the field of artistic image restoration, regarding artworks, blurred edges and inconsistent semantics in the restoration results have always been a severe challenge. AIGC provides practical solutions to address this challenge. Even if multimodal methods introduce text descriptions, the artistic intentions they understand are often superficial labels such as style and objects. For the deeper emotional expression of artworks, the

spirit of The Times, philosophical metaphors, and the accidental brushstrokes of artists during the creative process, which are advanced and abstract value dimensions, deep learning methods find it difficult to capture and reflect them. To resolve the issue where existing research lacks an understanding of deep semantics, leading to semantically incorrect content in restored artistic images, this paper proposes a method for art image restoration and semantic reconstruction driven by multimodal AIGC. First, to address the problem that GANs are prone to mode collapse and training collapse, this paper combines gated convolution with spectral normalisation to optimise the GAN. By introducing gated convolution to effectively distinguish different region information during the feature extraction phase, and incorporating spectral normalisation mechanisms into the discriminator network to constrain parameter matrices for optimising the training process. Based on this, a method for art image restoration and semantic reconstruction that integrates multimodal features (MSGAN) is proposed. The text disentanglement module of MSGAN can utilise information inherent in the damaged artistic images to disentangle textual descriptions and obtain key textual features helpful for restoration. This module enables the network to focus more on important information from the textual description, ensuring that the restoration results align as closely as possible with the text semantics. The designed image-semantic dual-channel reconstruction module can enhance the network's ability to predict image structures and textual semantics, improving the level of detail in restoration results. This paper conducts qualitative and quantitative experimental comparisons between MSGAN and existing methods on real datasets, demonstrating the effectiveness of the proposed approach, and verifies and analyses the design rationality and effectiveness of the modules through multiple ablation experiments.

2 Relevant technologies

2.1 Principles of image restoration

Image inpainting is a technique that automatically fills missing or damaged regions of an image using algorithms, with the core principle being to infer and reconstruct the missing parts by utilising surrounding pixels, textures, structures, etc., making the restored image visually and semantically natural and coherent. The most common form of image inpainting is to fill in the information already missing from the image through certain methods (Petrovici et al., 2018). Therefore, when repairing images with missing areas, it's not only necessary for the filled part to have similar texture details as the existing background, but also for the generated information to be semantically consistent with the background information (Zhao et al., 2023). This requires the image inpainting process to ensure that the restored image has detailed textures, coherent colour styles, and logical semantics overall.

In real-world scenarios, images containing large amounts of information are prone to damage or loss during processes such as acquisition, storage, and transmission due to uncontrollable factors or improper storage. This process is called image degradation and can be represented by the following equation.

$$P_M = P_0 \odot K, P_M \in (\Omega \cup \Phi) \quad (1)$$

where P_M represents the image under repair, M is the image mask, expressed as a 2D matrix with values of 0 or 1, missing areas Ω are assigned 0, existing areas Φ are assigned 1, and P_0 is the input image. Typically, designed image inpainting models are used to repair digital images. P denotes the output image, P_r represents the repaired region generated by the image inpainting model, overlapping the input image and the region generated by the image inpainting model results in a complete restored image.

$$\bar{P} = P_0 \odot K + P_r \odot (1 - K) \quad (2)$$

In real scenes, images may simultaneously have multiple degradation types such as noise, blurriness, loss, colour distortion, and compression artefacts, and the degree of degradation varies. Traditional single-task models have difficulty handling such situations where multiple degenerations are superimposed. For instance, when fuzziness and noise coexist, a single denoising or defuzzing model may not be able to optimise both types of degenerations simultaneously. As the image resolution increases, the difficulty of restoring texture details also rises. Simple interpolation can easily lead to blurring or breaking, while the recovery of high-frequency information requires the model to have a more refined feature extraction capability. Existing research visualises the repair process through attention mechanisms or introduces conditional generative adversarial networks, allowing users to control the repair style by inputting instructions. For instance, in the restoration of old photos, users can adjust the colour fidelity parameters to balance the sense of history and clarity.

2.2 Generative adversarial network

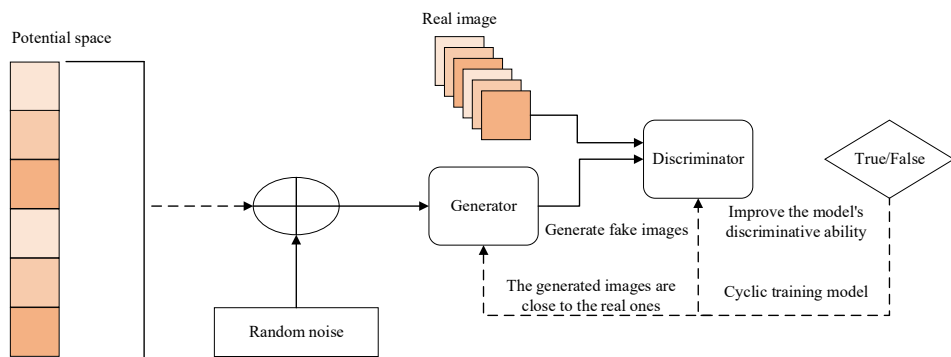
Unlike other networks in deep learning, the GAN network consists of two parts: a generative model (G) and a discriminative model (D), commonly referred to as the generator and discriminator (Aggarwal et al., 2021). It was proposed based on the Nash equilibrium in game theory. The principle is to allow the generator and discriminator to engage in adversarial games; through adversarial training, both networks' capabilities are enhanced. D enables G to learn to generate more realistic images even without a large amount of prior knowledge or distribution information. Ultimately, it allows the generated image from the model to achieve an indistinguishable between real and fake effects, meaning that D cannot differentiate the predicted image from the original image. Therefore, the above allows G and D to reach Nash equilibrium. The GAN network model is shown in Figure 1. Compared with classic deep learning models (such as autoencoders or pure discriminative models), generative adversarial networks have fundamentally changed the paradigm of image generation through their unique adversarial training mechanism, thereby significantly improving the generation quality.

Training a GAN can be viewed as a game where the generator continuously attempts to deceive the discriminator by producing increasingly realistic data, while the discriminator continuously improves its ability to distinguish between real and fake data. This process continues until it becomes difficult for the discriminator to differentiate real data from generated data, thus achieving a training equilibrium point. Equation (3) represents the GAN training process.

$$\min_G \max_D V(D, G) = E_{x \sim p_{data}(x)} [\log D(x)] + E_{x \sim p_z(z)} [\log (1 - D(G(z)))] \quad (3)$$

where z is the input random vector with prior distribution $p_z(z)$, x is the real sample, and $p_{data}(x)$ represents the data distribution of these samples. The generator G generates samples based on the input random vector z . The judgment result $D(x)$ for real samples and the judgment result $G(z)$ for generated samples $D(G(z))$ separately represent the probability that the discriminator believes these samples are genuine. During the training process of the GAN, G and D are updated alternately. The objective of training D is to maximise its capability to correctly classify real samples and fake samples produced by G , which can be achieved by maximising the accuracy $D(x)$ for classifying real samples and maximising $1 - D(x)$ for classifying generated samples. The training goal of G is to minimise $\log(1 - D(G(z)))$, meaning that G attempts to generate fake samples $G(z)$ which can be misidentified as real samples by D , thereby deceiving D .

Figure 1 The GAN network model (see online version for colours)



3 Optimisation of generative adversarial network based on gate convolution and spectral normalisation

3.1 Generator network based on gated convolution

In image inpainting and semantic reconstruction methods based on AGIC, the GAN is a commonly used generative network. However, in practical training, the GAN often experiences the discriminator reaching an ideal state too early, meaning the discriminator can always accurately identify whether the currently input image is real or fake, resulting in the generator's parameters being unable to be updated in time and effectively, thereby affecting the iterative training of the model and ultimately causing the generator to experience a near-zero gradient disappearance. To address the above issues, this chapter proposes optimising GAN (GSGAN) by combining gated convolution and spectral normalisation. First, we introduce gated convolution (Liu and Laili, 2018) into the generator network to replace traditional convolution; through the dynamic selection mechanism of gated convolution that can learn feature information from different positions and channels in images, effective differentiation of missing image regions is achieved. Second, a spectral normalisation mechanism is introduced into the discriminator network to constrain the spectral norm of parameter matrices for each layer in the discriminator, thereby controlling its Lipschitz constant and ultimately avoiding gradient disappearance and gradient explosion.

Unlike traditional convolutions, gated convolutions achieve distinction among different positions and channels in feature maps through their unique dynamic feature selection technique. While selecting features based on background, masks, and sketches, they also consider semantic segmentation across specific channels. Even in deep networks, gated convolution can still learn to highlight masked areas and sketch information for guiding the generation results. The generator network design draws on the encoder-decoder concept, consisting of an encoder, a middle part, and a decoder. The encoder is responsible for down-sampling the input image; to address the issue where traditional convolution cannot distinguish between missing regions and known region information during feature extraction, we use gated convolutions with self-learning soft mask mechanisms in place of traditional convolutions within the encoder. During feature extraction, the gated convolution dynamically selects based on background, masks, and sketches while also considering semantic segmentation across different channels to further improve restoration quality. The down-sampled feature information is processed through the middle part to extract low-dimensional abstract features, which consists of overlapping traditional convolution and dilated convolution layers that work together to further extract information from the down-sampled feature map. Finally, the decoder performs upsampling on these features to produce an initial generated image. The principle of gated convolutions is shown in equation (4) and equation (5), where W_g and W_f are weights and I represents input variables.

$$Gating_{y,x} = \sum \sum W_g \times I \quad (4)$$

$$Feature_{y,x} = \sum \sum W_f \times I \quad (5)$$

3.2 Spectral normalisation-based discriminator network

In GANs, the discriminator is responsible for evaluating the authenticity of images generated by the generator and is used during model training to guide parameter updates of the generator so that more realistic images can be produced. The discriminator network proposed in this chapter consists of a global discriminator and a local discriminator. The global discriminator takes the entire image as input, while the local discriminator takes the damaged region as input. Both are composed of convolutional neural networks and fully connected networks. By combining the global and local discriminators, the discriminator network makes judgments about whether the current input image is real or fake. Finally, gradient information is calculated based on the tensor values output by the linear layer of the discriminator and fed back to the generator network for parameter updates. To address issues such as vanishing gradients and mode collapse, this chapter introduces spectral normalisation mechanisms into the convolutional layers of the discriminator network to solve problems that are prone to occur in generative models.

In traditional GANs, the problem of solving Wasserstein distance (Li et al., 2020) is transformed into finding optimal solutions using Lipschitz continuous functions (Balashov, 2015). Spectral normalisation technology is proposed in this paper to ensure that the parameter matrices satisfy Lipschitz constraints without destroying their structures. If a function satisfies the Lipschitz constraint, its rate of change will not exceed a fixed constant. Viewing the discriminator as a multi-layer network, the n th layer can be represented as follows: a_n is the activation function for this layer, W is the network

parameter, and b is the bias. For convenience in derivation, we omit b , so equation (6) can be rewritten as equation (7).

$$x_n = a_n (W_n x_{n-1} + b_n) \quad (6)$$

$$x_n = D_n W_n x_{n-1} \quad (7)$$

Using a diagonal matrix to represent the ReLU activation function, denoted as D_n , when inputs are negative, the diagonal elements are 0; and when inputs are positive, the diagonal elements become 1. The neural network is represented as shown in equation (8).

$$F(x) = D_N W_N \dots D_1 W_1 x \quad (8)$$

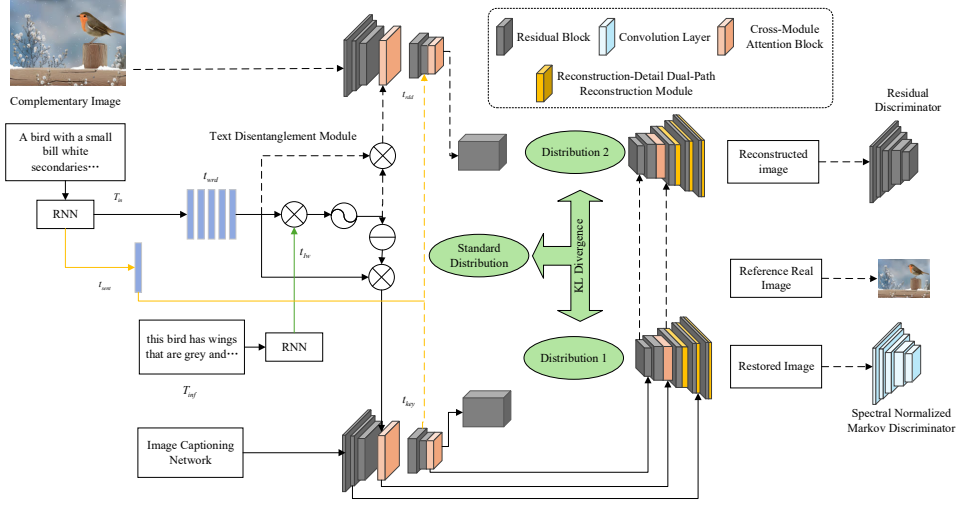
This divides each layer's network parameter matrix by its corresponding spectral norm, making the parameter matrices satisfy Lipschitz constraints with a constant of 1, achieving spectral normalisation for parameter matrices and improving the generative quality and diversity of GANs. Compared to other normalisation techniques, spectral normalisation only requires adjusting the Lipschitz constant. It is typically applicable in multi-layer neural network and CNN.

4 Art image restoration and semantic reconstruction based on GAN and multimodal features

4.1 Text decoupling module

Art images often contain specific semantic symbols and metaphorical relationships, while existing methods mostly rely on low-level visual features for inpainting, lacking understanding of these deeper semantics and easily leading to semantic logic errors in the restored content of art images. Therefore, this paper proposes an art image inpainting and semantic reconstruction method based on GSGAN and multimodal features (MGSGAN). The MGSGAN model structure is shown in Figure 2. By adopting the training idea of GSGAN, it consists of two generator networks trained in parallel and two discriminator networks; a text decoupling module is proposed to extract textual features beneficial for art image restoration, enhancing the cross-modal attention module's ability to integrate textual features and art image features. In addition, an art image-semantic dual-channel reconstruction module is proposed to enhance the prediction capabilities of art images and semantics.

During training, the repaired art image and reconstructed semantics are separately fed into two discriminators with real benchmark images, and network parameters of MGSGAN are updated through backpropagation. Dual-channel reconstruction modules are added to the last three layers of the repair network, mainly serving to enhance the model's overall prediction capabilities for art images and semantics. The image captioning network of MGSGAN is derived from reference (Gupta et al., 2021) and pre-trained on the corresponding dataset in this method, enabling it to generate textual descriptions from art images, i.e., converting art images into text. Parameters of the art image captioning network and recurrent neural network (RNN) in MGSGAN remain fixed during training and do not participate in backpropagation.

Figure 2 MSGGAN model structure (see online version for colours)

The text disentanglement module (TDM) proposed in this paper is used to decouple the textual descriptions of art images and extract information most helpful for restoration. The input of TDM consists of the textual description features of art images $t_{wrđ}$ and reasoning text features t_{lw} , where $t_{wrđ}$ originates from externally provided text descriptions T_{in} and t_{lw} originates from text descriptions inferred by the image captioning network T_{inf} . Since T_{inf} is generated from an incomplete art image, it mainly contains descriptions of visible features of the image and partial predictions of invisible feature descriptions. This differs from the information contained in T_{in} , which describes a complete art image. Therefore, if it is possible to disentangle specific descriptions of invisible features by utilising differences between their information, this would greatly benefit restoration. The key text features t_{key} and redundant text features $t_{rđđ}$ can be obtained through the following formula, where $Q(\cdot)$, $K(\cdot)$, and $V(\cdot)$ are three 1×1 convolutions with different weights.

$$t_{key} = t_{wrđ} + V(t_{wrđ}) \cdot [1 - \text{softmax}(Q(t_{wrđ}) \cdot K(t_{lw}))] \quad (9)$$

$$t_{rđđ} = t_{wrđ} + V(t_{wrđ}) \cdot \text{softmax}(Q(t_{wrđ}) \cdot K(t_{lw})) \quad (10)$$

where t_{lw} mainly describes the visible area, while $t_{wrđ}$ describes the entire artistic image. Through the information differences between these two text features, we can obtain specific textual features for the missing region, namely t_{key} . t_{key} represents the part of $t_{wrđ}$ and reasoning t_{lw} that is least similar, approximately equivalent to a real detailed description of the target restoration area. In contrast, $t_{rđđ}$ represents the most similar part between the two text features, nearly representing the actual description for the visible region with information contained in the input art image. Therefore, this method injects these two disentangled text features into different generator networks based on their distinct information and described regions. Among them, $t_{rđđ}$ is injected into the generator network responsible for reconstruction, while t_{key} is used by the generator network responsible for restoration.

4.2 Cross-modal attention module

Since text features and image features reside in different spatial feature spaces, a space transformation is required. The cross-modal attention block (CAB) designed in this paper is used to convert the text feature information into the art image information space, for a given image feature F_f and text feature t , as shown in equation (11).

$$s_{i,j}^{cross} = F_{fi}^T \cdot H(t_j) \quad (11)$$

where $s_{i,j}^{cross}$ represents the similarity of the i -th pixel on the art image to the j -th word in the text. F_{fi} indicates the image features corresponding to pixel i , t_j represents the textual feature information for the j -th word, and $H(\cdot)$ is a 1×1 convolution. Let N represent the number of pixels on the artistic image's feature map, then CAB scores are defined as follows.

$$\alpha_{j,i}^{cross} = \frac{\exp(s_{i,j}^{cross})}{\sum_{i=1}^N \exp(s_{i,j}^{cross})} \quad (12)$$

Finally, let L be the length of the text feature; then the output A of the CAB block can be calculated using equation (13).

$$A = \sum_{i=1}^L H(t_j) \alpha_{j,i}^{cross} \quad (13)$$

The text features obtained by MGSGAN through RNN can be divided into two types according to granularity: word features t_{word} and sentence features t_{sent} . Among them, t_{rdd} and t_{key} both belong to word features. The granularity of the word feature t_{word} is larger than that of the sentence feature t_{sent} , with stronger geometric information representation capabilities, making it more suitable for shallow images. The granularity of the sentence feature t_{sent} is smaller, with stronger semantic information representation capabilities, making it more suitable for deep images. Therefore, this chapter places CAB blocks at three locations in each of the two generators to convert different text feature information into art image feature spaces.

4.3 Dual-channel module for image restoration and semantic reconstruction

This paper proposes an image-semantic dual-channel reconstruction module (ISDRM) to enhance the prediction capabilities of image structure and semantic details. This module can be placed at any position in the decoder. The ISDRM consists of an encoding block, a decoding block, and two output layers. Among them, the encoding block and decoding block are composed of K residual blocks, responsible for the mutual conversion between deep features and shallow features without affecting resolution. Considering computational resources, K can be obtained as follows.

$$K = \log_2 \left(\frac{R_{Image}}{R_{Layer}} \right) + 1 \quad (14)$$

where R_{Image} and R_{Layer} represent the resolutions of the image and module layer respectively. Therefore, the closer the module is to the shallow network, the smaller K will be.

The two output layers of the ISDRM include a structural output layer and a detail output layer, which can generate structure images without semantics I_{sg} and inpainted images with rich textures I_{ge} . During training, $L1$ constraints are applied separately to these two outputs. Let I_{st} and I_{gt} be the ground truth of I_{sg} and I_{ge} , respectively; then the loss function is as follows.

$$L_{rec} = \|I_{sg} - I_{st}\|_1 + \|I_{ge} - I_{gt}\|_1 \quad (15)$$

ISDRM modules are incorporated into the last three layers of the decoder. Both I_{sg} and I_{ge} increase progressively with the layer index, while image structure and semantic details are progressively refined.

4.4 Loss function design

The objective function consists of several types of loss functions including reconstruction loss, adversarial loss, distribution regularisation function, and text-semantic alignment loss. Among them, the definition of the reconstruction loss is shown in equation (16). GSGAN introduces two discriminators for adversarial training: a residual discriminator and a spectral normalised Markov (SN-PatchGAN) discriminator. The generator responsible for reconstruction corresponds to the residual discriminator, using an L2 constraint to accomplish the image reconstruction task as follows.

$$L_g^{rec} = \|D^{rec}(I_{rec}) - D^{rec}(I_{gt})\|_2 \quad (16)$$

where $D^{rec}(\cdot)$ denotes the output of the residual discriminator, and I_{rec} and I_{gt} are the reconstructed artistic images and ground-truth artistic images, respectively. The generator responsible for inpainting corresponds to the SN-PatchGAN discriminator, using a relative average LS adversarial loss function to measure the probability that an image is real or fake relatively, with the calculation equation as follows.

$$L_g^{adv} = E_{I_{gt}} \left[\left(D^{adv}(I_{gt}) - E_{I_{ge}} D^{adv}(I_{ge}) + 1 \right)^2 \right] \\ + E_{I_{ge}} \left[\left(D^{adv}(I_{ge}) - E_{I_{gt}} D^{adv}(I_{gt}) - 1 \right)^2 \right] \quad (17)$$

where $D^{adv}(\cdot)$ denotes the output of the SN-PatchGAN discriminator, and I_{ge} is the inpainted image. Both discriminators use relative average LS adversarial loss functions as their discriminator loss functions.

GSGAN uses a Gaussian function to define distributions, where the distribution of the generator responsible for inpainting is defined as $p(z|I_m, t_{key}, t_{sent})$, and that of the generator responsible for reconstruction is $q(z|I_c, t_{rdd}, t_{sent})$, I_m and I_c are the input image and its corresponding complementary image, respectively. The distribution regularisation function uses KL divergence to regularise the learned importance sampling function onto a fixed latent space variable. The equation is as follows.

$$L_{KL} = -KL\left(p\left(z|I_m, t_{key}, t_{sent}\right)\|N(0, 1)\right) - KL\left(p\left(z|I_m, t_{key}, t_{sent}\right)\|q\left(z|I_c, t_{rdd}, t_{sent}\right)\right) \quad (18)$$

The text-semantic alignment loss function adopts the loss function from the deep attention multi-modal similarity model (DAMSM), used to constrain the consistency between the output artistic image and its descriptive text, defined as follows.

$$L_{TSA} = DAMSM(I_{rec}, T_{in}) + DAMSM(I_{ge}, T_{in}) \quad (19)$$

where T_{in} denotes the externally input textual description. Let λ_r , λ_g , λ_{KL} , and λ_{TSA} be hyperparameters for adjusting each loss function; then, the total objective loss function can be expressed by the following equation.

$$L = \lambda_r \sum_{i=1}^3 L_{rec}^i + \lambda_g (L_g^{rec} + L_g^{adv}) + \lambda_{KL} L_{KL} + \lambda_{TSA} L_{TSA} \quad (20)$$

5 Experiment and performance analysis

5.1 Analysis of art image restoration and semantic reconstruction results

The experimental platform of this paper is the Centos 7 system, using the Pytorch deep learning framework. The GPU is an NVIDIA TESLA V100 SXM2 with 32 GB of video memory, and the CUDA version is 11.4. To verify the effectiveness of the proposed method in this paper, experiments were conducted on the open Chinese traditional landscape painting dataset from Chinese-Landscape-Painting. This dataset contains 2,192 artistic images, including Chinese landscape paintings and artistic images by famous foreign painters. Each sample image corresponds to ten sentences as textual descriptions. During training, the dataset was divided into a training set, validation set, and test set at a ratio of 6:3:1. All input artistic images were adjusted to a resolution of 256×256 . The initial learning rate is 2×10^{-4} , with a total of 120 training epochs and a batch size of 32.

Four evaluation metrics are introduced in the experiments to examine the restoration effects for artistic image inpainting and semantic reconstruction, which include fr chet inception distance (FID), Kernel inception distance (KID), peak signal-to-noise ratio (PSNR), and structural similarity index (SSIM) (Liu et al., 2025). This paper first analyses the impact of masking rate on model performance. Figure 3 and Figure 4 show that when the masking rate is within the low occlusion range (0.01, 0.1], MGSGAN performs excellently in the PSNR and SSIM metrics, achieving approximately 37 dB and 0.97 respectively, indicating its strong ability to restore images with mild damage. As the masking rate increases to (0.1, 0). Even as the repair difficulty increases, the model's average PSNR remains above 26 dB and SSIM stays above 0.96, indicating its strong robustness for moderately damaged artistic images. At the same time, observing performance changes during training shows that when training reaches approximately 50 epochs, both PSNR and SSIM metrics basically stabilise, with model performance gradually becoming saturated; further training contributes little to improvement. This indicates that under the current network architecture and training configuration, the

model has already achieved stable and efficient restoration capabilities and demonstrates good adaptability and practicality for processing images of varying degrees of damage.

Figure 3 PSNR results at 10% and 20% mask rates in the test set (see online version for colours)

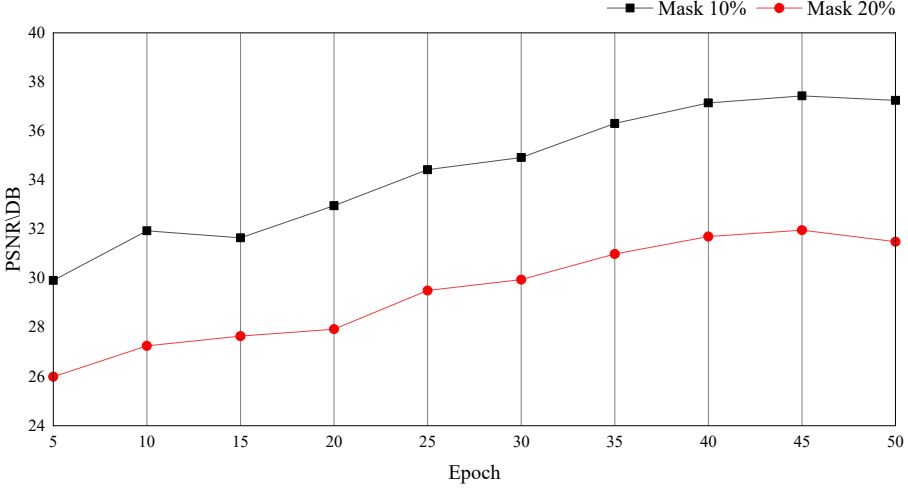
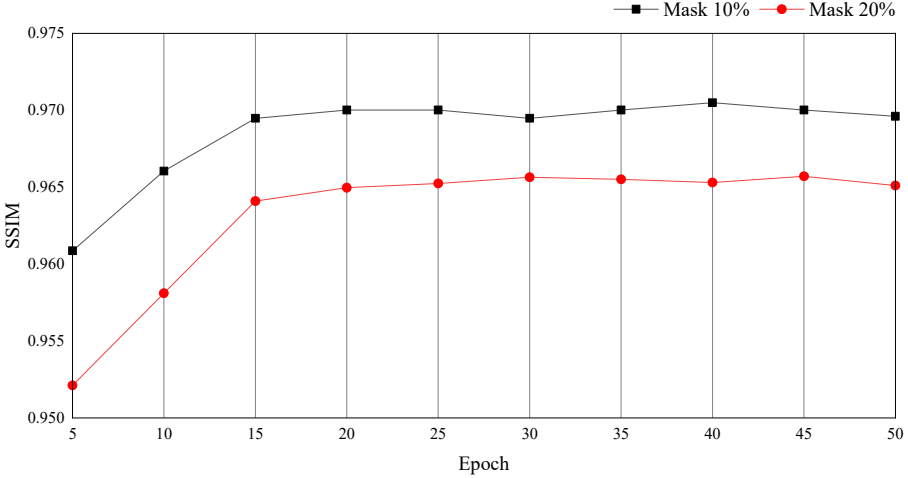


Figure 4 SSIM results for 10% and 20% mask coverage in the test set (see online version for colours)



5.2 Comparison of experimental results and analysis

To comprehensively compare MGSGAN with other methods, Table 1 lists quantitative comparison results among the MGSGAN model, SC-VAE (Zhang et al., 2021), CAFormer (Chen et al., 2022), DM-GAN (Ren et al., 2024), CGAN (Wang et al., 2024), and MDF-CL (Chen et al., 2025). In this table, ↓ indicates lower values are better while ↑ means higher values are better. MGSGAN clearly outperforms other models in both FID

and KID metrics and is very close to the optimal value in PSNR and SSIM metrics. FID and KID mainly measure visual effects perceived by human eyes, which are more relevant to semantic information of artistic images, thus they can be approximately regarded as indicators for evaluating textual description performance. In contrast, PSNR and SSIM relate more closely to baseline ground-truth images, measuring similarity with such reference images. The FID and KID values of MGSGAN are 3.01 and 0.09, achieving at least 24.37%, 35.71% reduction compared to the baseline models respectively. The PSNR and SSIM of MGSGAN are improved by at least 6%, 4.27% over the baseline model. This indicates that the artistic images restored by MGSGAN demonstrate significantly good visual effects to human eyes, with high integrity of image semantics.

Figure 5 The ranking distribution (see online version for colours)

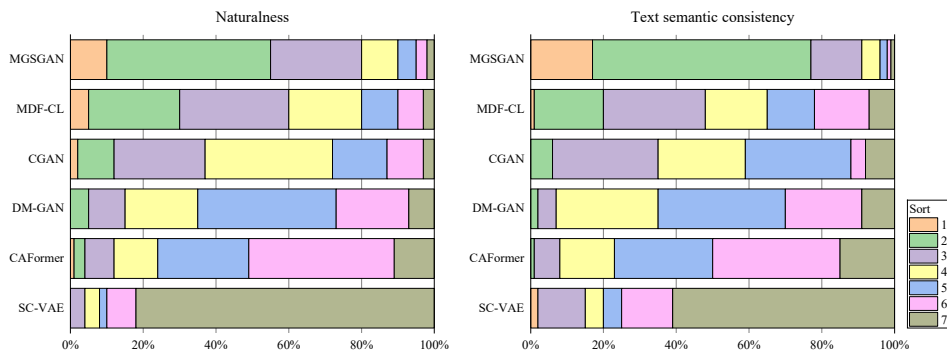


Table 1 Optimisation of green power consumption policy combinations

<i>Model</i>	<i>FID</i> ↓	<i>KID</i> ↓	<i>PSNR</i> ↑	<i>SSIM</i> ↑
SC-VAE	8.71	0.37	22.12	87.94
CAFormer	5.86	0.29	24.63	87.57
DM-GAN	5.58	0.24	25.39	89.96
CGAN	4.49	0.18	25.84	90.07
MDF-CL	3.98	0.14	26.02	90.84
MGSGAN	3.01	0.09	27.58	94.72

Finally, this experiment also introduces user studies on naturalness and textual semantic consistency for further quantitative comparison. One hundred artistic image sets are randomly selected, each set consisting of one text description, one baseline authentic artistic image, and six results restored by various methods. Each group of images is ranked in terms of naturalness and textual semantic consistency, with both the order between groups and within groups shuffled. The ranking distribution is shown in Figure 5. It can be observed that MGSGAN achieves the best performance among all methods in terms of naturalness and textual semantic consistency. In addition, it is worth noting that the average ranking of MGSGAN's textual semantic consistency is significantly higher than its naturalness. This indicates that MGSGAN can fully understand the content of text descriptions and repair missing areas according to the text description to generate realistic images consistent with the text description.

5.3 Analysis of ablation experiment results

In MGSGAN, TDM, ISDRM, and the CAB module are proposed separately. In this experiment, parts of the MGSGAN modules will be removed to verify their effects. Among them, CAB serves as a bridge between text features and artistic image features and is closely related to TDM. Therefore, it is bound with TDM, and when TDM is removed, CAB is also synchronously removed. The experiments in this subsection include three combinations: the MGSGAN method without the three modules (referred to as ‘BM’), adding ISDRM to BM, and adding TDM to the baseline model.

Table 2 Quantitative results of the melting experiment

<i>Index</i>	<i>BM</i>	<i>Join ISDRM</i>	<i>Join TDM</i>	<i>MGSGAN</i>
FID↓	9.15	6.08	4.37	3.01
KID↓	0.63	0.71	0.48	0.09
PSNR↑	18.46	20.19	25.38	27.58
SSIM↑	84.35	89.67	87.51	94.72

Table 2 shows the quantitative results of the ablation experiments. It can be seen that the addition of ISDRM improves PSNR and SSIM, but causes a slight decrease in metrics such as FID and KID. This is because ISDRM mainly works on the texture and details of artistic images, generating constraints that force the network to learn more refined textures from baseline authentic images; however, this also leads to a decline in the model’s ability to repair image semantics. Conversely, the addition of TDM significantly improves the FID and KID metrics while causing slight decreases in PSNR and SSIM. Due to the guidance provided by text descriptions through TDM, the semantic restoration capability of the model is enhanced; however, its detail processing capability becomes relatively weaker compared before. In order to ensure both semantic restoration capabilities and detail processing capabilities simultaneously, it is possible to consider adding SDRM on the basis of TDM, which can lead to significant improvements in indicators. Therefore, for image inpainting tasks, attention must be paid to both image detail processing capabilities and image semantic restoration capabilities.

6 Conclusions

In the field of digital art regeneration, the restoration and semantic reconstruction of artistic images are highly challenging tasks. Aiming at the lack of understanding of deep semantics in existing research, which leads to semantic errors in restored content of artistic images, this paper proposes a method for restoring artistic images and reconstructing their semantics under multimodal AIGC. First, to address the problem of mode collapse and training crashes commonly encountered with GANs, Gated Convolution is combined with spectral normalisation to optimise the GAN. The traditional convolution in the feature extraction part is replaced by gated convolution, which utilises its inherent learnable soft mask mechanism to dynamically select information from different regions within an image. The spectral normalisation mechanism is introduced into each convolutional layer of the discriminator network, ensuring that the parameters of each layer satisfy the Lipschitz constraint, thereby

optimising model training stability. Then, a method for restoring and reconstructing semantics in artistic images by integrating multimodal features, named MGSGAN, was designed. The text disentanglement module in MGSGAN can utilise information inherent to damaged artistic images to disentangle text descriptions, obtaining special textual features beneficial for restoration. The cross-modal attention module of MGSGAN makes the network focus more on important information within text descriptions, ensuring that the restoration results align as closely as possible with the text. The dual-channel reconstruction module can enhance the network's ability to predict image structure and semantic details, thereby improving the accuracy of restoration results. Experimental results show that the proposed method reduces FID and KID by at least 24.37% and 35.71%, respectively, demonstrating superior detail recovery capability and semantic coherence, enabling effective handling of artistic image restoration tasks under complex semantic scenarios.

Declarations

All authors declare that they have no conflicts of interest.

References

- Aggarwal, A., Mittal, M. and Battineni, G. (2021) 'Generative adversarial network: an overview of theory and applications', *International Journal of Information Management Data Insights*, Vol. 1, No. 1, pp.10–17.
- Amaranageswarao, G., Deivalakshmi, S. and Ko, S-B. (2022) 'Joint restoration convolutional neural network for low-quality image super resolution', *The Visual Computer*, Vol. 38, No. 1, pp.31–50.
- Balashov, M. (2015) 'Maximization of a function with Lipschitz continuous gradient', *Journal of Mathematical Sciences*, Vol. 209, No. 1, pp.12–18.
- Barbu, T. and Favini, A. (2014) 'Rigorous mathematical investigation of a nonlinear anisotropic diffusion-based image restoration model', *Electronic Journal of Differential Equations*, Vol. 129, pp.1–9.
- Bhele, S., Shriramwar, S. and Agarkar, P. (2023) 'An efficient texture-structure conserving patch matching algorithm for inpainting mural images', *Multimedia Tools and Applications*, Vol. 82, No. 30, pp.46741–46762.
- Chen, G., Wen, Z. and Hou, F. (2023) 'Application of computer image processing technology in old artistic design restoration', *Heliyon*, Vol. 9, No. 11, pp.15–27.
- Chen, L., Xiong, Q., Zhang, W., Liang, X., Gan, Z., Li, L. and He, X. (2025) 'Multi-modal degradation feature learning for unified image restoration based on contrastive learning', *Neurocomputing*, Vol. 616, pp.12–24.
- Chen, Z., Zhang, Y., Gu, J., Kong, L. and Yuan, X. (2022) 'Cross aggregation transformer for image restoration', *Advances in Neural Information Processing Systems*, Vol. 35, pp.25478–25490.
- Gupta, V., Sambyal, N., Sharma, A. and Kumar, P. (2021) 'Restoration of artwork using deep neural networks', *Evolving Systems*, Vol. 12, No. 2, pp.439–446.
- Kumar, P. and Gupta, V. (2023) 'Restoration of damaged artworks based on a generative adversarial network', *Multimedia Tools and Applications*, Vol. 82, No. 26, pp. 40967–40985.
- Kumar, P., Gupta, V. and Grover, M. (2024) 'Dual attention and channel transformer based generative adversarial network for restoration of the damaged artwork', *Engineering Applications of Artificial Intelligence*, Vol. 128, pp. 40967–40985.

- Li, J., Huo, H., Liu, K. and Li, C. (2020) 'Infrared and visible image fusion using dual discriminators generative adversarial networks with Wasserstein distance', *Information Sciences*, Vol. 529, pp.28–41.
- Liang, B., Jia, X.-x. and Lu, Y. (2021) 'Application of adaptive image restoration algorithm based on sparsity of block structure in environmental art design', *Complexity*, Vol. 2021, No. 1, pp.63–68.
- Liu, A. and Laili, Y. (2018) 'Balance gate controlled deep neural network', *Neurocomputing*, Vol. 320, pp.183–194.
- Liu, H., Shao, M., Wan, Y., Liu, Y. and Shang, K. (2025) 'SeBIR: semantic-guided burst image restoration', *Neural Networks*, Vol. 181, pp.68–74.
- Lyu, Q., Zhao, N., Yang, Y., Gong, Y. and Gao, J. (2024) 'A diffusion probabilistic model for traditional Chinese landscape painting super-resolution', *Heritage Science*, Vol. 12, No. 1, pp.97–105.
- Ma, H., Liu, D. and Wu, F. (2022) 'Rectified wasserstein generative adversarial networks for perceptual image restoration', *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 45, No. 3, pp.3648–3663.
- Marivani, I., Tsiligianni, E., Cornelis, B. and Deligiannis, N. (2022) 'Designing CNNs for multimodal image restoration and fusion via unfolding the method of multipliers', *IEEE Transactions on Circuits and Systems for Video Technology*, Vol. 32, No. 9, pp.5830–5845.
- Petrovici, M.-A., Damian, C. and Coltuc, D. (2018) 'Maximum entropy principle in image restoration', *Advances in Electrical and Computer Engineering*, Vol. 18, No. 2, pp.77–84.
- Ren, H., Sun, K., Zhao, F. and Zhu, X. (2024) 'Dunhuang murals image restoration method based on generative adversarial network', *Heritage Science*, Vol. 12, No. 1, pp.39–45.
- Wang, C., Pan, J., Wang, W., Dong, J., Wang, M., Ju, Y. and Chen, J. (2023) 'Promptrestorer: a prompting image restoration method with degradation perception', *Advances in Neural Information Processing Systems*, Vol. 36, pp.8898–8912.
- Wang, R. and Tao, D. (2016) 'Non-local auto-encoder with collaborative stabilization for image restoration', *IEEE Transactions on Image Processing*, Vol. 25, No. 5, pp.2117–2129.
- Wang, X., Ao, Z., Li, R., Fu, Y., Xue, Y. and Ge, Y. (2024) 'Super-resolution image reconstruction method between Sentinel-2 and Gaofen-2 based on cascaded generative adversarial networks', *Applied Sciences*, Vol. 14, No. 12, pp.50–63.
- Wang, X., Bylinskii, Z., Hertzmann, A. and Pepperell, R. (2020) 'Toward quantifying ambiguities in artistic images', *ACM Transactions on Applied Perception (TAP)*, Vol. 17, No. 4, pp.1–10.
- Yue, Z., Wang, J. and Loy, C.C. (2024) 'Efficient diffusion model for image restoration by residual shifting', *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 47, pp.116–130.
- Zha, Z., Yuan, X., Wen, B., Zhang, J., Zhou, J. and Zhu, C. (2020) 'Image restoration using joint patch-group-based sparse representation', *IEEE Transactions on Image Processing*, Vol. 29, pp.7735–7750.
- Zhai, L., Wang, Y., Cui, S. and Zhou, Y. (2023) 'A comprehensive review of deep learning-based real-world image restoration', *IEEE Access*, Vol. 11, pp.21049–21067.
- Zhan, F., Yu, Y., Wu, R., Zhang, J., Lu, S., Liu, L., Kortylewski, A., Theobalt, C. and Xing, E. (2023) 'Multimodal image synthesis and editing: the generative AI era', *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 45, No. 12, pp.15098–15119.
- Zhang, C., Barbano, R. and Jin, B. (2021) 'Conditional variational autoencoder for learned image reconstruction', *Computation*, Vol. 9, No. 11, pp.114–121.
- Zhang, Y., Sun, L., Yan, C., Ji, X. and Dai, Q. (2018) 'Adaptive residual networks for high-quality image restoration', *IEEE Transactions on Image Processing*, Vol. 27, No. 7, pp.3150–3163.
- Zhao, L., Wang, K., Zhang, J., Wang, A. and Bai, H. (2023) 'Learning deep texture-structure decomposition for low-light image restoration and enhancement', *Neurocomputing*, Vol. 524, pp.126–141.