applied sciences

MDPI

Article

Feature Separation and Fusion to Optimise the Migration Model of Mural Painting Style in Tombs

Meng Wu [1,2,*,†], Minghui Li [1,*,†] and Qunxi Zhang [3]

[1] School of Information and Control Engineering, Xi'an University of Architecture and Technology, Xi'an 710055, China
[2] Institute for Interdisciplinary and Innovate Research, Xi'an University of Architecture and Technology, Xi'an 710055, China
[3] Shaanxi History Museum, Xi'an 710061, China; qxzh98@sina.com
* Correspondence: wumeng@xauat.edu.cn (M.W.); liminghui@xauat.edu.cn (M.L.)
† These authors contributed equally to this work.

Abstract: Tomb murals are different from cave temple murals and temple murals, as they are underground cultural relics, their painting style is unique, solemn, and austere, and the performance image is characterised by simple colours, low contrast, and fewer survivors. During the digital restoration process, it is important to have sufficient reference samples to ensure the accuracy of the restoration. In addition, the style of mural paintings in the tombs varies greatly from other styles of murals and types of word paintings. Therefore, learning the unique artistic style of tomb murals, providing stylistically consistent training samples for digital restoration, and overcoming the problems of dim lighting and complex surface granularity of tomb murals are all necessary for research. This paper proposes a generative adversarial network algorithm that separates and fuses style features to enhance the generative network's ability to acquire image information. The algorithm extracts underlying and surface style feature details of the image to be tested and conducts fusion generation experiments. The generative network's parsing layer modifies the input noise tensor and optimises the corresponding weights to prevent misalignment between drawing lines and fresco cracks. Finally, to optimise the fresco generation effect, we add the corresponding loss function in the discriminator. The tomb murals dataset was established for experiments and tests, and quantitatively and qualitatively analysed with other style migration models, and SSIM, FID, LPIPS and NIQE were used as evaluation indexes. The results were 0.97, 269.579, 0.425 and 3.250, respectively, and the effect of style migration of this paper's method was significantly higher than that of the control group model.

Keywords: deep learning; tomb murals; style migration; generative adversarial networks; feature fusion

check for updates

Citation: Wu, M.; Li, M.; Zhang, Q. Feature Separation and Fusion to Optimise the Migration Model of Mural Painting Style in Tombs. *Appl. Sci.* **2024**, *14*, 2784. https://doi.org/10.3390/app14072784

Academic Editor: Andrea Prati

Received: 3 March 2024
Revised: 23 March 2024
Accepted: 24 March 2024
Published: 26 March 2024

Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Tomb murals are a form of wall painting known for their long history, diversity of subject matter and wide-ranging artistic significance [1] These murals are a valuable repository of shared artistic and cultural heritage across the country and around the world. However, disseminating mural art presents more challenges than other forms of painting, such as oil painting and drawing. This is due to the unique nature of murals and the difficulties involved in completing the painting process. As shown in Figure 1, the tomb mural location is poorly lit. The murals are on both sides of the tomb passage, which is susceptible to damage by cuts and scrapes. The murals are usually of great length and need to be captured and photographed in batches of sub-lenses. Due to various factors, including time, each mural may differ. However, modern technology has enabled the restoration of tomb murals for public enjoyment. The restoration and digital display of the murals have used digital inpainting and stylistic relocation to preserve their integrity.

Relocating the style of the mural art in the crypts is a challenging task due to flaking, disease, and other natural or man-made factors. Modern relocation techniques can easily erase or distort features of mural art. Therefore, it is important to approach this task with caution.

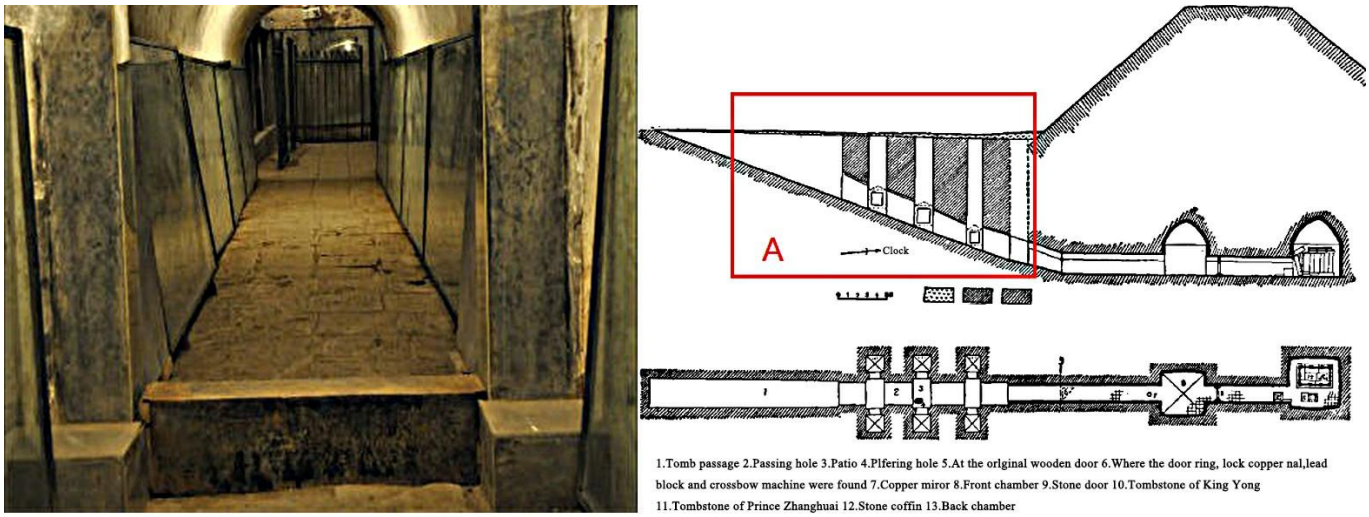


Figure 1. Diagram of the excavation site and layout of the mural in the tombs. (**a**) Tomb mural scene. (**b**) The location of the mural in the tomb.

For preserving tomb murals, digital inpainting is an effective approach. The restoration of fragmentary samples of tomb murals via the internet can reduce the cost of manual restoration and can prevent potential restoration errors. Recently, Wu et al. implemented mural restoration using a deep learning network with added dual attention [2]. Figure 2 shows the restored mural split-screen of the tomb. Nonetheless, mural restoration algorithms necessitate numerous training samples as a foundation. The application of style migration algorithms for training sample provision has become a prevalent method. Specifically, mural style migration is considered beneficial for augmenting the dataset used for digital inpainting of tomb murals. This method provides a multitude of samples with distinct styles and textures, thus amplifying the training dataset of the restoration model. The restoration model can achieve a more comprehensive and robust feature representation, leading to improved quality and accuracy in the restoration outcomes. In addition, the mural-style migration samples emulate the damage conditions on the tomb murals, including cracks, fading, and missing elements. Consequently, the restoration model can accurately recognise damage patterns and analyse the damage to the murals. This facilitates a more precise restoration process. Finally, the images resulting from mural style migration can serve as evaluation criteria for restoration outcomes. The quality and accuracy of the restoration are measured quantitatively by comparing the similarity and consistency between the restored outcomes and the generated samples. The use of this assessment technique can provide quantitative metrics to help evaluate the performance and potential for improvement of the restoration model.

In recent years, deep learning techniques have been continuously used by researchers for style migration experiments, such as Gatys, who successfully performed style migration [3] on ordinary photographs by separating and reorganising the extraction of image style and content by convolutional neural networks, respectively. However, the application of this technique to tomb murals resulted in chaotic image colour fills and imbalanced image gaps, which could not solve the problem of large area restoration during image restoration. The dominant colour is extracted from the input image and converted into palette-quantised chroma and luminance channels [4]; contour shapes are extracted by the Laplacian of Gaussians filtering (LoG, LoG); the tensor quantisation is optimised to

generate the image colour [5]; the contours are adaptively smoothed to generate the texture by flow field computation; and finally the colour images and drawn textures are mixed to produce the final result [6]. However, due to the large gap between the Western oil brush painting and the tomb mural, the image produced by the application of this method is prone to texture distortion, which is not conducive to the treatment of the fine cracks in the mural. Chen et al. migrated into the image style of cartoon painting [7] by adding cartoon semantic loss [1] and cartoon edge confrontation loss and using a simple patch level discriminator, but the cartoon painting is simple in the use of colour and exaggerated in its allegory, which is very different from the tomb mural in its profound expression and rich colour, and the repair effect is poor in the case of a large break in the small dataset of mural image repair. All of these methods mentioned above can generate images of significantly higher quality in the corresponding domains. However, when these techniques are applied to tomb paintings, the visual quality of the images is compromised by the presence of noise and staining due to disease in the tomb paintings, as well as the variable applicability of these methods, which leads to the confusion of the elements in the stylised images generated. Therefore, the artistic style of tomb murals is fundamentally different from that of ordinary paintings, so the commonly used style migration model is not suitable for tomb murals.



Figure 2. The current situation of missing information on murals.

Tomb murals adorn the walls of tombs, showcasing the artistic and religious aspects of the era. Striving to preserve this unique art form and express their own style, artists used brush and ink to capture the essence of the times. The murals depict religious beliefs and secular ideas that were clandestine in nature. By reflecting the zeitgeist, these murals highlight the significant shifts in history over the different epochs. The technique of creating murals is based on the traditional Chinese painting method of white drawing, in which thin lines are used to outline a highly generalised piece of work [8]. The colours employed within murals are unique and reflect the cultural development and material constraints of the time. Over time, however, the colours have gradually faded, resulting in significant changes in the saturation and overall appearance of the artwork. Most importantly, murals are rare and challenging to acquire due to their historical specificity. As a result, compiling the necessary dataset for deep learning model training is a crucial issue at present. Consequently, the style migration method was developed.

In order to achieve better results in generating tomb mural images, this paper introduces an innovative style migration model from real photographs to tomb mural art. We

combine the style generation network architecture [9] with the proposed identity protection measures to achieve the style migration of tomb fresco art. Initially, it was found that although many encoder-extracted image features are employed to calculate contrast loss [10] in the current market, the model's performance is limited due to the decline of image features from inter-domain style differences and model optimisation difficulties. Therefore, we believe that an iterative mapping approach that combines stylistic features and latent vectors to construct the latent space can effectively overcome these problems and improve the overall output quality.

We modify the regularisation on the original generator and added Perceptual Path Length (PPL) [11] and a type of feature separation to achieve our goal. In addition, as the painting technique of tomb murals differs from other styles, we simulate their line and colour style by line loss and colour loss. To ensure the generated stylised images have high visual quality, we propose using a semantic loss to control it. Finally, to mitigate the issue of model training overfitting, the accuracy of the discriminator is regulated through the inclusion of a micro-enhanceable enhancement. The primary achievements of this research can be summarised as follows:

1. We construct a style migration model based on the original StyleGAN2 model, which can efficiently learn the style mapping from real sketch photographs to tomb mural art images for image style migration. Our method outperforms existing methods to generate high-quality images of tomb mural art styles.
2. The generator proposes a separation and fusion of stylistic features to alleviate the negative impacts of style discrepancies across domains. Additionally, the discriminator integrates three types of losses, namely, line texture loss, colour loss, and semantic loss, which positively impact the quality of the images generated and the optimisation of the models.

In this research, we generate a dataset of tomb murals using real-world photographs and images of tomb mural art. The aim is to improve the effectiveness of the proposed method in training and testing, and to aid future research on the transfer of artistic styles in tomb murals.

2. Related Work

2.1. Image Style Migration

Due to the significant evolution of GPUs, research into deep learning and neural networks has become popular once again. Researchers propose Neural Style Transfer (NST) by combining texture migration techniques with deep learning models, which is currently a prominent area of research. Gatys et al. conducted an investigation into the first efficient and comprehensive neural approach. They identified two distinct elements—style and content—and advocated their integration for the creation of the new images using a pre-trained convolutional neural network (CNN) [12]. This enables style migration from one image to another by minimising loss of content and style. However, online optimisation is a time-consuming process. Currently, the most frequently employed deep learning methods that are based on online models are algorithmic networks featuring pre-trained structures (e.g., VGG) [1] and utilising the approach introduced by Gaty et al.

To remedy the inefficiency of Gatys' model, Johnson and Ulyanov presented two distinct proposals. Instead of optimising the parameters of the resultant images as Gatys et al. did, Johnson et al. suggested a Feed-forward (Feed-forward) approach that focuses on optimising the model instead [13]. Specifically, a forward propagation neural network is trained on the stylised image before the actual migration algorithm is performed; every time the stylised migration task needs to be completed, only the target image is used as input; and the trained neural network is subjected to one forward propagation computation, which will achieve the desired effect. In addition, Gatys et al. innovatively proposed the concept of semantic loss (perceptual loss) function, which improves the pixel-by-pixel

loss method mainly used in the past by integrating semantic information to calculate image loss.

Ulyanov et al. proposed an alternative network structure that utilises a multi-scale architecture as input for the generative network [14]. The generative network produces the target image texture information that is then fed into the previously trained interpretation network. This process results in stylised fusion to eventually obtain the target image. The research also found that normalising individual images greatly improves the quality of stylisation compared to batch images.

In the mural style migration experiments using the above models, the training of the Johnson and Ulyanov models was significantly ahead of the Gatiss model in terms of speed. Regularised training processing optimisation techniques allow the fast modelling of each input image while training the model, thus increasing the speed of model training convergence and reducing the time cost. However, both models also have some limitations because they require prior pre-training on the style of tomb murals using a forward network and an interpretation network. As a result, whenever it is necessary to migrate to a new style, pre-training is required for each model adaptation.

With the booming development of artificial intelligence and big data in recent years, various underlying models have seen considerable improvements and optimisations, and expanded to new areas. Zhang et al. improved AdaIN in style control by fixing the image optimisation target and setting cross-channel correlation, and comprehensively implemented Adaptive Style Modulation (AdaSM) [15] to enhance global style control, but large numbers of artefacts appeared in the field of mural applications. Fernandez et al. introduced style migration into the field of reinforcement learning, and utilised the Neural Policy Style Migration algorithm (NPST) [16] to achieve style migration to the field of intelligent robotics, but the model is not applicable to static mural effects. Yu et al. added the Covariance Attention Network (CovAttN) [17] to achieve stylistic uniformity between different patches of the image in terms of stylistic fusion, but stylistic confusion occurs when facing mural paintings where styles may be misaligned, which affects the final generation effect. Li et al. proposed a two-branch feature processing module to enhance the overall features of the image [18] and preserve the object contours and an interpolation module with spatial awareness to achieve semantic consistency of the generated content, but texture disappearance occurs when using murals with darker stylistic colours and fuzzy texture details.

2.2. Style Feature Style Migration

Feed-forward neural networks based on the research of Gatys et al. are being used extensively to handle style migration tasks. Various transform-based methods implementing encoder–decoder architecture have been proposed. The AdaIN [19] technique is one of the most traditional methods that uses an adaptive instance normalisation layer to derive the mean and variance of features, thus enabling the module to learn spatial features from both shallow and deep features. Building upon AdaIN, the Whitening and Colouring Transform (WCT) [20] optimises by replacing variance with covariance. OptimalWCT [21] further improves results by utilizing a more general closed-form solution. Similarly, Linear Style Transfer (LST) [22] proposes a linear transformation of cross-domain features to solve the generic style migration problem. Additionally, ArtFlow [23] implements reversible neural flow to solve the problem of content leakage during transformation. Additionally, another approach to solving the image style transformation problem is the patch-based approach, where pattern features are used to beautify content features [24]. Specifically, StyleSwap [25] substitutes each activation patch of the content image with a corresponding style patch.

With the widespread use of attention mechanisms [26], attention-based methods have been introduced to the style migration field, resulting in a variety of attention-based models [27]. Firstly, when computing attention towards style within the feature space, taking into account the necessity of extracting features from various layers, SANet [28]

merges both local and global style patterns to incorporate style features that match the content features. Based on SANet model features, IEContraAST [29] can amalgamate internal and external learning [30], and comparative learning [31]. It uses SANet as its backbone and learns to incorporate human-perceived styles through the discriminator's external learning [32] to ensure precise content and style delivery. Furthermore, MANet [33] and mcnet [27] utilise multi-adaptive and multi-channel correlation techniques, respectively, to enhance feature fusion performance. PAMA [34] introduces progressive attentional manifold alignment, which relocates style features dynamically using repetitive attentional operations. Furthermore, within the style migration domain, there have been proposed conversion methods that utilise self-attention mechanisms and positional encoding [35]. With regards to the style transfer task, StyTr2 [36] has modified the converter structure and suggested the implementation of content-aware positional encoding. To ensure optimal style transformation, StyleFormer [37] has modified the transformer structure by utilising a combination of style libraries and parameterisation. Additionally, Zhu et al. [38] have proposed a novel form of adversarial loss, employing generative adversarial networks (GANs) and introducing image translation techniques that do not require pairwise data. Methods that leverage multi-style models are capable of transforming images into various styles without necessitating model retraining. These modelling techniques are versatile, and require less time to train models for new styles.

2.3. Image Quality Analysis Metrics

In this paper, we use four objective evaluation metrics to quantitatively compare the effects of the stylistic migration of images, three quality evaluation metrics with reference images, and one quality evaluation metric without reference images. Structural similarity (SSIM), Learned Perceptual Image Patch Similarity (LPIPS), and the Fréchet inception distance (FID) were selected for reference image evaluation metrics, and the Naturalness Image Quality Evaluator (NIQE) was selected for no-reference image evaluation metrics.

The structural similarity index (SSIM) [39] is a metric used to compare the degree of similarity between two images. The SSIM measurement system consists of three main contrast modules: luminance, contrast, and structure. The closer its value is to 1, the more similar the two images are, i.e., the less the content structure is distorted during the migration process. The SSIM function is a combination of the Luminance–Contrast function, the Contrast–Contrast function and the Structure–Contrast function with the following formula:

$$f(l(x,y),c(x,y)s(x,y)) = \frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2)}{\left(\mu_x^2 + \mu_y^2 + c_1\right)\left(\sigma_x^2\sigma_y^2 + c_2\right)} \tag{1}$$

where μ_x and μ_y are the average grey levels of the image, σ_x and σ_y are the standard deviation of the grey levels of the image, σ_{xy} is the covariance, and c_1 and c_2 are constants.

The distance score (FID) [40] is a method for extracting high-level image information through the top layer of a pre-trained neural network, and is used to measure the similarity of features between the real image and the generated image. The proposers of FID use the pre-trained Inception V4 [41] model to extract 2048-dimensional vectors as the feature representation of the image before the fully connected layer, and evaluate the degree of similarity between the feature vectors of the real and generated images by calculating the mean and covariance distance between them. When the features of the generated image and the real image are more similar, i.e., the smaller the mean and covariance, the smaller the value of FID, which reflects the degree of similarity between the images. The specific mathematical expression formula is:

$$FID = \left\|\mu_r - \mu_g\right\|^2 + \mathrm{T}_r\left(\Sigma_x + \Sigma_g - 2\left(\Sigma_x\Sigma_y\right)^{1/2}\right) \tag{2}$$

where μ_r is the average of the real image characteristics, μ_g is the mean of the generated image characteristics, T_r is the trajectory of the real image characteristics, Σ_x is the covariance matrix of the real image characteristics, and Σ_y is the covariance matrix of the generated image features.

Learned Perceptual Image Patch Similarity (LPIPS) [42], also known as "perceptual loss", is a metric for evaluating the perceptual similarity between images, which can be obtained by LPIPS, also known as "perceptual loss", is a metric for evaluating the perceptual similarity between images by forcing the generator to learn to reconstruct a fake image into a real image by allowing the generator to reverse map in a fake image. LPIPS uses a pre-trained VGG neural network to extract features from an image and normalise the features. Then, their perceived similarity is determined by calculating the L2 distance between the feature representations of the generated image and the real image on each fully connected layer. A lower value of LPIPS indicates that the two images are more similar, and conversely, the greater the difference. The specific mathematical expression formula is:

$$d(x,y) = \frac{1}{L}\sum_{l=1}^{L}\|w_l \odot (F_l(x) - F_l(y))\|_2 \tag{3}$$

where w_l is the weighting point of the layer, and F_l is the feature that was extracted.

The Naturalness Image Quality Evaluator (NIQE) [43] is a reference-free image quality evaluation algorithm that aims at evaluating the naturalness of an image, i.e., whether the image looks like a product of a natural scene. NIQE generates a simple but highly regularised Natural Scene Statistics (NSS) model based on a set of "quality-aware" features and fits it to the MVG model. The NIQE metric of the test image is then expressed as the distance between the Multivariate Gaussian (MVG) model of the NSS features extracted from the test image and the Multivariate Gaussian model of the quality-aware features extracted from the natural image corpus. The lower value represents the better quality of the image. The NIQE expression can be expressed as:

$$D(\nu_1, \nu_2, \Sigma_1, \Sigma_2) = \sqrt{(\nu_1 - \nu_2)^T \left(\frac{\Sigma_1 + \Sigma_2}{2}\right)(\nu_1 - \nu_2)} \tag{4}$$

where ν_1, ν_2 , Σ_1 , and Σ_2 denote the mean vector and covariance matrix of the natural MVG model and the test image MVG model, respectively. This matrix can be obtained from standard maximum likelihood estimation.

3. Method

3.1. Network Architecture

The network structure of the algorithm used in this paper is shown in Figure 3. In this section, for the input image, firstly, the feature separation is realised based on the StyleGAN2 [9] generator mapping network, and then the style mixing and feature fusion using intermediate vectors are used to combine the identity features with the expression features so as to realise the adequate training between different features of the input image. For the discriminator, this paper adds three loss functions to help optimise the discriminator, respectively: texture loss, colour loss, and semantic loss. The experimental results prove that the model can solve the problems of the degradation of image features due to inter-domain style differences and limiting the performance of the model due to difficulties in model optimisation.

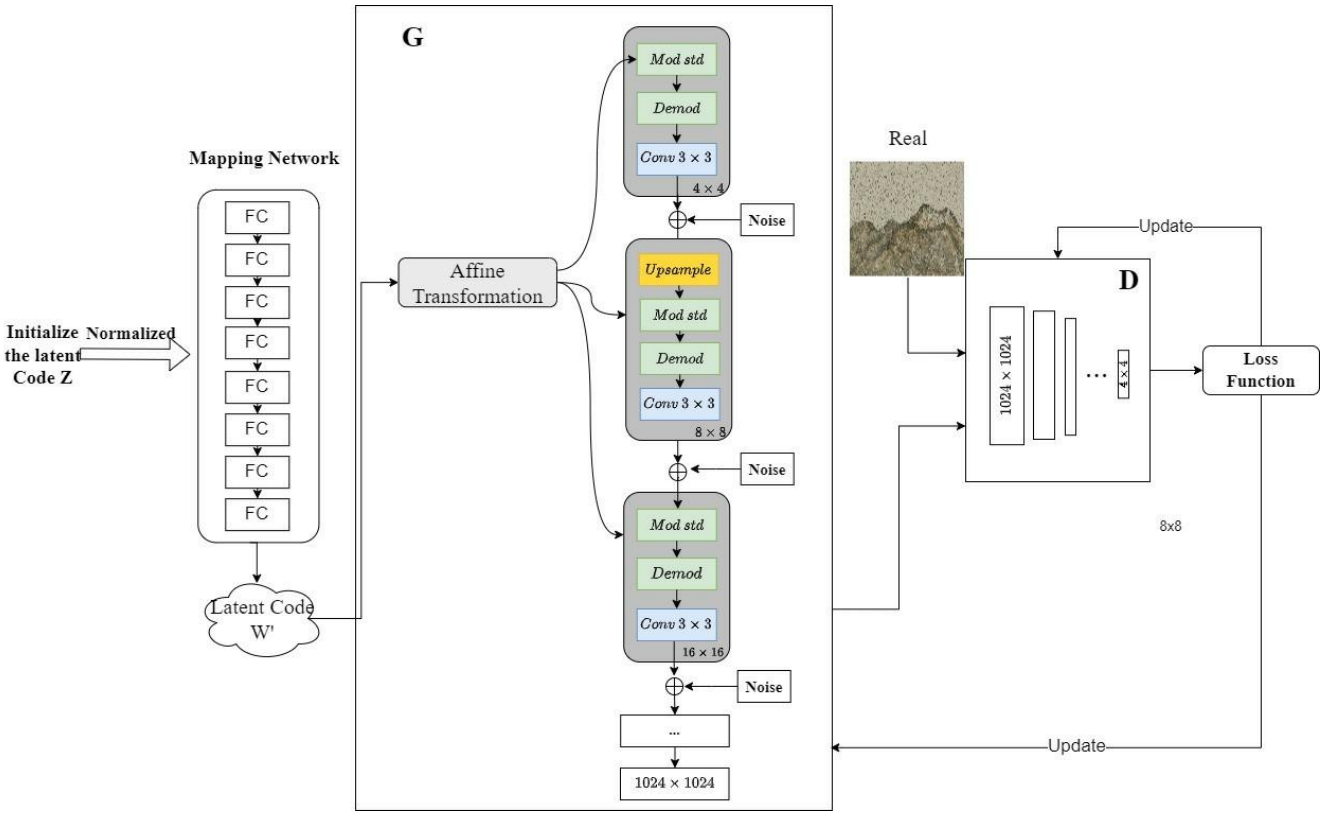


Figure 3. Algorithmic network structure of this paper.

3.2. Style Feature Separation and Fusion

In order to synthesise mural images from tomb murals in a targeted manner, this paper intends to distinguish between low- and high-level detail features depicted in the murals. The low-level features pertain to personal identification information and depict the figure's identity status within the tomb mural. There exists a notable correlation among diverse features found within the images depicted on tomb murals. This correlation further complicates the process of synthesizing the character's identity both qualitatively and quantitatively. However, the feature-decoupling technique serves as an effective measure to neutralise the impact of this correlation upon expression synthesis. This research aims to synthesise low-level images of tomb murals or to locate and quantitatively control fine high-level features in order to decouple the intertwined data variables in the original data space. Despite the presence of multiple features in the entire image, conventional generative adversarial networks are restricted to generating the entire image based on a 512-dimensional feature potential code. As it is infeasible to depict the entire 512-dimensional vector, the feature vectors are prone to entanglement. If each characteristic is assigned to one dimension of the vector, the other attributes can only be regulated conjointly by merging several dimensions. This largely affects the feature stability for the fine control of expression synthesis, which cannot be achieved with a single feature. Therefore, it is argued in this paper that the autonomous and precise control of the features of tomb murals can be achieved using feature decoupling. In this paper, we use the mapping network of the StyleGAN2 generator for feature separation. This means that a mapping network consisting of eight fully connected layers is used. The mapping network converts the input feature codes into intermediate vectors that have the same form as the input vectors. The intermediate vectors are then fed into the StyleGAN2 18-layer generative network, which generates 18 control vectors after passing through the generative network. The vectors produced by the trained generative model do not need to follow the distribution of the training data because the inputs to the generator are encoded intermediate vectors,

thus reducing the correlation between features. Figure 4 shows the structural framework of the implementation.

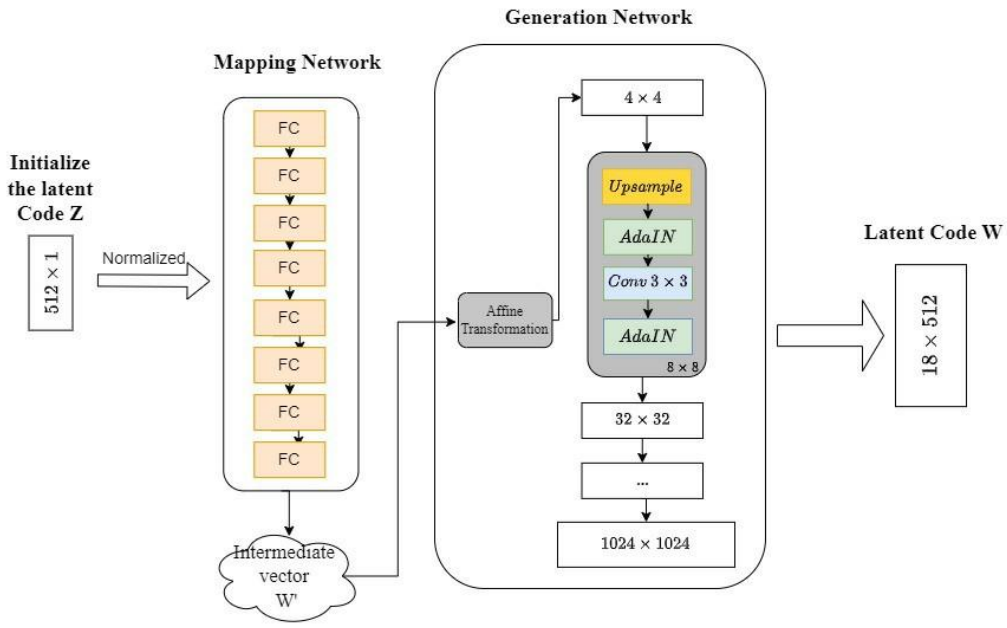


Figure 4. Structural framework for feature separation.

In this paper, the advantage of ResNet50 [44] is sufficient network depth to extract more feature information in different layers. Therefore, we implement feature code extraction of real images using ResNet50 to optimise the start vector in the feature separation framework [45].

We feed the features extracted from ResNet50 into the generator for image generation and adjusted the feature code for each level of image generation to match the image at that level. The chosen generator has nine levels, and each level has two levels of feature codes to control the direction of image synthesis. Thus, after a complete run of the generator, 18 × 512-dimensional feature latent codes can be obtained. In the whole feature separation framework, the Euclidean distance between the original image features and the generated image features is calculated as a loss function, and then the features of the original and generated images [46] are extracted using the VGG19 [1] network. Next, the final tuned feature code is used as input to the generator, and model optimisation is performed based on the Adam optimiser until the loss value reaches the preset value. At this point, the feature latent codes used in each stage of generation are considered ideal latent codes, and the generated image obtained by using the feature codes as inputs is considered the optimal image.

The original StyleGAN2 superimposes noise on all resolution layers of the synthetic mesh from low to high through noise broadcasting as a way of increasing the diversity of basic to advanced features of the image, e.g., increasing the diversity of features such as skin colour, hair colour, background colour to pose, face shape, wear, etc., of a face image. However, for the use of StyleGAN2 to generate images of tomb murals with specific features, due to the historical nature of the murals, the surface of the murals is often accompanied by a large area of noise and dirt, and sometimes, there are even scratches and missing parts,and these advanced features are often removed by the model discriminator as errors or artefacts after the addition of the noise-stacking module, which destroys the uniqueness of the original tomb mural, thus the need to reduce the diversity of advanced features in the image.

In response to the above problems of StyleGAN2 in the application of cryptographic images, this paper removes the noise superposition in the 128 × 128 to 1024 × 1024 resolution layers and retains the noise superposition in the 64 × 64 and lower-resolution layers in

order to increase the noise weights of the low-resolution layers and suppress the noise weights of the high-resolution layers. Figure 5 shows the optimised generator structure of this paper, and it can be clearly seen that there is no noise-stacking module in the 128 × 128 and higher-resolution layers.

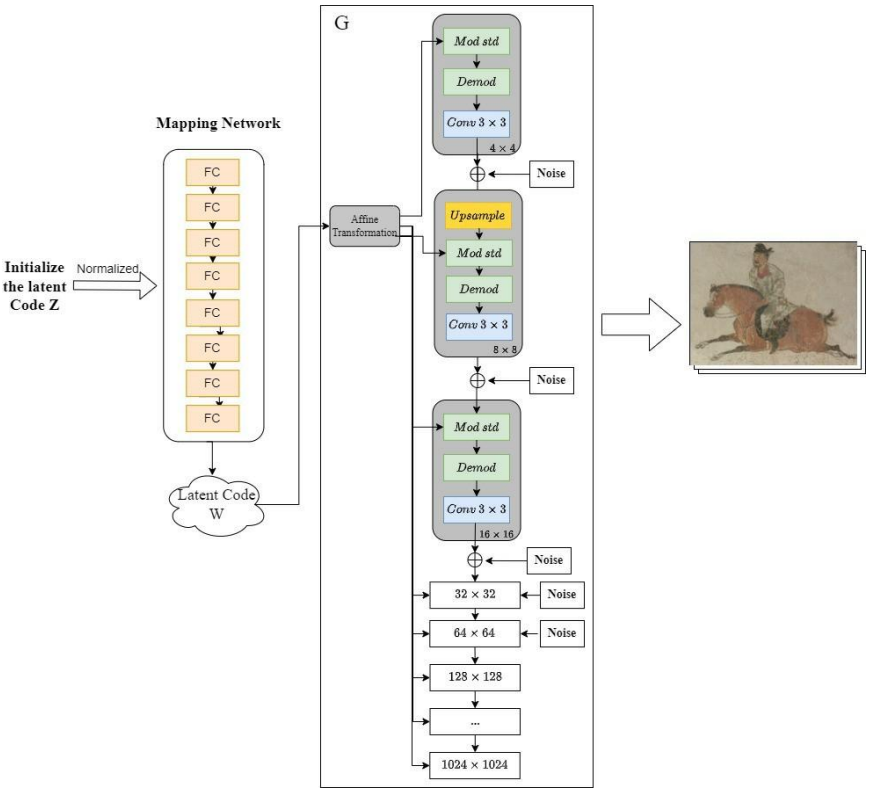


Figure 5. General framework for the generator after feature fusion and noise addition.

The StyleGAN2 generative network uses mapped latent space vectors at each level of resolution, which can lead to some correlation between features learned at different levels. The generative model uses two randomly selected input vectors and creates an intermediate vector W′ for each input vector to reduce the correlation. The entire generative network uses the first input vector to train a number of network layers, and then switches to a different input vector at a random point in the network. By randomly swapping feature vectors, the dependency of the network on the synthetic network is reduced, as well as the connections between different features. Although it does not always improve the performance of the model, this idea provides a very interesting side effect by allowing the network to constantly mix the attributes of different images. Two photographs, A and B, are used as inputs to this network model. The remaining features are then extracted from image B, while extracting the base-level features from image A. The network is then able to extract the remaining features from image B, while the low-level features are extracted from image A. The two sets of extracted features are then compared. The two sets of extracted features are then combined with the overall image used to create the overall image that is fused into the new cryptographic image features image style; the fine spatial resolution style is created by using the latent factors of the source image B. The resulting image, for example, all the colours (textures and light) and finer high-level features, can be derived from the high-level style of the source image B, including poses, textures, smudges, and scratches. And coarse stylistic spatial resolution (e.g., 4 × 4 to 8 × 8) can be generated from source image A by obtaining images generated using the latent factors.

3.3. Regularisation Optimisation

Path length regularisation (PPL) [47] is a regularisation technique that is often used to control the complexity of neural network models. In deep neural networks, the high complexity of the model often leads to overfitting problems. To solve this problem, path length regularisation effectively limits the complexity of the model and improves its generalisation performance by restricting the information propagation paths in the neural network. The method has been widely used in computer vision, natural language processing and other fields.

The basic idea of the method is to control the complexity of the model by limiting the path length from each neuron in the neural network to the output layer. Specifically, the method calculates the path lengths from each neuron to the output layer and weights and sums these path lengths with a penalty term. This penalty term can be understood as a constraint on the complexity of the model, which can effectively avoid the occurrence of the overfitting phenomenon.

The advantage is that no additional hyperparameters are introduced compared to other regularisation methods. This means that there is no need to manually adjust the hyperparameters when using the method, and the default parameters can be used directly. In addition, the method has a relatively small computational complexity, which makes it suitable for large deep neural networks. The formula (5) is as follows:

$$L_{pl} = \sum \left[\frac{1}{\epsilon^2} d(G_t(slerp(z_1, z_2)), G_{t+\epsilon}(slerp(z_1, z_2))) \right] \tag{5}$$

where ϵ is the subdivision sub-segment, which can be viewed as the step size, and is generally replaced by 1×10^{-4}, z_i is the intermediate hidden code obtained by the hidden space random code after the mapping network, d denotes the perceptual distance, which is generally computed in the discriminator, t denotes a certain point in time, $t \in (0, 1)$, $t + \epsilon$ denotes the next point in time, G is the generator, and $slerp$ denotes the linear interpolation, that is, interpolation on the latent space is interpolated according to the parameter t.

In the main loss function, the regularisation term and the adversarial loss of the GAN network are in one expression and can be optimised simultaneously. However, during training, it can be observed that the regularisation term is computed less frequently than the main loss function, so in this paper, we use inert regularisation by evaluating the regularisation term in a separate regularisation process that is run once every k training iterations.

The internal state of the Adam optimiser is shared between the loss term and the regularisation term, which means that the optimiser processes the gradient of the main loss first, and after k iterations processes the gradient of the regularisation term again. In other words, 16 iterations of the main loss function are performed before an iteration of the regularisation term is performed.

3.4. Texture Loss

In order for the discriminator to successfully identify the line drawing features of the tomb paintings simulated by the generator during discrimination, and to avoid degradation of the model performance due to differences in the number of sample categories, we use texture loss to improve the line texture consistency between the generated image and the input image. This is performed by extracting the edge mapping between the original image and the produced image using the pretrained edge extraction network HED [48], and using this edge mapping to compute the balanced cross entropy to obtain the texture loss. We calculate the texture loss as:

$$\begin{aligned} L_l(G, X) = E_{x \sim X} [& -\frac{1}{N} \sum_{i=1}^{N} \mu E(x)_i \log E(G(x))_i + \\ & (1 - \mu)(1 - E(x)_i) \log (1 - E(G(X))_i)] \end{aligned} \tag{6}$$

where N is the edge pixels extracted from the original or generated image, E is the edge extraction network HED, and μ is the weights.

3.5. Colour Loss

Due to the wash of history, the colours of the tomb mural are no longer bright and frivolous, so instead of pursuing the authenticity of the colours, the emphasis is on the decorative nature of the colours of the mural as a whole. In order to simulate the colours of the mural, a HSV hue channel-based colour loss is proposed. The hue channel's [49] histogram vector is initially extracted to represent the image's colour characteristics. Then, the cosine similarity between the generated image's hue histogram vector and the tomb's murals is calculated to obtain the colour loss. We express the colour loss as:

$$L_c(G, X, Y) = E_{x \sim X, y \sim Y}\left[\frac{H_u(G(x)) \cdot H_u(y)}{\|H_u(G(x))\| \times \|H_u(y)\|}\right] \tag{7}$$

where H_u is an 8-partitioned histogram vector of the tonal channels extracted from the image.

3.6. Semantic Loss

The subject matter of the tomb murals is mostly related to contemporary religion and politics, which results in the absence of certain content elements. In the case of landscape and architectural murals, in order to emphasise the landscape and architectural and even human elements of the mural, painters of the time achieved this by reducing the depiction of other larger, but less important, features, such as the sky, water and ground textures. As a result, the direct stylisation of real photographs that contain a lot of sky and ground can cause distortion and artefacts in these areas. To solve this problem, we have developed a semantic loss to avoid over-processing the relevant areas of the stylised image.

First, we use a pre-learned classification model [50] with exceptional scene recognition to obtain regions such as sky, ground or water in the image. Next, we calculate the structural similarity (SSIM) [51] between the actual photo and the produced image to obtain the semantic loss value. The equation for semantic loss is as follows:

$$L_{se}(G, X) = E_{x \sim X}[1 - SSIM(M_x \times x, M_x \times G(x))] \tag{8}$$

where M_x is the mask of the sky and the water region for x that was generated by the pre-learned classification model.

3.7. Total Losses

Adversarial learning and optimisation are utilised during the training phase to produce adversarial networks. The generative network G and the discriminative network D engage in a continuous game to update and optimise the network parameters. The generative network must have a low error rate to create "virtual samples" as similar as possible to real samples and enable the discriminative network to distinguish them as authentic data classes. The iterative optimisation of a generative adversarial network can be characterised as a problem that is simultaneously very large and very small. The ultimate aim of optimisation is to achieve a Nash equilibrium between the two parties involved. The discriminative network is required to constantly enhance its ability to discriminate between real and generated samples with utmost accuracy. Moreover, the optimisation objective involves achieving high accuracy for the discriminative network as expressed in Equation (9):

$$L_{GAN} = E_{x \sim P_{data}(x)}[\lg D(x)] + E_{z \sim P_z(z)}[\lg (1 - D(G(Z)))] \tag{9}$$

where x is the true sample and z denotes the input n-dimensional random noise. $P_{data}(x)$ is the true sampled distribution of the data, $D(x)$ is the probability that the input sample x will be identified as the true sample, $P_z(z)$ is the prior distribution of the probability,

$D(G(Z))$ is the likelihood of classifying the input sample as a counterfeit (i.e., not real) sample is being measured. The Nash equilibrium is reached by the network when the discriminator cannot differentiate between the various data source categories.

The total loss in this paper is expressed in terms of:

$$L(G, D, X, Y) = \lambda L_{GAN}(G, D, X, Y) + \lambda_1 L_l(G, X) + \lambda_2 L_c(G, X, Y) + \lambda_3 L_{se}(G, X) \quad (10)$$

where the GAN models loss, where the discriminator is derived from a normalisation function that aims to normalise and trim the StyleGAN2 output to fit the log-likelihood distance between features after s-type activations. In order to minimise the loss, data processing requires a large amount of GPU memory.

4. Experiments

4.1. Experimental Setup and Datasets

In this paper, we use the PyTorch framework to construct network models, and the experimental equipment is a desktop with Win10 system, Intel I9 9900 K 5.0 GHz CPU, 64 G RAM, and NVIDIA RTX 2080Ti for GPU acceleration, Intel, Santa Clara, CA, USA. During the training process, the image size is set to 512 × 512 as the input, the Adam optimiser is used for the training model optimisation iteration, and the warm-up adjustment strategy [52] is used to set the learning rate to 0.0005 when the model reaches optimal convergence. We set the batch size to 8 and train our network with 200,000 iterations.

Due to the serious lack of tomb mural data samples, in order to meet the model training needs, we create a non-public dataset. The metadata sample is the polo picture mural in the tomb of Prince Zhang Huai of the Tang Dynasty that was collected and standardised by the Shaanxi History Museum. We divide the polo picture of the tomb of Prince Zhang Huai of the Tang Dynasty into mural blocks with a pixel size of 256 × 256. The data are preprocessed, and 744 mural samples with rich texture details are screened out, and digital expansion methods [53] such as translation, flipping, random cropping, etc., are used to expand the number to 7000 as the basic training dataset. At the same time, in order to increase the versatility of the model, we select 5000 photos from the public datasets Vangogh2 Photo and Monet2Photo as supplements and add them to the basic dataset. The dataset is divided into a training set, a test set, and a verification set according to 8:1:1 to train the network model.

4.2. Comparison and Analysis with Learning Methods

This section is divided into two parts; on the one hand, the qualitative results of the mural dataset are analysed to verify the validity of the style migration model in this paper. On the other hand, quantitative metrics are used to demonstrate the change in performance specific metrics for style migration improvement using this method.

Subjective and Quantitative Analysis of Each Model

To evaluate the effectiveness of the improved model in this paper, the same dataset is used for training and testing in both qualitative and quantitative experiments. Figure 6 shows the comparison of this paper's method with other style migration methods, including Gatys [3], AdaIN [1], SANet [24], Artflow [23], CycleGAN [54], CUT [10], ChipGAN [55], and LseSim [56]. Gatys and Artflow are instance-based style transformation methods; they are typically applied in the field of nature photographs. CycleGAN and CUT are image transformation methods characterised by the need for a relatively symmetric image training domain. LseSim proposes a new spatially adaptive loss function that learns the image feature domain invariant spatial correlation map training. SANet introduces a spatial attention mechanism to achieve style integration. ChipGAN is for the migration of ink painting styles by introducing a loss function about ink strokes. And the structural similarity, perceptual similarity, distance score, and natural quality assessment of the results of each experiment are counted. The statistical data are shown in Table 1, and the statistical

results are represented using line graphs as shown in Figures 7–10. Our method is clearly superior to other models, and these metrics prove the excellence of our model.

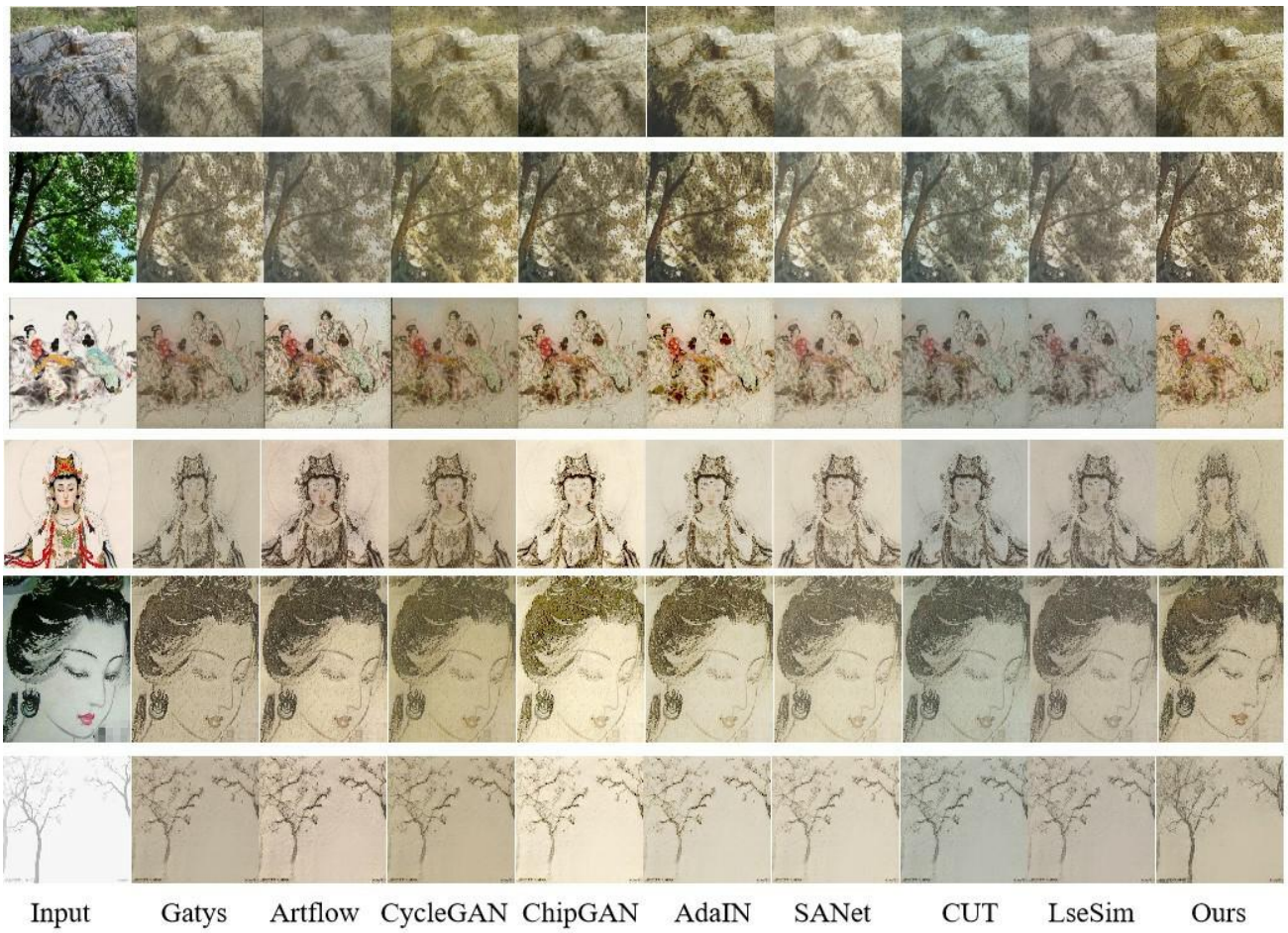


Figure 6. Results of the qualitative comparison with state-of-the-art methods.

As can be seen from the comparison results in Figure 6, Gatys uses iterative image optimisation to obtain clearer conversion results, but the method takes a long time for model training and loses too much time when used for bulk data expansion. AdaIN inputs images with different contents and styles; it can achieve the overall transfer of styles, but the network training will ignore the feature that there is sometimes a certain local correlation between the contents and styles, which may produce some halos at the edges of the image, affecting the overall visual effect of the generated image. The SANet network employs an attention mechanism for the effective suppression of the repetition of some image textures to ensure the uniqueness of the image texture information, but the stylistic features of the stylised images can adversely affect the generation process, e.g., the noise in the background of the generated image is more pronounced, thus affecting the natural presentation of the generated image. The Artflow network follows the distribution of stylistic features so that the low-level feature intervals of the generated image are fully presented, but the network is trained with high-level features in the high-dimensional space of the image through feature-clustering methods, which distorts the structure of the content image and makes the generated image produce a significant deformation, which is not in line with the generative purpose of the migration method. In contrast, our approach achieves better results. By removing the noise insertion module for in the high-dimensional space, it strengthens the connection similarity between the low-level features and reduces the dot artefacts and content distortion that often occur in the original network generation.

Moreover, the quality of the generated images is significantly improved due to the addition of three loss functions in the discriminator to optimise the model performance.

Table 1. Quantitative comparison of different methods, FID, SSIM, LPIPS and NIQE assays. Best results are shown in bold.

Method	SSIM↑	FID↓	LPIPS↓	NIQE↓
Gatys	0.93	355.525	0.555	5.771
Artflow	0.83	323.090	0.508	5.210
CycleGAN	0.86	295.235	0.453	4.623
ChipGAN	0.92	343.747	0.592	4.058
AdaIN	0.88	330.744	0.498	4.553
SANet	0.96	330.206	0.501	5.905
CUT	0.94	**258.743**	0.467	3.748
LseSim	0.91	370.869	0.491	3.712
Ours	**0.97**	269.579	**0.425**	**3.250**

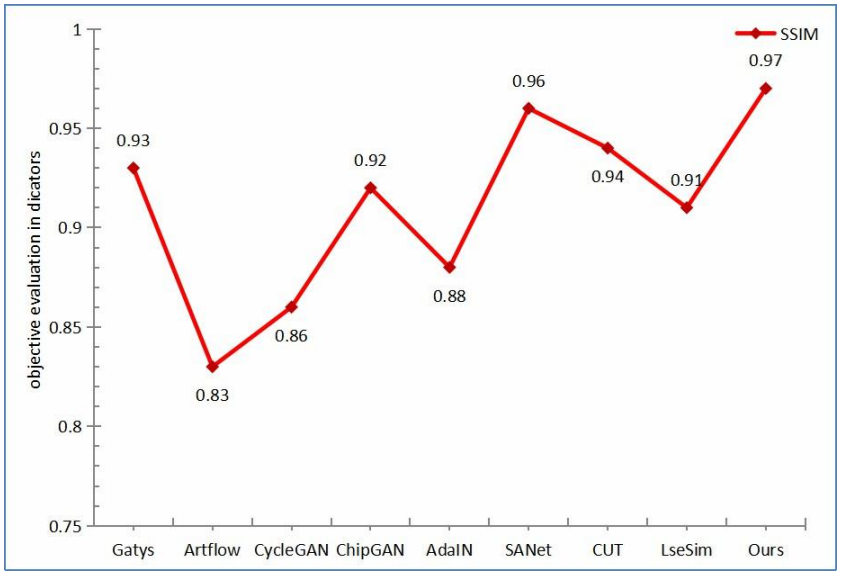


Figure 7. Indicators of structural similarity of experimental results.

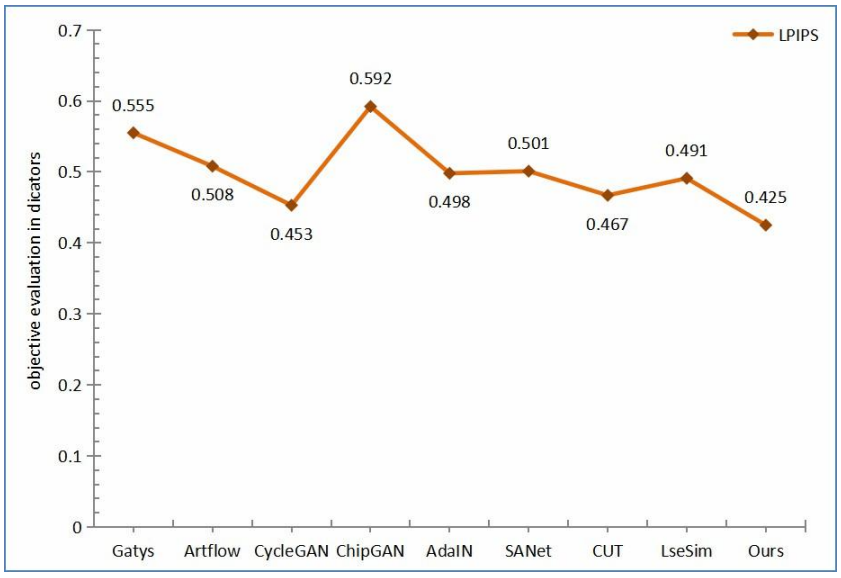


Figure 8. Perceived similarity indicators of experimental results.

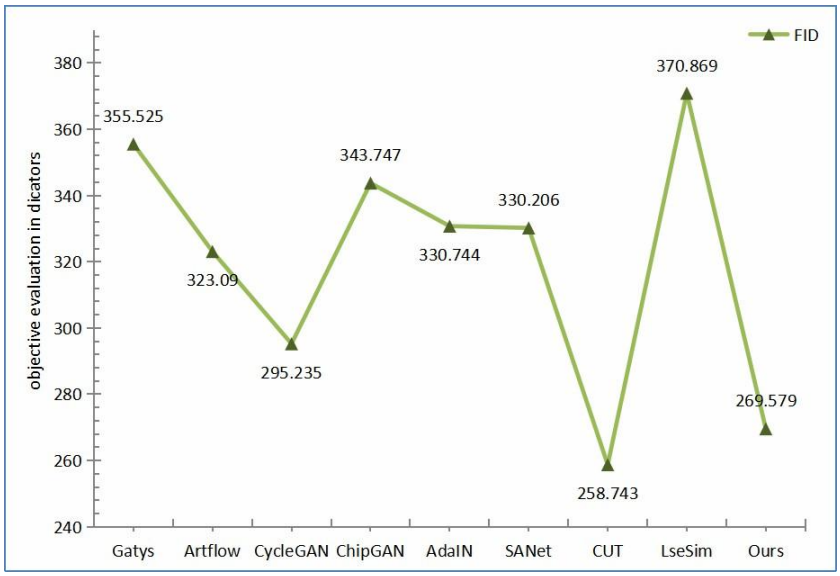


Figure 9. Distance score metrics for experimental results.

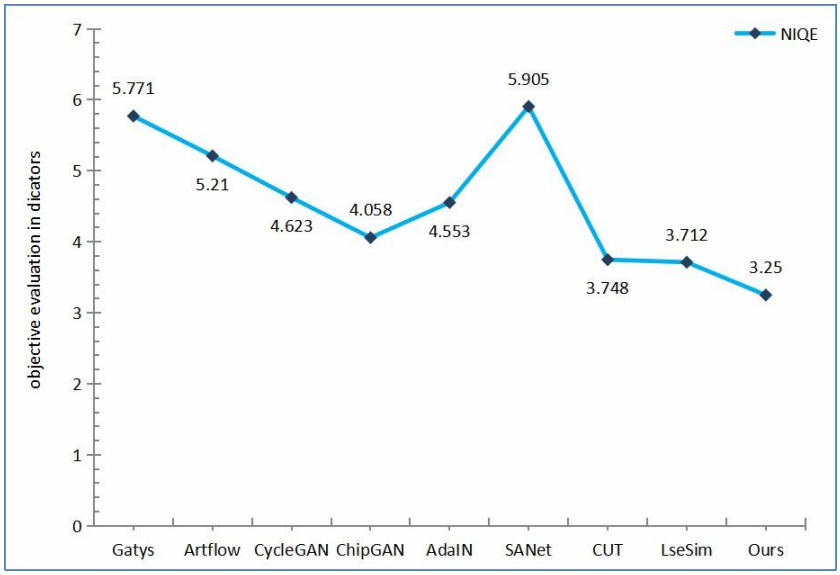


Figure 10. Indicators of natural quality assessment of experimental results.

As can be seen from Figure 9, our model is not as good as the CUT model in terms of FID metrics due to the concept of comparative learning introduced in the CUT model, which maximises access to the common structure between images by comparing the mutual information (MUTI) [57] between the input image and the output image. The model is more effective in calculating the FID score when the image structure is more homogeneous and clear. However, the training time of the model is long, and the requirement of the image input is high.

4.3. Experiment on Ablation

4.3.1. Style Feature Separation Weight Coefficients

In this study, the feature layers after feature decoupling are first analysed from the available research data, and the features with different depths are selected for fusion computation to guide the subsequent image synthesis process. In image synthesis, feature layers at different resolution levels contain different visual features of the tomb paintings. The quality of image synthesis can be affected by the number of features selected for fusion, and we choose to use feature codes from the first eight feature layers for fusion. This

is because these features have a more significant role in influencing the quality of the generated images, especially when dealing with important features of the murals, such as cracks, noise and stains. In order to increase the diversity of the generated data, for each computed feature layer combination, the remaining multiple features show a lower level of influence. Thus, the weight coefficients of the feature inputs play a decisive role in the impact of the model.

Figure 11 shows the results of the synthesised images when influenced by various feature-weighting coefficients. Table 2 demonstrates the numerical statistics of the objective metrics SSIM, FID, PSNR and NIQE for the experimental results. The various degrees of expression under different weighting factors are depicted in Figure 11a–c. The degree of distribution of stain texture in the mural is the main visual cue for this expression. As can be seen from the rate of change in size plots, the detailed texture becomes more balanced and careful with a weighting factor starting from -1.5, but the level of detail in the image deliberately omits the information about the cracks that is necessary for the fresco. When the coefficient is 1.0, the colour and texture information of the image is best presented. However, the coefficients should not be too high or they will distort the information received. The data in Table 2 show the same result, with all metrics performing optimally when the weights are 1.0. Even the NIQE metrics show a sharp increase in values when the feature weights are too large and the generated images deviate from the style that the model needs to achieve.

It is clear from the flowchart that the feature distribution density of the source image affects the size of such representations. However, in order to maintain the stability of the identity and texture information in the image, the value of the weight coefficients should be kept in the range of $-1.0 \sim 1.0$.

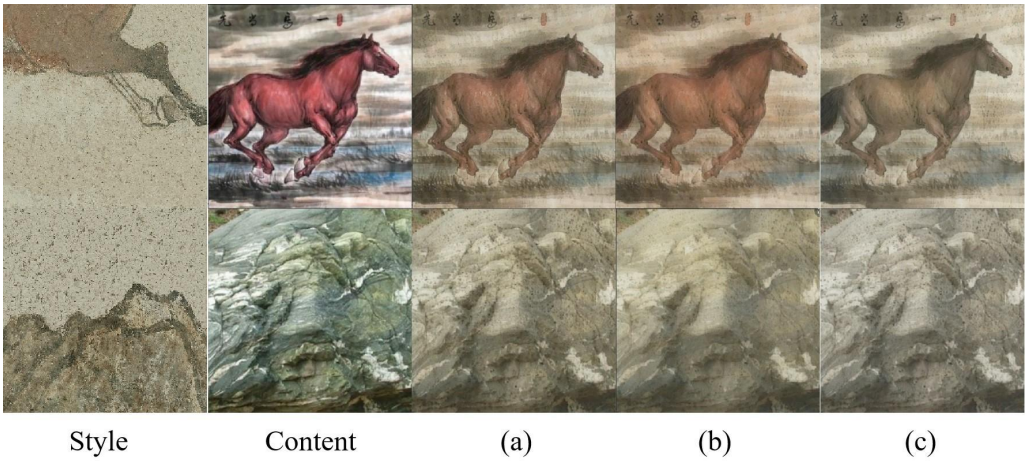


Figure 11. Synthesised images with different weighting factors, where (**a**) Coef = -1.5, (**b**) Coef = 1.0, (**c**) Coef = 2.5.

Table 2. Quantitative results of our method with different weighting parameters added. The best values are shown in bold.

Method	Coef = -1.5	Coef = 1.0	Coef = 2.5
SSIM↑	0.76	**0.77**	0.72
FID↓	326.013	**297.150**	336.078
PSNR↑	23.41	**23.95**	24.02
NIQE↓	3.543	**3.155**	4.191

4.3.2. Addition of a Random Noise Layer

By comparing the original StyleGAN2 model with the results generated by the research in this paper with the human eye, the image generated by the original StyleGAN2 is blurred, the surface colour blocks are different, and the texture details are missing, whereas

the image generated by the improved model proposed in this paper is much clearer, and the texture details of the edges and surface of the mural image are more obvious, and the transition of the surface colours is smooth. As StyleGAN2 inputs noise at each resolution layer of the synthetic network, the basic features such as the surface colour and texture of the mural change significantly, resulting in blurred or even missing features, while this paper controls the changes in the advanced features by suppressing the noise inputs at the high resolution layer so as to make the basic features clearer and more in line with the real situation. Therefore, the algorithm proposed in this paper can provide more realistic and objective samples for the generation or migration of tomb murals. Figure 12 shows the synthetic image with different resolution noise added. When the noise is added only below 16 × 16, the detailed information of the image cannot be presented in detail. And when adding noise at 128 × 128 layers and above, the generated image will have a lot of redundant information, which destroys the beauty of the overall structure of the information and fails to satisfy the migration purpose. As can be seen from the specific metrics in Table 3, the PSNR, SSIM, and NIQE metrics perform well with less fluctuation as the Gaussian noise increases, and the FID metrics are minimised and optimised to the highest degree at 64 × 64 resolution, but continuing to increase the noise tensor at higher levels produces a greater degree of degradation. The intuitive analysis and quantitative results demonstrate the effectiveness of our proposed method in terms of the effect of different resolutions of the noise tensor.

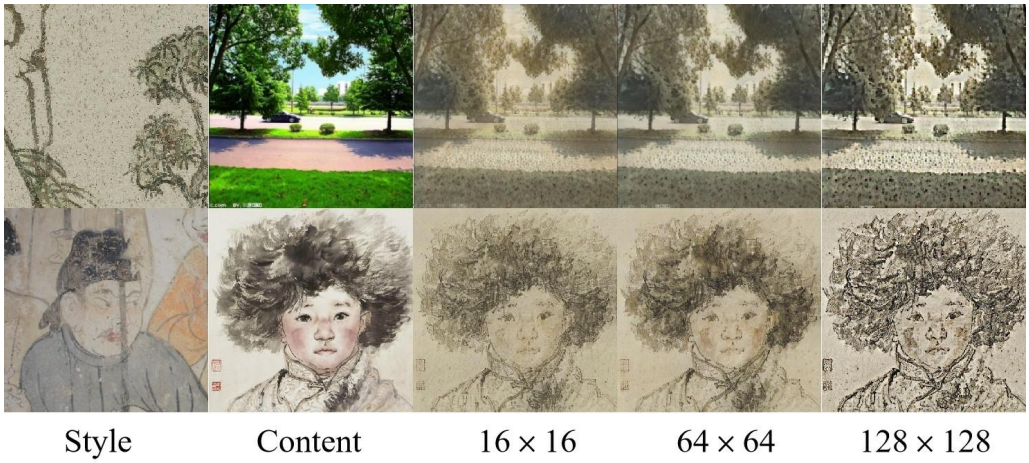


Figure 12. Synthesised image with noise addition with different resolutions.

Table 3. Quantisation results of our method with different noise tensor additions. The best mean results are in bold.

Method	16 × 16	64 × 64	128 × 128
SSIM↑	0.846	**0.864**	0.797
FID↓	328.324	**280.449**	380.833
PSNR↑	20.40	**20.42**	19.31
NIQE↓	3.285	**3.284**	3.355

4.3.3. Loss Function Ablation Experiments

The following figure shows the image generated after the proposed loss function compared with the generated image without the addition of this function module: by adding the colour loss and line loss, the generated image retains most of the salient regional information consistent with the content image, and the migration effect is more natural and the colours are more appealing, which is more in line with the characteristics of the murals of the tomb. The addition of semantic loss reduces the distortion that may be caused by missing elements in the generation of the mural to ensure the quality of the visual aspects of the stylised images of the tomb murals. Figure 13 shows the generation effect of each

function module. Where w/o L_l refers to the model generation effect of removing the texture loss function, w/o L_{se} refers to the model generation effect of removing the colour loss function, and w/o L_c refers to the model generation effect of removing the semantic loss function. When the L_l function is missing, a large amount of texture information of the generated result is removed by the model as error information, and the image display is incomplete. When the L_{se} function is missing, the migration effect is not natural enough, and the colours are more rigid and less natural. When the L_c function is missing, the mural generation result may be distorted due to the missing necessary elements, which affects the migration generation effect. Table 4 shows the quantitative analysis results of the ablation experiments for each loss function, from which it can be seen that our proposed method shows outstanding performance in all the metrics comparisons, and generates images with better quality.

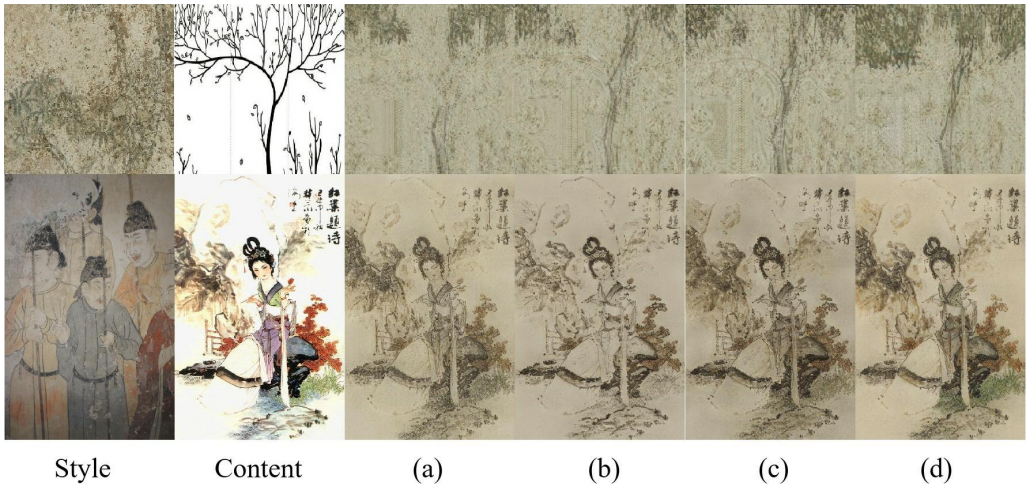


Figure 13. Generated ablation effects for each function module. Where (**a**) ours w/o L_l, (**b**) ours w/o L_{se}, (**c**) ours w/o L_c, (**d**) ours.

Table 4. Quantitative results of loss function ablation experiments. Best results are shown in bold.

Method	SSIM↑	FID↓	LPIPS↓	PSNR↑
Ours w/o L_l	0.802	335.331	34.440	19.43
Ours w/o L_{se}	0.800	342.972	40.284	19.59
Ours w/o L_c	0.861	376.027	34.989	20.48
Ours	**0.871**	**310.503**	**32.667**	**20.91**

4.4. Application Analysis in Digital Restoration

In order to verify the excellent results of the generative model generation results in digital restoration, we add the generated images of this paper's model to the original training dataset for simple data augmentation and input into the digital restoration model for training and testing. The test results with other commonly used style migration model generated by the tomb style image results are input into the digital restoration model for comparison; the comparison results are shown in Figure 14. Meanwhile, in order to better visualise the excellence of this paper's model in the field of digital restoration, we use SSIM, FID, IPIPS, and NIQE metrics to analyse the results in terms of quality, and the results, as shown in Table 5, show that our method significantly outperforms the other baselines in terms of the various quality metrics, with the FID metrics being slightly worse than with the AdaIN model. These specific indicators strongly demonstrate the effectiveness of our proposed method.

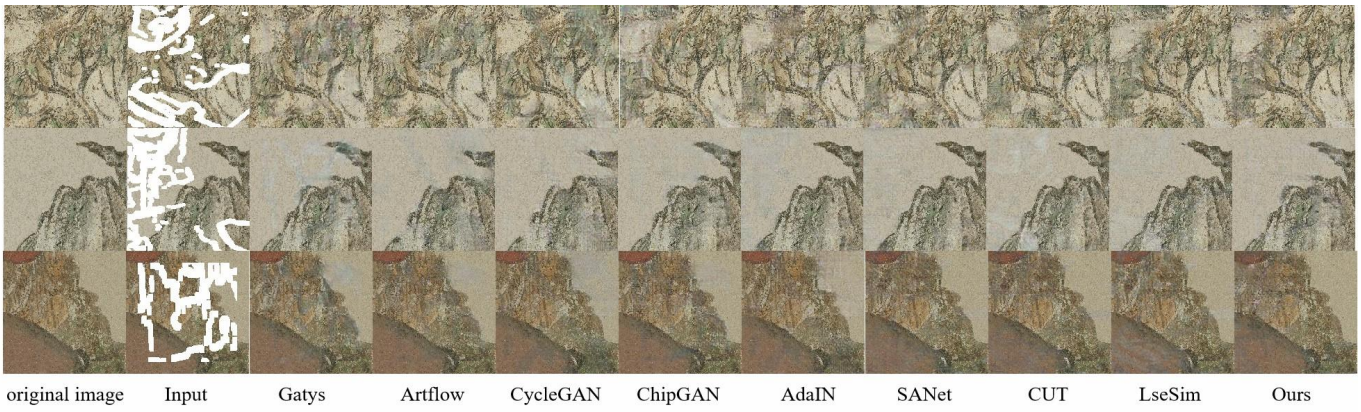


Figure 14. Digital restoration effects for various styles of migration models.

Table 5. Quantitative effects of different models for digital restoration. Best results are shown in bold.

Method	SSIM↑	FID↓	LPIPS↓	NIQE↓
Gatys	0.819	481.126	0.272	4.205
Artflow	0.863	527.819	0.257	4.096
CycleGAN	0.822	511.409	0.198	4.173
ChipGAN	0.855	501.745	0.222	4.414
AdaIN	0.859	**425.532**	0.366	4.305
SANet	0.867	437.335	0.208	4.275
CUT	0.864	459.213	0.250	4.148
LseSim	0.868	476.218	0.225	4.538
Ours	**0.871**	458.260	**0.166**	**4.032**

5. Conclusions

This paper proposes a generative model for a tomb mural style based on feature separation and fusion. The model aims to address the high cost of digital restoration and the lack of training samples, while also improving the restoration of tomb mural style samples. First, it is proposed to separate image features and detail features before fusion to improve the performance of the model, and to adjust the noise insertion module of the parsing layer in the generator framework to preserve the noise, cracks, etc., that are unique to tomb murals, and to optimise the generation quality of the generator in a targeted way. Then, texture loss, colour loss and semantic loss functions are used in the original discriminator to constrain the model to learn the style of tomb murals. Extensive experiments on synthetic and real datasets proved the effectiveness of our method and achieved the most excellent results; objectively, our model, in terms of generative effectiveness, improved with SSIM by 0.05, FID optimisation was second only to the CUT model, and the LPIPS and NIQE were also optimised by reductions of 0.125 and 1.364. When using our model for digital restoration, SSIM was improved by 0.052, the FID optimisation index was second only to the AdaIN model, and LPIPS and NIQE also obtained 0.2 and 0.506 reduction optimisation, which fully proves the feasibility of our method.

In the future, we will focus on more detailed and precise stylistic migration of chamber murals, such as generating stylised images in a given cave or dynasty.

Author Contributions: Conceptualisation, M.W.; methodology, M.W.; writing—review and editing, M.W.; software, M.L.; writing—original draft preparation, M.L.; validation, Q.Z.; resources, Q.Z.; data curation, Q.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China (No. 61701388), the Cross-disciplinary Fund of Xi'an University of Architecture and Technology (No. X2022082), (No. X20230085) and the Fund of the Ministry of Housing and Urban-Rural Development (No. Z20230826).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Restrictions apply to the availability of these data. Data were obtained from Shaanxi History Museum and are available from the authors with the permission of Shaanxi History Museum.

Acknowledgments: We are thankful for the help of Huaidong Zhao from the School of Art of Xi'an University of Architecture and Technology. Zhao was responsible for the verification part, verifying that the inpainting results followed the artistic standards. He assisted in the inpainting of the murals under the ancient artistic style.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.
2. Wu, M.; Chang, X.; Wang, J. Fragments Inpainting for Tomb Murals Using a Dual-Attention Mechanism GAN with Improved Generators. *Appl. Sci.* **2023**, *13*, 3972. [CrossRef]
3. Gatys, L.A.; Ecker, A.S.; Bethge, M. Image style transfer using convolutional neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 2414–2423.
4. Levin, A.; Lischinski, D.; Weiss, Y. Colorization using optimization. In *ACM SIGGRAPH 2004 Papers*; ACM: New York, NY, USA, 2004; pp. 689–694.
5. Brox, T.; Van Den Boomgaard, R.; Lauze, F.; Van De Weijer, J.; Weickert, J.; Mrázek, P.; Kornprobst, P. *Adaptive Structure Tensors and Their Applications*; Springer: Berlin/Heidelberg, Germany, 2006.
6. Semmo, A.; Limberger, D.; Kyprianidis, J.E.; Döllner, J. Image stylization by oil paint filtering using color palettes. In Proceedings of the Workshop on Computational Aesthetics, Girona, Spain, 18–20 May 2015; pp. 149–158.
7. Chen, Y.; Lai, Y.K.; Liu, Y.J. Cartoongan: Generative adversarial networks for photo cartoonization. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 9465–9474.
8. Wang, W.; Li, Y.; Ye, H.; Ye, F.; Xu, X. DunhuangGAN: A Generative Adversarial Network for Dunhuang Mural Art Style Transfer. In Proceedings of the IEEE International Conference on Multimedia and Expo (ICME), Taipei, Taiwan, 18–22 July 2022; pp. 1–6.
9. Karras, T.; Laine, S.; Aittala, M.; Hellsten, J.; Lehtinen, J.; Aila, T. Analyzing and improving the image quality of stylegan. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 8110–8119.
10. Park, T.; Efros, A.A.; Zhang, R.; Zhu, J.Y. Contrastive learning for unpaired image-to-image translation. In Proceedings of the Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, 23–28 August 2020; pp. 319–345.
11. Karras, T.; Laine, S.; Aila, T. A Style-Based Generator Architecture for Generative Adversarial Networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–19 June 2019; pp. 4401–4410
12. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. ImageNet classification with deep convolutional neural networks. *Commun. ACM* **2017**, *60*, 84–90. [CrossRef]
13. Ulyanov, D.; Lebedev, V.; Vedaldi, A.; Lempitsky, V. Texture networks: Feed-forward synthesis of textures and stylized images. *arXiv* **2016**, arXiv:1603.03417.
14. Johnson, J.; Alahi, A.; Li, F.F. Perceptual losses for real-time style transfer and super-resolution. In *Computer Vision–ECCV 2016: Proceedings of the 14th European Conference, Amsterdam, The Netherlands, 11–14 October 2016*; Proceedings, Part II 14; Springer: Berlin/Heidelberg, Germany, 2016; pp. 694–711.
15. Zhang, Y.; Hu, B.; Huang, Y.; Gao, C.; Wang, Q. Adaptive Style Modulation for Artistic Style Transfer. *Neural Process. Lett.* **2023**, *55*, 6213–6230. [CrossRef]
16. Fernandez-Fernandez, R.; Victores, J.G.; Gago, J.J.; Estevez, D.; Balaguer, C. Neural policy style transfer. *Cogn. Syst. Res.* **2022**, *72*, 23–32. [CrossRef]
17. Yu, X.; Zhou, G. Arbitrary style transfer via content consistency and style consistency. *Vis. Comput.* **2024**, *40*, 1369–1382. [CrossRef]
18. Li, X.; Pu, Y.-Y.; Zhao, Z.-P.; Xu, D.; Qian, W.-H. Content semantics and style features match consistent artistic style transfer. *J. Graph.* **2023**, *44*, 699.
19. Huang, X.; Belongie, S. Arbitrary style transfer in real-time with adaptive instance normalization. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 1501–1510.
20. Li, Y.; Fang, C.; Yang, J.; Wang, Z.; Lu, X.; Yang, M.H. Universal style transfer via feature transforms. In Proceedings of the 31st International Conference on Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; Volume 30.
21. Lu, M.; Zhao, H.; Yao, A.; Chen, Y.; Xu, F.; Zhang, L. A closed-form solution to universal style transfer. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 5952–5961.
22. Li, X.; Liu, S.; Kautz, J.; Yang, M.H. Learning linear transformations for fast image and video style transfer. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 3809–3817.

23. An, J.; Huang, S.; Song, Y.; Dou, D.; Liu, W.; Luo, J. Artflow: Unbiased image style transfer via reversible neural flows. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 862–871.
24. Sheng, L.; Lin, Z.; Shao, J.; Wang, X. Avatar-net: Multi-scale zero-shot style transfer by feature decoration. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 8242–8250.
25. Chen, T.Q.; Schmidt, M. Fast patch-based style transfer of arbitrary style. *arXiv* **2016**, arXiv:1612.04337.
26. Zhang, H.; Goodfellow, I.; Metaxas, D.; Odena, A. Self-attention generative adversarial networks. In Proceedings of the International Conference on Machine Learning, Long Beach, CA, USA, 9–15 June 2019; pp. 7354–7363.
27. Deng, Y.; Tang, F.; Dong, W.; Huang, H.; Ma, C.; Xu, C. Arbitrary video style transfer via multi-channel correlation. In Proceedings of the AAAI Conference on Artificial Intelligence, Virtual, 2–9 February 2021; Volume 35, pp. 1210–1217.
28. Park, D.Y.; Lee, K.H. Arbitrary style transfer with style-attentional networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 5880–5888.
29. Chen, H.; Zhao, L.; Wang, Z.; Zhang, H.; Zuo, Z.; Li, A.; Xing, W.; Lu, D. Artistic style transfer with internal-external learning and contrastive learning. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 26561–26573.
30. Park, S.; Yoo, J.; Cho, D.; Kim, J.; Kim, T.H. Fast adaptation to super-resolution networks via meta-learning. In Proceedings of the Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, 23–28 August 2020; pp. 754–769.
31. Wu, H.; Qu, Y.; Lin, S.; Zhou, J.; Qiao, R.; Zhang, Z.; Xie, Y.; Ma, L. Contrastive learning for compact single image dehazing. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 19–25 June 2021; pp. 10551–10560.
32. Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative adversarial networks. *Commun. ACM* **2020**, *63*, 139–144. [CrossRef]
33. Deng, Y.; Tang, F.; Dong, W.; Sun, W.; Huang, F.; Xu, C. Arbitrary style transfer via multi-adaptation network. In Proceedings of the 28th ACM International Conference on Multimedia, Seattle, WA, USA, 12–16 October 2020; pp. 2719–2727.
34. Luo, X.; Han, Z.; Yang, L.; Zhang, L. Consistent style transfer. *arXiv* **2022**, arXiv:2201.02233.
35. Ma, Y.; Zhao, C.; Li, X.; Basu, A. RAST: Restorable arbitrary style transfer via multi-restoration. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, 3–7 January 2023; pp. 331–340.
36. Deng, Y.; Tang, F.; Dong, W.; Ma, C.; Pan, X.; Wang, L.; Xu, C. Stytr2: Image style transfer with transformers. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 11326–11336.
37. Wu, X.; Hu, Z.; Sheng, L.; Xu, D. Styleformer: Real-time arbitrary style transfer via parametric style composition. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 14618–14627.
38. Dinh, L.; Krueger, D.; Bengio, Y. Nice: Non-linear independent components estimation. *arXiv* **2014**, arXiv:1410.8516.
39. Lucey, P.; Cohn, J.F.; Kanade, T.; Saragih, J.; Ambadar, Z.; Matthews, I. The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops, San Francisco, CA, USA, 13–18 June 2010; pp. 94–101.
40. Salimans, T.; Goodfellow, I.; Zaremba, W.; Cheung, V.; Radford, A.; Chen, X. Improved techniques for training GANs. In Proceedings of the 30th International Conference on Neural Information Processing Systems, Barcelona, Spain, 5–10 December 2016; pp. 2234–2242.
41. Szegedy, C.; Ioffe, S.; Vanhoucke, V.; Alemi, A. Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning. In Proceedings of the AAAI Conference on Artificial Intelligence, San Francisco, CA, USA, 4–9 February 2017.
42. Zhang, R.; Isola, P.; Efros, A.A.; Shechtman, E.; Wang, O. The unreasonable effectiveness of deep features as a perceptual metric. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 586–595.
43. Wu, L.; Zhang, X.; Chen, H.; Wang, D.; Deng, J. VP-NIQE: An opinion-unaware visual perception natural image quality evaluator. *Neurocomputing* **2021**, *463*, 17–28. [CrossRef]
44. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27– 30 June 2016; pp. 770–778.
45. Abdal, R.; Qin, Y.; Wonka, P. Image2stylegan: How to embed images into the stylegan latent space? In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 4432–4441.
46. White, T. Sampling generative networks: Notes on a few effective techniques CoRR. *arXiv*, **2016**, arXiv:1609.04468.
47. Hang, T.; Yang, H.; Liu, B.; Fu, J.; Geng, X.; Guo, B. Language-guided face animation by recurrent StyleGAN-based generator. *IEEE Trans. Multimed.* **2023**, *25*, 9216–9227. [CrossRef]
48. Xie, S.; Tu, Z. Holistically-nested edge detection. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 1395–1403.
49. Ibraheem, N.A.; Hasan, M.M.; Khan, R.Z.; Mishra, P.K. Understanding color models: A review. *ARPN J. Sci. Technol.* **2012**, *2*, 265–275.
50. Zhou, B.; Zhao, H.; Puig, X.; Xiao, T.; Fidler, S.; Barriuso, A.; Torralba, A. Semantic understanding of scenes through the ade20k dataset. *Int. J. Comput. Vis.* **2019**, *127*, 302–321. [CrossRef]
51. Hore, A.; Ziou, D. Image quality metrics: PSNR vs. SSIM. In Proceedings of the 20th International Conference on Pattern Recognition, Istanbul, Turkey, 23–26 August 2010; pp. 2366–2369.

52. Xiong, R.; Yang, Y.; He, D.; Zheng, K.; Zheng, S.; Xing, C.; Zhang, H.; Lan, Y.; Wang, L.; Liu, T. On layer normalization in the transformer architecture. In Proceedings of the International Conference on Machine Learning, Virtual Event, 13–18 July 2020; pp. 10524–10533.
53. Zhang, Y.; Zhou, D.; Hooi, B.; Wang, K.; Feng, J. Expanding small-scale datasets with guided imagination. *arXiv* **2022** arXiv:2211.13976.
54. Zhu, J.Y.; Park, T.; Isola, P.; Efros, A.A. Unpaired image-to-image translation using cycle-consistent adversarial networks. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2223–2232.
55. He, B.; Gao, F.; Ma, D.; Shi, B.; Duan, L.Y. Chipgan: A generative adversarial network for chinese ink wash painting style transfer. In Proceedings of the 26th ACM international conference on Multimedia, Seoul, Republic of Korea, 22–26 October 2018; pp. 1172–1180.
56. Zheng, C.; Cham, T.J.; Cai, J. The Spatially-Correlative Loss for Various Image Translation Tasks. *Methods* **2021**, *56*, 22.
57. Dai, S.; Ye, K.; Zhao, K.; Cui, G.; Tang, H.; Zhan, L. Constrained Multiview Representation for Self-supervised Contrastive Learning. *arXiv* **2024**, arXiv:2402.03456.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.