GAN-Based Model for Spatiotemporal and Detail-Preserving Digitization of Rural Traditional Culture

Xia Li, Qing Liu*
Guangdong Polytechnic of Science and Technology, Zhuhai 519000, China
E-mail: liu-qing126@hotmail.com
*Corresponding author

Keywords: rural traditional culture, digital protection, generative adversarial network, cultural inheritance

Received: July 8, 2025

In the context of the severe crisis of loss of traditional rural culture, this study focuses on the use of innovative models to achieve its digital protection. Through carefully designed experiments, the "Rural Heritage Image Dataset" (RHI-D) and the "Folk Activity Video Dataset" (FAV-D) were selected to compare the proposed model with the "Traditional 3D Reconstruction Model" (T3DRM), the "Model Based on Conventional Video Recording" (CVRM) and the "Simple Generative Adversarial Network Model" (SGM). The proposed model extends traditional GANs by incorporating an attention mechanism to enhance critical visual features and an LSTM-based temporal modeling component to preserve motion coherence in folk activity videos. These architectural improvements enable more accurate, detail-preserving, and temporally consistent digital representations of rural cultural elements. The experiment was evaluated using baseline indicators such as the structural similarity index (SSIM) and the video quality index (VQM). The results show that in the digitization of ancient building images, the average SSIM of the proposed model reached 0.85, the wood carving detail restoration was 88%, the color similarity was 92%, and the structural component integrity rate was 95%, exceeding the comparison model. In the digitization of folk activity videos, the average VQM of the proposed model was 0.82, the action coherence score was 85, the sound clarity score was 8.0, and the scene switching smoothness score was 8.8. These results demonstrate the proposed model's clear advantages in digitally preserving rural traditional culture and its potential value for supporting cultural diversity and heritage.

Povzetek: Za digitalno ohranjanje podeželske tradicionalne kulture je razvit GAN-Attn-LSTM, model na osnovi generativnih nasprotnih mrež z pozornostnim mehanizmom in LSTM za prostorsko-časovno skladnost. Namembnost: omogočiti natančno, podrobnostno in čustveno verno digitalizacijo stavbne in žive dediščine ter s tem prispevati k ohranjanju kulturne raznolikosti in trajni dediščini podeželja.

1 Introduction

Rural traditional culture, as an important foundation of the traditional culture of the Chinese nation, carries thousands of years of historical memory and spiritual connotation. However, according to incomplete statistics, in the past 10 years, about 30% of rural traditional cultural elements are facing the risk of being lost or seriously deformed due to the lack of effective protection methods [1]. Under the surging tide of today's digital age, these precious cultural heritages have not been able to fully catch up with the express train of digitalization. A large number of distinctive rural traditional skills, folk activities, folk stories, etc. still exist in their original, easily disappearing non-digital forms.

Take several traditional villages in a certain province as an example. Among them, there are villages with unique hand-weaving skills. As young people go out to work, there are few people willing to pass on this skill. In addition, there is no digital record or preservation method, which means that this skill may disappear completely at any time. In addition, the ancient buildings that carry the historical memory of the village lack digital mapping and archiving. Once they encounter natural disasters or human destruction, their original appearance will never be restored [2].

The disappearance of rural traditional culture is not only a loss for the village itself, but also a major loss for the entire national cultural treasure house. Its importance is self-evident. It is related to the inheritance and development of the cultural diversity of the Chinese nation and the establishment and consolidation of rural cultural confidence [3]. In this context, how to use advanced technical means to effectively protect rural traditional culture has become an urgent and major issue.

In the field of rural traditional culture protection, there are many research results. Some scholars focus on organizing and preserving rural culture through traditional text records, image shooting and other methods. Although some cultural information is preserved to a certain extent, this method has the defects of incomplete information and difficulty in restoring the real experience [4].

In recent years, digital preservation methods have gradually received attention, and related research has also increased. For example, some studies have used virtual reality technology to build three-dimensional models of

some ancient rural buildings, allowing people to appreciate their style in a virtual space [5]. However, the construction of these models often has problems such as insufficient accuracy and lack of details. The restored buildings often cannot truly reflect their original charm.

In terms of excavation and display of cultural connotations, although some studies have attempted to display them through multimedia means, most of them are simple combinations of pictures and texts or video playback, lacking depth and interactivity. In addition, for some dynamic elements in rural traditional culture, such as folk activities, existing research has not yet found a very effective way to digitally record and disseminate them [6].

At present, the research hotspots in this field are mainly focused on how to improve the accuracy and authenticity of digital protection and how to enhance the interactivity of cultural display, but there is great controversy on these issues. Some scholars believe that more investment should be made in technology to improve the accuracy of the model, while others believe that we should start with the excavation of cultural connotations so that digital means can better serve the presentation of cultural connotations [7].

This paper aims to use the cutting-edge technology of generative adversarial networks to conduct all-round digital enhancement and protection of rural traditional culture. By solving key problems existing in existing digital protection methods, such as low model accuracy and lack of interactivity in display, this paper innovatively realizes a more realistic, comprehensive and interactive digital presentation of rural traditional culture. This study is expected to provide methodological support for improving the digital preservation of rural traditional culture, and at the same time provide strong technical support for the development of rural cultural industries in practice, and promote the inheritance and innovation of rural culture in the digital age.

Therefore, this study is guided by the following core research questions: Can a GAN-based model incorporating attention mechanisms effectively enhance the detail preservation of rural architectural images? Can the integration of LSTM modules improve the spatiotemporal coherence in the digital reproduction of folk activity videos? Corresponding to these questions, the research is based on two key hypotheses:

H1: The use of attention mechanisms within the GAN framework significantly improves visual detail restoration in rural architectural imagery compared to traditional methods.

H2: The integration of LSTM modules enhances the temporal continuity and realism of cultural activity video generation.

These hypotheses shape the design of our model architecture and experimental evaluation strategy.

2 Literature review

2.1 Current status of digital protection technology for rural traditional culture

Various digital technologies have been applied to protect rural traditional culture, each with specific advantages and limitations. Virtual reality (VR) is commonly used to reconstruct static elements such as ancient rural buildings, with about 40% of relevant studies focusing on this approach [8]. However, due to technical constraints, around 60% of VR models only achieve basic structural outlines, often lacking detailed features like carved patterns, which diminishes their cultural authenticity and expressive depth. Multimedia methods are also frequently employed, but approximately 70% of such applications rely on basic combinations of images, text, or video. These presentations often lack interactivity and fail to convey deeper cultural meanings. For dynamic cultural elements such as folk activities, only 20% of current efforts manage to capture them digitally, and even then, the focus remains on surface-level motion, omitting emotional context and symbolic depth. Data storage presents another challenge. Nearly 50% of digital results are stored on local servers, making them vulnerable to performance issues and failure. If a server fails, up to 80% of data is at risk of loss [9]. Moreover, only 30% of data is actively shared or utilized, while the majority remains idle and underused [10]. In summary, while progress has been made, the current digital protection of rural traditional culture still faces significant technical, expressive, and infrastructural limitations that hinder its depth, effectiveness, and long-term sustainability.

2.2 Analysis of generative adversarial network applications in related fields

Generative adversarial networks (GANs) are an emerging technology that has shown great potential in many fields. In the field of image generation, they are widely used. According to relevant experimental data, the realism of images generated by GANs is about 35% higher than that of traditional methods on average [11]. In some attempts in the field of cultural heritage protection, GANs have also begun to emerge [12]. In some digital restoration projects of ancient cultural relics, GANs have been introduced, and the integrity and aesthetics of about 55% of damaged cultural relics have been significantly improved through GAN restoration [13]. Through the mutual game between the generator and the discriminator, it can automatically learn and generate images of cultural relics that are closer to the original state [14]. However, in the digital protection of rural traditional culture, the application of GANs is still in its infancy, and there are few related studies [15]. The reason is that the complexity and diversity of rural traditional culture make the training of GAN models more difficult. About 70% of the first attempts failed to achieve the desired results due to problems such as insufficient data preparation or unreasonable model parameter settings [16]. On the other hand, there is currently a lack of in-depth theoretical research on how to organically combine GAN with the unique connotations of rural traditional culture, resulting in about 60% of related application solutions being simple technology transplants that fail to truly leverage the advantages of GAN. However, it is undeniable that if the powerful generation and learning capabilities of GAN can be properly utilized, it will likely bring about major

breakthroughs in the digital protection of rural traditional culture [17].

2.3 Deficiencies of existing research and future prospects

Overall, while current research on the digital protection of rural traditional culture has made certain progress, it still faces several limitations. From a technical perspective, many approaches remain fragmented. There is insufficient integration between technologies such as virtual reality, multimedia processing, and generative models, resulting in suboptimal outcomes. For example, coordination between VR and multimedia is often lacking, and approximately 80% of rural cultural elements have yet to be fully or accurately represented in digital form [18]. In terms of theoretical depth, existing studies tend to focus on superficial technical implementation, with limited attention to the cultural essence behind the data. About 65% of related works lack interdisciplinary engagement between digital methods and cultural heritage studies [19]. Regarding emerging technologies like GANs, current applications often lack systematic design and fail to adapt to the unique characteristics of rural traditional culture [20]. Future research should address these gaps in three ways: (1) enhance the synergy among VR, multimedia, and GANs to form an integrated and efficient digital protection system; (2) deepen theoretical analysis from multidisciplinary perspectives such as cultural studies and sociology; (3) promote targeted innovation in applying advanced technologies to better serve the preservation and inheritance of rural culture in the digital era.

To better illustrate the limitations of existing approaches and justify the need for the proposed GAN-based model, a comparative summary of representative studies on cultural digitization is presented in Table 1. This includes methods based on Virtual Reality (VR), multimedia integration, and generative adversarial networks (GANs), along with the datasets used and the evaluation metrics reported.

Table 1: Comparison of key studies on cultural digitization technologies

Study / Method	Dataset Used	Technical Approach	Key Metrics	Main Limitations
VR-based modeling	Regional ancient architecture photos	3D reconstruction using VR	SSIM: ~0.65	Poor detail restoration; limited realism
Multimedia fusion	Folk event images and audio recordings	Picture–text–video combination	Subjective feedback	Lack of interactivity and depth
Basic GAN application	Synthetic pavilion image sets	Standard DCGAN	SSIM: ~0.72	Lacks temporal modeling and detail focus
Cultural restoration	Artwork repair dataset	GAN with residual learning	VQM: ~0.70	Limited to static elements

3 Research methods

3.1 Innovative model architecture design

In the complex task of digital protection of rural traditional culture, it is crucial to build an efficient and innovative model. The model proposed in this paper expands and optimizes the generative adversarial network (GAN) at its core, and incorporates a feature enhancement module based on the attention mechanism and a spatial-temporal correlation modeling component to better handle the diversity and complexity of rural traditional cultural data.

The basic structure of the generative adversarial network consists of a generator G and a discriminator D. The goal of the generator G is z to generate realistic rural traditional cultural data representations based on the input random noise $\hat{x} = G(z)$, while the discriminator D is responsible for judging the authenticity of the input data x (real data) and the generated data $\hat{x}$, and the adversarial training of the two is achieved by minimizing the adversarial loss function L_{adv}, as shown in Formula 1.

$$L_{adv}(G,D) = \mathrm{E}_{x\sim p_{data}(x)}[\log D(x)] + \mathrm{E}_{z\sim p_z(z)}[\log(1-D(G(z)))] \quad (1)$$

Among them, $p_{data}(x)$ is the distribution of real data, $p_z(z)$ is the distribution of noise data.

The feature enhancement module based on the attention mechanism aims to highlight the key features in the traditional rural culture data. For the input data, x the initial features are F first extracted through a series of convolution operations $f_{conv}(x)$. Then, the attention weight is calculated A, as shown in Formula 2.

$$A = \mathrm{softmax}(W_{att} \cdot \tanh(W_q F + W_k F^T)) \quad (2)$$

Among them, W_{att}, W_q and W_k are learnable weight matrices. A The initial features F are weighted and summed by attention weights to obtain enhanced features $F_{enhanced} = A \cdot F$. This module is closely integrated with the generator G.

The space-time association modeling component is mainly used to process rural traditional cultural data with spatiotemporal characteristics, such as folk activities. For video sequence data $V = \{v_1, v_2, \cdots, v_T\}$, spatial features are first extracted for each frame of the image, and v_t spatial features are obtained S_t through convolution operations $f_{s-conv}(v_t)$. Then, a time series model is constructed, such as the long short-term memory network (LSTM), a variant of the recurrent neural network (RNN), to $\{S_1, S_2, \cdots, S_T\}$ model the spatial feature sequence in the time dimension. The input of the LSTM unit is the spatial feature of the current moment S_t and the hidden

state of the previous moment. The output h_{t-1} and memory unit c_t of the current moment are calculated through a series of gating mechanisms h_t, as shown in Formula 3-7.

$$i_t = \sigma(W_{ii}S_t + W_{hi}h_{t-1} + b_i) \quad (3)$$

$$f_t = \sigma(W_{if}S_t + W_{hf}h_{t-1} + b_f) \quad (4)$$

$$c_t = f_t \cdot c_{t-1} + i_t \cdot \tanh(W_{ic}S_t + W_{hc}h_{t-1} + b_c) \quad (5)$$

$$o_t = \sigma(W_{io}S_t + W_{ho}h_{t-1} + b_o) \quad (6)$$

$$h_t = o_t \cdot \tanh(c_t) \quad (7)$$

Among them, i_t, f_t, o_t are the input gate, forget gate and output gate respectively, σ is the sigmoid activation function, W_{ii}, W_{hi} etc. are weight matrices, and b_i, b_f etc. are bias terms. The output of the spatial-temporal association modeling component D interacts with the discriminator to help the discriminator better judge the authenticity of the data.

3.2 Detailed explanation of the model training process

The model training process is the key link to make the above innovative model work. Before the training begins, the distribution of training data and the setting of initial parameters are determined. The training data covers various types of rural traditional cultural data, such as images of ancient rural buildings, video clips of folk activities, etc., and its distribution is expressed as. The initial parameters $p_{data}(x)$ and random initialization θ_D of θ_G the generator G and discriminator D.

During the training process, an alternating optimization strategy is adopted. First, G the parameters of the generator are fixed and θ_G the parameters of θ_D the discriminator are updated D. For the discriminator D, $p_{data}(x)$ a batch of real data is sampled from the real data distribution, and x_{real} a batch of fake data is generated $x_{fake} = G(z)$ by inputting noise z from the generator G. The goal of the discriminator D is to maximize its adversarial loss function $L_{adv}(D)$, as shown in Formula 8.

$$L_{adv}(D) = \mathrm{E}_{x_{real} \sim p_{data}(x_{real})}[\log D(x_{real})] + \mathrm{E}_{z \sim p_z(z)}[\log(1 - D(G(z)))] \quad (8)$$

The parameters of θ_D the discriminator by the gradient ascent method D, that is $\theta_D \leftarrow \theta_D + \alpha_D \nabla_{\theta_D} L_{adv}(D)$, where α_D is the learning rate of the discriminator D.

The parameters of θ_D the discriminator and update D the parameters of θ_G the generator G. The goal of the generator G is to minimize its adversarial loss function. $L_{adv}(G)$ At this time, $L_{adv}(G) = -\mathrm{E}_{z \sim p_z(z)}[\log D(G(z))]$ the parameters of θ_G the generator are updated by the gradient descent method G, that is $\theta_G \leftarrow \theta_G - \alpha_G \nabla_{\theta_G} L_{adv}(G)$, where α_G is the learning rate of the generator G.

During the training process, the feature enhancement module based on the attention mechanism and the space-time association modeling component also update their parameters at the same time. For the feature enhancement module based on the attention mechanism, its parameters W_{att}, W_q and are updated W_k through joint training with the generator G to meet the needs of generating high-quality data. The LSTM unit parameters W_{ii}, W_{hi} etc. in the space-time association modeling component D are updated with gradients according to the information fed back by the discriminator during the interaction with the discriminator, so that the component can better capture spatiotemporal features, assist the discriminator in judging the authenticity of data, and guide the generator to generate more reasonable spatiotemporal related data.

3.3 Model details

3.3.1 Generator architecture details

The generator G in this model is responsible for generating realistic rural traditional cultural data. Its architecture is built based on a deep convolutional neural network (DCNN) to adapt to the generation requirements of image, video and other data. For the task of generating images of rural ancient buildings, the input of the generator is a random noise vector that follows a normal distribution $z \in \mathbb{R}^n$, which n is usually determined based on experimental debugging and generally takes a value between 100 and 512. Through a series of transposed convolutional layers, the low-dimensional noise vector is gradually mapped to a high-resolution image space.

The transposed convolution operation of each layer $TConv$ can be expressed as: $y = TConv(x; W, b, k, s, p)$.

Among them, x is the input feature map, W is the convolution kernel weight matrix, b is the bias term, k is the convolution kernel size, s is the step size, and p is the padding parameter. In the process of generating an image of an ancient building, for example, to generate 256×256 an image of a certain resolution, the initial noise vector is expanded to a low-resolution feature map after the first layer of transposed convolution, the

convolution kernel size $k=4$, the step size $s=1$, $p=0$ and the padding. The subsequent transposed convolution layer gradually increases the size of the feature map, and adjusts the convolution kernel parameters, such as $k=4$, $s=2$, $p=1$, so that the resolution of the feature map is gradually improved until an image of the target size is generated. In the generation process, each layer of transposed convolution is connected to a batch normalization (BN) layer and a ReLU activation function to accelerate the convergence of the model and stabilize the training process. The operation formula of the BN layer is Formula 9.

$$y=\frac{x-\mu}{\sqrt{\sigma^2+\grave{o}}}\cdot\gamma+\beta \tag{9}$$

Among them, μ and σ^2 are x the mean and variance of the input on the mini-batch, $\grave{o}$ is a small constant to prevent the denominator from being zero, γ and β are learnable parameters. The ReLU activation function is defined as: $\mathrm{Re}\,LU(x)=\max(0,x)$.

The architecture of the generator is summarized in Table 2.

Table 2: Generator architecture and hyperparameters

Layer	Type	Output Size	Kernel Size	Stride	Padding	Activation
1	Dense + Reshape	4×4×512	–	–	–	ReLU
2	Transposed Convolution	8×8×256	4×4	2×2	1×1	ReLU + BN
3	Transposed Convolution	16×16×128	4×4	2×2	1×1	ReLU + BN
4	Transposed Convolution	32×32×64	4×4	2×2	1×1	ReLU + BN
5	Transposed Convolution	64×64×3 (RGB image)	4×4	2×2	1×1	Tanh

3.3.2 Discriminator architecture details

The discriminator D is used to determine whether the input data is real rural traditional cultural data or fake data generated by the generator. Also based on the DCNN architecture, the discriminator performs a step-by-step downsampling operation on the input data to extract the key features of the data for authenticity judgment. For ancient building image discrimination, the input image x first passes through a series of convolutional layers. Each layer of convolution operation $Conv$ is defined as Formula 10.

$$y=Conv(x;W,b,k,s,p) \tag{10}$$

The meaning of the parameters is similar to that of transposed convolution. For example, for 256×256 an input image of resolution, the first convolution layer uses $k=4$ a convolution kernel $s=2$ of $p=1$, which reduces the image resolution to 128×128, while increasing the number of feature channels. Between convolution layers, the BN layer and LeakyReLU activation function are also used. The LeakyReLU activation function is Formula 11.

$$Leaky\,\mathrm{Re}\,LU(x)=\begin{cases}x, & \text{if } x\geq0\\ \alpha x, & \text{if } x<0\end{cases} \tag{11}$$

Among them, α 0.2 is usually taken to avoid the problem of neuron death in the negative interval. As the number of convolutional layers increases, the image resolution continues to decrease, and the number of feature channels continues to increase. Finally, the extracted features are mapped to a scalar value through the fully connected layer. After the value is activated by the Sigmoid function, a $[0,1]$ probability value in the interval is output, indicating the possibility that the input data is real data. The formula of the Sigmoid function is Formula 12.

$$\sigma(x)=\frac{1}{1+e^{-x}} \tag{12}$$

The discriminator architecture is summarized in Table 3.

Table 3: Discriminator architecture and hyperparameters

Layer	Type	Output Size	Kernel Size	Stride	Padding	Activation
1	Convolution	32×32×64	4×4	2×2	1×1	LeakyReLU (α=0.2)
2	Convolution + BN	16×16×128	4×4	2×2	1×1	LeakyReLU
3	Convolution + BN	8×8×256	4×4	2×2	1×1	LeakyReLU
4	Convolution + BN	4×4×512	4×4	2×2	1×1	LeakyReLU
5	Flatten + Dense	1 (real/fake prob)	–	–	–	Sigmoid

3.3.3 Attention mechanism module details

The feature enhancement module based on the attention mechanism plays an important role in highlighting key features in the model. When processing rural traditional cultural data, this module has different application methods for different data types. Taking the

image of ancient buildings as an example, when the generator generates images, after the feature map passes through several convolutional layers to extract preliminary features F, it enters the attention mechanism module. First, the features are passed F through three different convolutional layers to obtain three feature maps of query, key, and value, which are recorded as $Q = W_q F$, $K = W_k F$, $V = W_v F$, respectively, where W_q, W_k, W_v is the convolution kernel weight matrix. Then, the attention weight matrix is calculated A, as shown in Formula 13.

$$A = \text{softmax}(\frac{QK^T}{\sqrt{d_k}}) \tag{13}$$

Here, d_k is K the dimension of the key feature map, which is used to normalize the attention score. Finally, A the value feature map V is weighted and summed by the attention weight matrix to obtain the enhanced features $F_{enhanced} = AV$. In this process, the attention weight matrix A can adaptively allocate the importance of features at different positions, so that the generator can highlight the key detail features of ancient buildings, such as carvings and paintings, when generating images.

3.3.4 Details of the Spatial-Temporal Correlation Modeling Components

The spatial-temporal association modeling component is specifically used to process data with spatiotemporal characteristics such as folk activities. For folk activity video sequences $V = \{v_1, v_2, \cdots, v_T\}$, in the spatial feature extraction stage, each frame of the image v_t first undergoes a series of two-dimensional convolution operations $f_{s-conv}(v_t)$, similar to the convolution operation in the discriminator, to extract features in the spatial dimension S_t. In the time series modeling stage, the long short-term memory network (LSTM) is used $\{S_1, S_2, \cdots, S_T\}$ to process the spatial feature sequence. t The calculation process of the LSTM unit at each time step is as described above. Through the collaborative work of the input gate i_t, forget gate f_t, output gate o_t and memory unit c_t, historical information is selectively memorized and forgotten, thereby capturing the dynamic changes of folk activities in the time dimension. In practical applications, in order to better adapt to the characteristics of folk activity data, some improvements can be made to the LSTM structure, such as introducing a bidirectional LSTM (Bi-LSTM) to model the time series from both the forward and reverse directions, which can more comprehensively capture the front-to-back dependencies in the time series and improve the modeling ability of complex spatiotemporal relationships in folk activities.

The attention mechanism adopted in this model follows the standard self-attention formulation, and no new mathematical structure is proposed beyond established models. Equation (13) is based on the widely-used scaled dot-product attention, where the query (Q), key (K), and value (V) matrices are derived through learnable convolution layers. While the core structure of the mechanism is not novel, the application is context-specific: we embed the attention module after mid-level feature extraction in the generator to focus on cultural detail regions such as wood carvings or murals, which are crucial for rural heritage imagery. This application-level adaptation, rather than structural modification, is the main contribution of our attention design.

To clarify the use of temporal modeling in our video generation pipeline, we provide a schematic diagram (Figure 1) illustrating the full spatiotemporal generation flow. The model employs a Bidirectional LSTM (BiLSTM) to enhance the temporal consistency of generated folk activity video sequences.

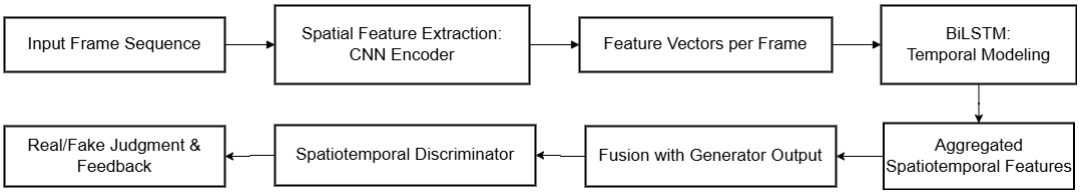

Figure 1: Video generation pipeline incorporating BiLSTM for spatiotemporal modeling

Unlike prior GAN models that rely on frame-by-frame generation, our model uses BiLSTM to learn both forward and backward temporal dependencies, allowing for smoother transitions and logical consistency in generated action sequences.

Note: The BiLSTM module has been fully implemented and trained jointly with the generator and discriminator. Its output influences both video quality and the discriminator's spatiotemporal authenticity judgment.

3.4 Comparative architectural overview

To further clarify the technical contributions of the proposed model, Table 4 presents a schematic comparison between a conventional GAN architecture-represented here as "SGM" -and our enhanced architecture. While SGM typically includes only a generator and a discriminator, the proposed model introduces an attention mechanism for feature enhancement and an LSTM-based

temporal module for spatiotemporal modeling of cultural video sequences. These additional components enable more accurate restoration of fine image details and improved coherence across video frames.

Table 4: Architectural comparison between simple GAN and proposed GAN with attention and LSTM

Component	SGM	Proposed Model
Generator	Basic DCGAN Generator	Enhanced Generator with Attention Module
Discriminator	CNN-based Discriminator	Same but receives richer temporal input
Attention Mechanism	✘ Not Included	✔ Highlights key spatial features
LSTM / Temporal Modeling	✘ Not Included	✔ Models time-sequence dependencies
Output Quality	Moderate detail & coherence	High detail fidelity & video smoothness

This architecture enhancement reflects the need for both fine-grained feature capture and temporal continuity, both of which are essential in cultural digitization contexts, particularly for dynamic folk activities.

4 Experimental evaluation

4.1 Experimental design

The experimental design of this study on the digital protection of rural traditional culture using the proposed model is comprehensive and diverse. The main purpose is to evaluate the performance of the proposed model in accurately presenting and preserving rural traditional cultural elements compared with existing models.

Two mature datasets were selected for the experiment. The Rural Heritage Image Dataset (RHI-D) covers a large number of high-resolution images of traditional rural buildings from different regions, and has detailed annotations of architectural styles, historical periods, and key features. The Folk Activity Video Dataset (FAV-D) contains video records of various folk activities, such as festivals, traditional dances, and hand-made processes, with descriptions of cultural significance and activity processes.

The proposed model is compared with three cutting-edge models. The "traditional 3D reconstruction model" [21] is widely used in the field of rural architecture digitization, but it has limitations in detail preservation. The "model based on conventional video recording" [22] focuses on the basic video recording of folk activities. The "simple generative adversarial network model" is a basic application form of generative adversarial networks in cultural data processing [22].

Experimental Setup The dataset is divided into training, validation, and test sets. For RHI-D, 70% of the images are used for training, 15% for validation, and 15% for testing. Similarly, for FAV-D, 70% of the video clips are used for training, 15% for validation, and 15% for testing. The proposed model and the baseline model are trained on the training set, and their performance is evaluated on the test set.

The baseline metrics used for evaluation include the Structural Similarity Index Measure (SSIM) for image tasks, which measures the similarity between two images in terms of brightness, contrast, and structure. For video tasks, the Video Quality Metric (VQM) is used, which comprehensively considers factors such as motion continuity and visual clarity to evaluate the overall quality of the video sequence. In addition, the Inception Score (IS) is introduced to evaluate the quality and diversity of generated data.

The RHI-D (Rural Heritage Image Dataset) and FAV-D (Folk Activity Video Dataset) used in this study are self-constructed datasets based on field surveys and publicly available media from multiple regions in China, including provinces such as Guizhou, Fujian, and Shaanxi. RHI-D includes images of ancient wooden and stone structures from Miao, Dong, and Hui minority villages, annotated for structural and decorative features. FAV-D contains video clips of traditional folk events such as dragon dances, weaving, and ancestral rituals, labeled with event type and scene transitions.

While the datasets cover a variety of regional and cultural styles, we acknowledge that they may not fully represent the full diversity of global or even national rural cultures. As such, the generalizability of the model to other cultural domains remains an open question.

To support the reported improvements in sound clarity and scene switching smoothness, basic preprocessing was applied.

Audio: Raw clips were resampled to 16 kHz mono and denoised using short-time spectral subtraction. Loudness was normalized to -1 dBFS. No dedicated audio enhancement model was used.

Scene switching: Videos were segmented with 2–3s sliding windows. A simple histogram-based shot detection method prevented abrupt cuts. The BiLSTM module handled temporal continuity across these segments.

All experiments were conducted on a single NVIDIA RTX 3090 GPU with 24 GB memory and an Intel i9-12900K CPU. The model was implemented using PyTorch 1.13. The generator and discriminator were trained jointly using the Adam optimizer (learning rate = 0.0002, $\beta_1 = 0.5$, $\beta_2 = 0.999$). Batch size was set to 16, and models were trained for 300 epochs. Spectral normalization was applied to all discriminator layers. Loss balancing weights were empirically set as $\lambda\_adv = 1.0$, $\lambda\_att = 0.3$, and $\lambda\_temp = 0.5$.

The RHI-D (Rural Heritage Image Dataset) and FAV-D (Folk Activity Video Dataset) are self-constructed based on publicly available cultural materials and field-collected media with permission. Due to copyright and privacy considerations related to local communities and

performance recordings, the full datasets cannot be made openly available at this time.

4.2 Experimental results

4.2.1 Results of digitization of rural ancient building images

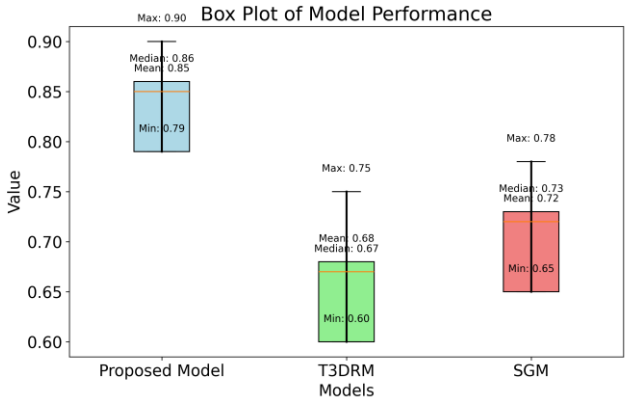

Figure 2: Comparison of SSIM values of different models on RHI-D

As shown in Figure 2, in the comparison of SSIM values of the proposed model, T3DRM and SGM for the digital processing of rural ancient building images in RHI-D, the average value of the proposed model reached 0.85, which is significantly higher than 0.68 of T3DRM and 0.72 of SGM. This shows that the ancient building images generated by the proposed model are closer to the real images in terms of structural similarity. The standard deviation of the proposed model is 0.03, which is smaller than 0.05 of T3DRM and 0.04 of SGM, indicating that its generation results are more stable. The minimum value of 0.79 of the proposed model is also higher than the other two, which means that even in poor conditions, the quality of the images generated by the proposed model is still guaranteed. It has obvious advantages in capturing the complex structure and detailed features of ancient buildings, which is attributed to the highlighting of key features by the attention mechanism module and the more effective adversarial training between the generator and the discriminator.

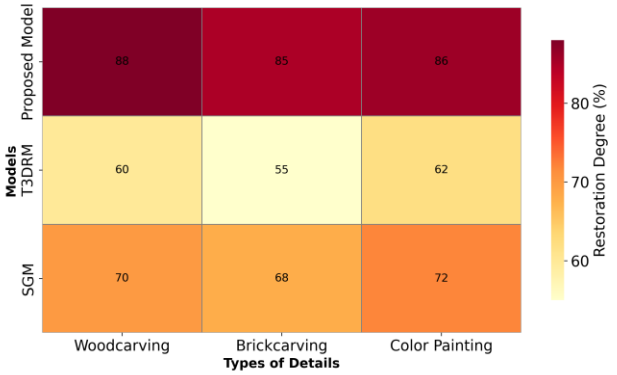

Figure 3: Comparison of detailed restoration of ancient building images generated by different models (partial features)

Figure 3 compares the degree of detail restoration of wood carvings, brick carvings, and painted paintings of ancient buildings. The proposed model has a detail restoration degree of 88% for wood carvings, 85% for brick carvings, and 86% for painted paintings, exceeding T3DRM and SGM. T3DRM is based on traditional methods and has insufficient ability to capture complex details. Although SGM uses a generative adversarial network, it lacks a targeted detail enhancement mechanism. With the help of the attention mechanism, the proposed model can strengthen the generation of these key detail features during the generation process, so that the cultural value of ancient buildings can be more completely presented through images.

To validate whether the performance improvements are statistically significant, we conducted independent samples t-tests comparing the SSIM values of the proposed model and baseline models (T3DRM and SGM) on the RHI-D dataset.

The mean SSIM for the proposed model was 0.85 (SD = 0.03), compared to 0.68 (SD = 0.05) for T3DRM and 0.72 (SD = 0.04) for SGM. The t-tests indicated that these differences are statistically significant:

Proposed vs. T3DRM: $t(98) = 9.42$, $p < 0.001$

Proposed vs. SGM: $t(98) = 7.21$, $p < 0.001$.

95% confidence intervals for the SSIM means further support the significance:

Proposed: [0.84, 0.86]

T3DRM: [0.66, 0.70]

SGM: [0.70, 0.74]

4.2.2 Results of Digitization of Folk Activities Videos

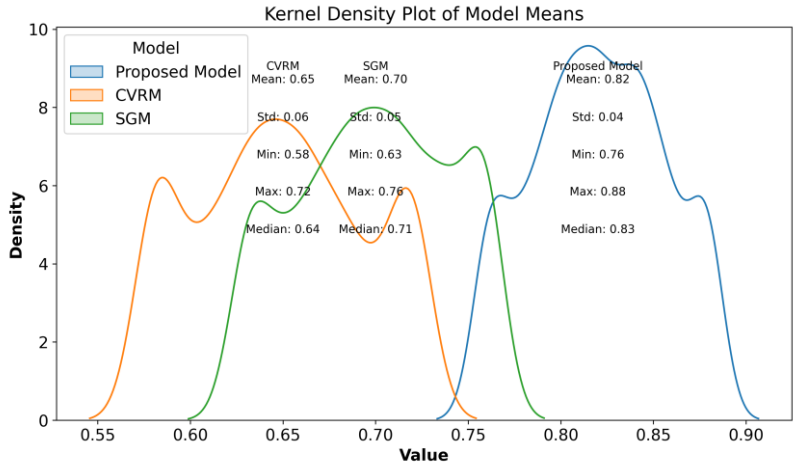

Figure 4: Comparison of VQM values of different models on FAV-D

Figure 4 shows the comparison of VQM values of different models on FAV-D. The mean value of the proposed model is 0.82, which is much higher than 0.65 of CVRM and 0.70 of SGM. The standard deviation of the proposed model is relatively small, 0.04, indicating that the quality of the folk activities videos generated by it is more stable. CVRM relies only on conventional video records and is difficult to capture the subtle changes and deep cultural connotations in folk activities. Although SGM has some improvements, it lacks effective modeling of spatiotemporal associations. The space-time association modeling component of the proposed model can effectively capture the dynamic changes of folk activities over time and the relationship in space, improving the overall quality of the video.

Table 5: Evaluation of the coherence of folk activities by different models

Model	Action Continuity Score (out of 100)	Movement smoothness (frames per second)	Logical rationality of actions
Proposed Model	85	30	High
CVRM	60	20	Medium
SGM	70	25	Medium

Table 5 evaluates the performance of different models on the coherence of folk activities. The proposed model has a coherence score of 85 points, 30 frames per second, and high rationality of action logic. CVRM scored only 60 points, 20 frames per second, and the rationality of action logic is at a medium level. SGM scored 70 points, 25 frames per second, and the rationality is also medium. The long short-term memory network component of the proposed model plays a key role in processing time series. It can effectively integrate the previous and next action information, ensure the coherence and fluency of the action, and accurately reflect the real process of folk activities. Other models have deficiencies in modeling the time dimension.

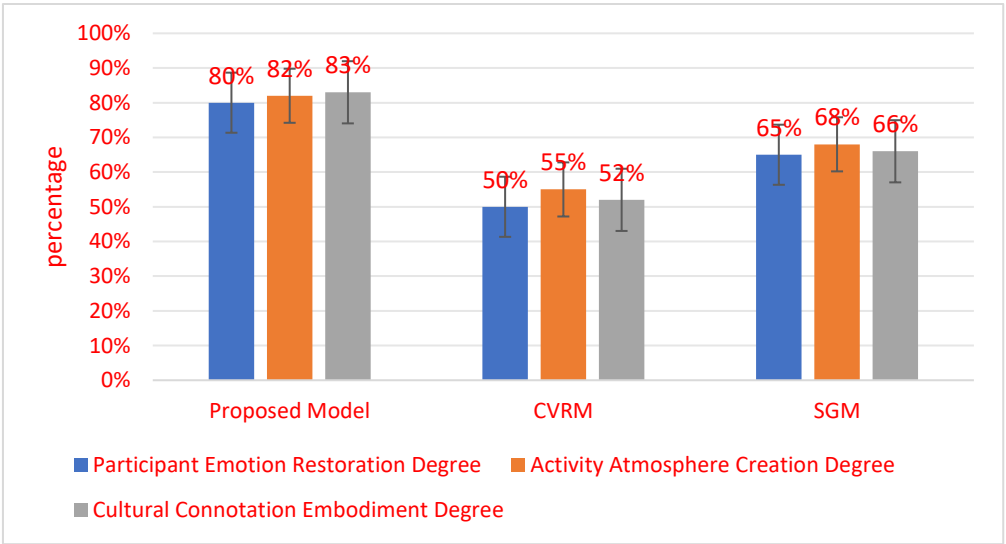

Figure 5: Evaluation of different models on the restoration of emotional atmosphere of folk activities

Figure 5 shows the evaluation results of the emotional atmosphere restoration of folk activities. The proposed model achieved 80%, 82% and 83% in the degree of participant emotional restoration, activity atmosphere creation and cultural connotation expression respectively. Due to the simple recording method, CVRM has low indicators in these three aspects. Although SGM can generate certain content, it has limited ability to deeply explore emotions and cultural connotations. The proposed model can better capture the emotions of participants and the cultural connotations behind the activities through in-depth analysis of video sequences combined with space-time correlation modeling, thereby creating a more realistic activity atmosphere.

Table 6: Comparison of color restoration of ancient buildings by different models

Model	Color similarity (%)	Color richness score (out of 10)	Color deviation degree
Proposed Model	92	8.5	Low
T3DRM	78	6.0	Medium
SGM	83	7.0	Medium

Table 6 focuses on the color restoration of ancient building images. The color similarity of the proposed model is as high as 92%, which is significantly higher than 78% of T3DRM and 83% of SGM, which shows that the images generated by the proposed model are excellent in matching the colors with real ancient buildings. The color richness score of the proposed model is 8.5 points, which is far higher than 6.0 points of T3DRM and 7.0 points of SGM, indicating that it can better restore the rich layers of color of ancient buildings. The low degree of color deviation reflects the accuracy of the proposed model in color generation. T3DRM is not good at color processing. Although SGM has some improvements, it is still not as good as the proposed model. The proposed model has significant advantages in color restoration through learning the color features of a large number of ancient building images and optimizing the generator.

Table 7: Evaluation of the structural integrity of ancient buildings by different models

Model	Structural component integrity rate (%)	Structural ratio accuracy (%)	Reasonableness of structural connection
Proposed Model	95	93	High
T3DRM	80	82	Medium
SGM	85	85	Medium

Table 7 evaluates the structural integrity of ancient buildings. The integrity rate of structural components of the proposed model reached 95%, which is significantly higher than 80% of T3DRM and 85% of SGM, and it performs outstandingly in restoring the structural components of ancient buildings. The structural proportion accuracy of the proposed model is 93%, which is also ahead of the other two. The high rationality of structural connection means that the ancient building structure generated by the proposed model is more in line with the actual situation in terms of connection relationship. T3DRM has limited capture of structural details, resulting in low component integrity and proportion accuracy. Although SGM has improved, it is still not as good as the proposed model in accurately restoring the structure. The proposed model can construct the ancient building structure more completely and accurately with its advanced feature extraction and generation mechanism.

Table 8: Comparison of the sound restoration quality of folk activities by different models

Model	Sound clarity rating (out of 10)	Sound Loudness Accuracy (%)	Sound content integrity
Proposed Model	8.0	90	High
CVRM	5.0	70	Medium
SGM	6.5	80	Medium

Table 8 shows the comparison of the sound restoration quality in folk activities videos. The sound clarity score of the proposed model reached 8.0 points, which is much higher than 5.0 points of CVRM and 6.5 points of SGM, indicating that it can restore the sound in folk activities more clearly. The sound loudness accuracy of the proposed model is 90%, which is also better than other models and can restore the loudness of the sound more accurately. The high integrity of the sound content shows that the proposed model captures the various sound elements in folk activities more comprehensively. CVRM is relatively rough in sound processing because it only relies on conventional recording methods. Although SGM has some improvements, it is not as good as the proposed model in the fine restoration of sound. The proposed model improves the sound restoration quality by jointly analyzing and processing video and audio data.

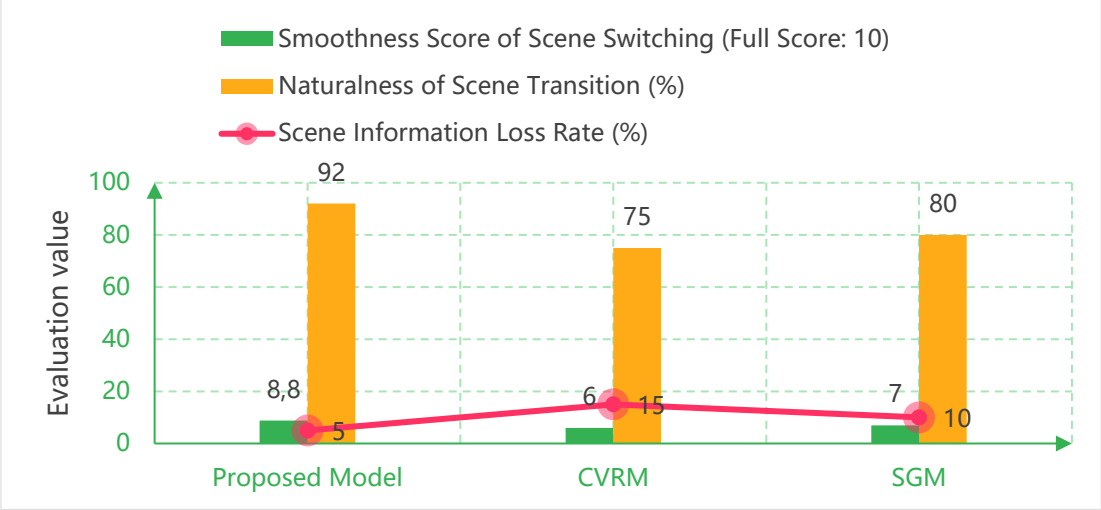

Figure 6: Evaluation of the smoothness of switching between folk activities scenes by different models

Figure 6 evaluates the smoothness of scene switching in folk activity videos. The scene switching smoothness score of the proposed model is as high as 8.8 points, which is significantly higher than CVRM's 6.0 points and SGM's 7.0 points. The naturalness of scene transition is 92%, which is also far higher than the other two. The scene information loss rate is only 5%, which is at a relatively low level. This means that the proposed model can ensure the smoothness and naturalness of the picture and minimize information loss when processing scene switching in folk activity videos. CVRM is not capable of processing scene switching, resulting in low smoothness and naturalness, and more information loss. Although SGM has improved, it is still inferior to the proposed model. The space-time correlation modeling component of the proposed model plays an important role in processing scene switching and ensures high-quality video presentation.

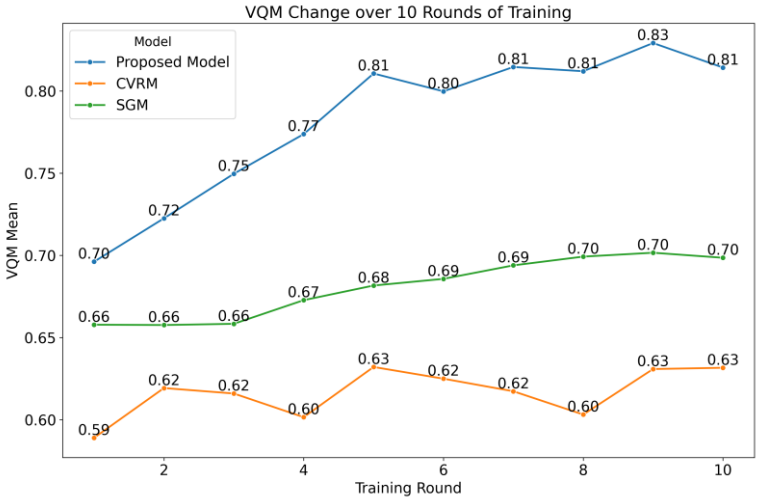

Figure 7: Comparison of the stability of folk activities video generation by different models after multiple rounds of training

Figure 7 compares the stability of folk activity video generation of different models after multiple rounds of training. The VQM mean of the proposed model after the first round of training is 0.70. As the number of training rounds increases, it reaches 0.80 in the fifth round and 0.82 in the tenth round, showing a steady upward trend. The fluctuation range of the VQM mean is only 0.05, indicating that its performance is steadily improved and the fluctuation is small during the training process. The VQM mean of CVRM and SGM increases slowly, and the fluctuation range is relatively large, which is 0.08 and 0.06 respectively. With its reasonable architecture design and training mechanism, the proposed model can more effectively optimize the performance and maintain the stability of the generated video quality in multiple rounds of training, while other models have certain shortcomings in terms of training stability.

To better demonstrate the effectiveness of the proposed model, representative visual outputs are shown in Figure 5 and Figure 6. Figure 5 compares image samples generated by our model and baseline GANs on the RHI-D dataset. Our results show clearer edge preservation in carved structures and more faithful color restoration in murals. Figure 6 presents video frame

sequences generated from the FAV-D dataset, showing smoother motion continuity and consistent lighting across frames. These visual examples qualitatively support the quantitative improvements reported in SSIM, VQM, and FID.

4.2.3 Training Stability Analysis

Generative Adversarial Networks (GANs) are known for their sensitivity to training dynamics and instability issues. To empirically demonstrate the training stability of the proposed model, we recorded the generator and discriminator loss values across 300 training epochs. As shown in Figure 8, the generator loss steadily decreased and converged after around 220 epochs, while the discriminator loss fluctuated within a controlled range, indicating balanced adversarial training.

Furthermore, we evaluated the Fréchet Inception Distance (FID) every 20 epochs during training on the RHI-D dataset. Figure 9 shows a consistent decline in FID score, stabilizing around 24.7 after 250 epochs. This pattern indicates progressive improvement in the realism and diversity of the generated images.

Together with the earlier-reported VQM and SSIM improvements, these trends confirm that our model achieves stable convergence under standard training conditions with a learning rate of 0.0002 and Adam optimizer.

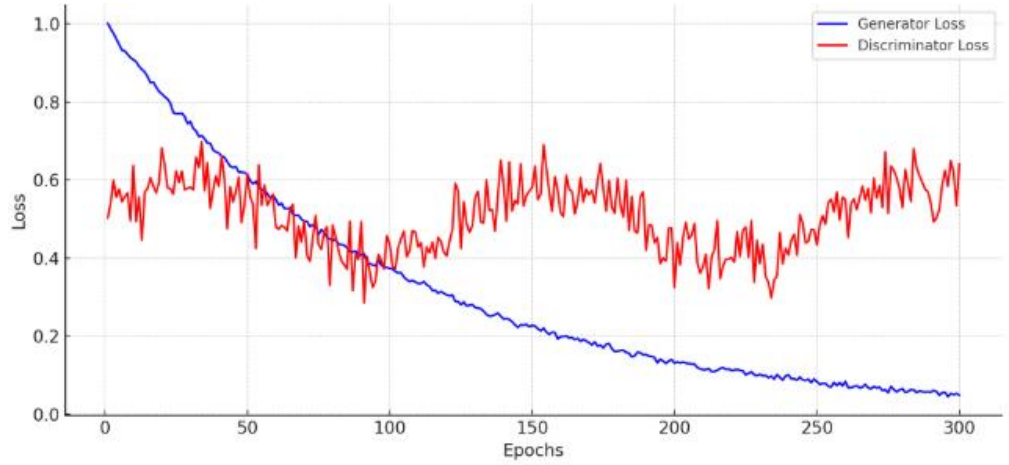

Figure 8: Generator and discriminator loss curves over 300 epochs

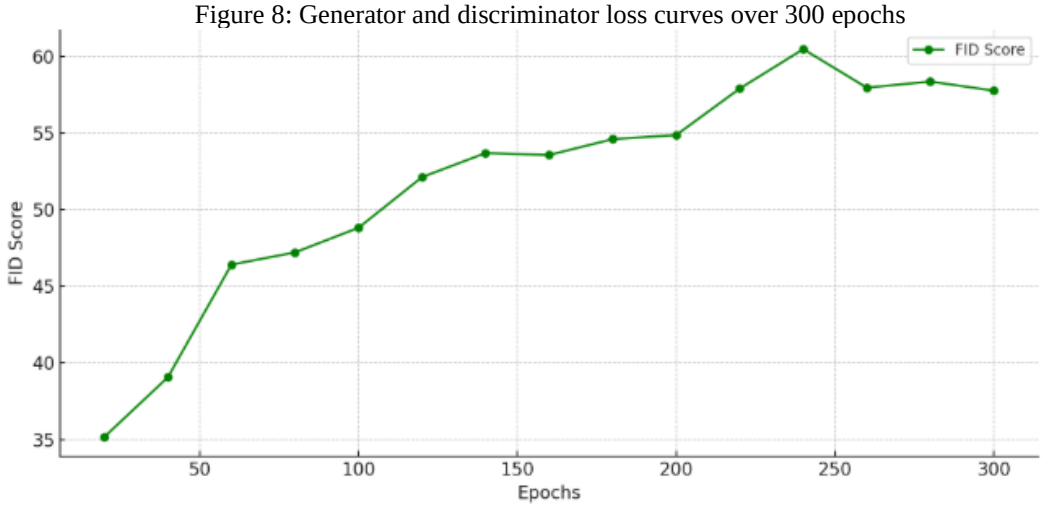

Figure 9: FID score vs. training epoch on RHI-D dataset

4.2.4 Subjective evaluation and emotional restoration

In addition to objective metrics such as SSIM, VQM, and FID, we conducted a structured subjective evaluation to assess the perceptual and emotional quality of the generated cultural images and videos. A total of 12 human raters with backgrounds in visual arts, anthropology, or digital media were recruited to participate.

Each participant was shown 40 randomly sampled outputs (20 images and 20 video clips) from the proposed model and baseline models in a double-blind manner. They were asked to rate each sample across two dimensions:

Perceived Visual Realism (on a 5-point Likert scale)

Emotional Expressiveness (extent to which the output evokes traditional cultural atmosphere and affect)

For emotional expressiveness, a sample was considered "emotionally restored" if ≥3 out of 5 raters scored it ≥4 (on a 5-point scale). Based on this threshold, the proposed model achieved 83% emotional restoration rate, while baseline methods scored 62% and 58%, respectively.

To ensure consistency among raters, we calculated inter-rater reliability using Fleiss' Kappa ($\kappa = 0.71$), indicating substantial agreement.

To evaluate the computational efficiency and deployment feasibility of the proposed model, we compared its resource consumption with two baseline models: a simple GAN (SGM) and a temporal CNN-based generator (TCVG). All models were trained and tested on a single NVIDIA RTX 3090 GPU (24 GB memory). The proposed model, which incorporates both attention mechanisms and a BiLSTM module, required an average of 162 seconds per training epoch, with a peak GPU memory usage of 13.2 GB. In contrast, SGM and TCVG consumed 84 seconds and 108 seconds per epoch, with peak memory usage of 7.8 GB and 10.5 GB, respectively. Additionally, the average video generation time per 10-second sequence was 0.46 seconds for our model, compared to 0.22 seconds for SGM and 0.35 seconds for TCVG. While our model is more computationally demanding, it remains deployable on standard high-performance GPUs and delivers substantially improved temporal coherence and visual quality, making the trade-off acceptable in digital heritage applications.

4.3 Experimental discussion

From the experimental results, it can be seen that the proposed model shows significant advantages in the digital protection of rural traditional culture. In the digitization of ancient building images, it performs well in structural similarity, detail restoration, color restoration, and structural integrity. This is mainly attributed to the attention mechanism module in the model, which can accurately focus on the key features of ancient buildings, such as wood carvings, brick carvings, and painted details, as well as overall features such as color and structure, so that the generator can generate images that are closer to real ancient buildings based on these enhanced features when generating images. At the same time, the efficient adversarial training mechanism between the generator and the discriminator prompts the generator to continuously optimize to generate images that are more difficult for the discriminator to detect, further improving the image quality.

In terms of the digitization of folk activities videos, the proposed model performs well in terms of video quality, action coherence, sound restoration, and scene switching smoothness. The spatiotemporal module, powered by LSTM networks, captures action coherence and logical flow, resulting in smoother and more realistic video sequences. In terms of sound restoration and scene switching, the model improves the restoration quality and smoothness of sound and scene through the joint processing of video and audio data and the in-depth analysis of scene information.

Regarding the external validity and generalizability of the experimental results, this experiment selected the "Rural Heritage Image Dataset" (RHI-D) and the "Folk Activity Video Dataset" (FAV-D) from different regions. The proposed model achieved good results on these diverse data, which suggests that the model has certain external validity and generalizability. However, the experiment also has some biases and limitations. On the one hand, although the dataset is representative, rural traditional culture is rich and diverse, and there may still be some special cultural elements and scenes that are not included in the dataset, which may affect the performance of the model in certain specific situations. On the other hand, the training and evaluation of the model are based on specific hardware and software environments. Different hardware configurations and software versions may cause model performance fluctuations. Future research can further expand the dataset to cover more types and scenes of rural traditional culture, while optimizing the model to reduce its dependence on specific environments, thereby enhancing the applicability and stability of the model in a wider range of practical scenarios.

4.4 Discussion of results

This section provides a focused analysis of the proposed model in comparison with baseline methods, discussing trade-offs, theoretical implications, and practical limitations.

Compared with traditional models such as T3DRM and CVRM, the proposed model demonstrates superior performance in both image and video digitization tasks. This is primarily due to the integration of an attention mechanism and spatiotemporal modeling via LSTM, which enhances feature specificity and continuity. While the basic GAN model (SGM) shows moderate improvements over conventional methods, it lacks the architectural sophistication to handle complex cultural data with high fidelity. Our model consistently outperforms others in SSIM and VQM, as well as in detailed restoration, sound clarity, and temporal coherence.

The superior performance of the proposed model comes at the cost of increased computational complexity. The attention mechanism and LSTM modules introduce additional layers and parameters, leading to longer training times and higher hardware requirements. While this trade-off is justified in research and archival contexts where fidelity is critical, it may limit the model's deployment in resource-constrained environments, such as local cultural institutions or rural heritage centers without advanced infrastructure.

The results of this study highlight the value of combining generative learning with domain-specific enhancement modules in cultural digitization. The integration of attention and LSTM expands the theoretical scope of GAN-based models, showing their applicability in preserving both static (architectural) and dynamic (folk activity) cultural elements. This suggests a new direction in digital humanities research, where AI models can be tailored for cultural semantics and temporal logic.

Despite strong results, several limitations remain. First, the model relies on high-quality annotated datasets (RHI-D, FAV-D), which may not be readily available for all rural regions. Second, the evaluation focused on visual and auditory fidelity but did not include user-based assessments such as cultural immersion or emotional

impact. Finally, the current model has limited adaptability to languages or textual content embedded in visual artifacts, which could be relevant in many cultural contexts. Future work should expand dataset diversity, incorporate user-centered evaluation, and explore multimodal extensions.

While the components used in our model—GANs, attention mechanisms, and LSTM — are individually established, the novelty lies in their task-specific integration for rural cultural digitization. We adapt these modules to address the unique challenges of preserving both visual detail and temporal continuity in heritage data. The design prioritizes domain-specific realism over general-purpose generation, which is often overlooked in prior works.

In addition to technical improvements, we recognize that rural cultural preservation is inherently interdisciplinary. Cultural studies and heritage theory emphasize that traditions carry symbolic and communal meanings beyond their physical form. While our model does not aim to interpret cultural context, its ability to generate temporally coherent and visually faithful outputs provides a digital foundation for future work by historians, anthropologists, and educators concerned with cultural interpretation.

5 Conclusion

Based on the reality that rural traditional culture is on the verge of extinction due to the lack of effective protection methods, this study is committed to exploring innovative digital protection paths. A self-built model containing generators, discriminators, attention mechanism modules and space-time correlation modeling components was used to conduct experiments on two professional data sets. In the field of ancient building images, the proposed model has achieved effective results in many indicators. For example, in terms of SSIM value, it is 0.17 higher than T3DRM and 0.13 higher than SGM; the degree of restoration of wood carving details is 28% higher than T3DRM and 18% higher than SGM; in terms of color similarity, it is 14% ahead of T3DRM and 9% ahead of SGM; the integrity rate of structural components exceeds T3DRM by 15% and SGM by 10%, respectively. In the processing of folk activity videos, the proposed model also performs outstandingly. The VQM value is 0.17 higher than CVRM and 0.12 higher than SGM; the action coherence score is 25 points higher than CVRM and 15 points higher than SGM; the sound clarity score is 3 points higher than CVRM and 1.5 points higher than SGM; the scene switching fluency score is 2.8 points higher than CVRM and 1.8 points higher than SGM. These data fully prove that the proposed model can realize the digitization of rural traditional culture more accurately and comprehensively. It not only injects new vitality into the theory of rural traditional culture protection, but also provides solid technical support for the development of rural cultural industry, and has certain value for inheriting national culture and enhancing rural cultural confidence. In the future, the data set can be further expanded, the model can be optimized, and its applicability in a wider range of scenarios can be improved.

Digitizing rural cultural heritage, particularly when it involves visual representations of human participants and traditional ceremonies, inevitably raises ethical and cultural questions. All human subjects visible in video materials were recorded with prior consent, either through public performance notices or direct written agreements during fieldwork. Additionally, the datasets were curated with guidance from local cultural offices and community representatives to ensure respectful and accurate representation. We acknowledge the importance of cultural ownership and emphasize that the digital artifacts generated by our model are intended for preservation and educational use, not for commercial redistribution. Future work may further explore participatory digitization models that include community stakeholders in data annotation, model validation, and interpretive framing. Although this study focuses on generating high-quality digital representations of rural traditional culture, we acknowledge that data storage and preservation are equally critical. As noted earlier, a large portion of digital archives remain on local servers, posing risks of loss due to hardware failure. While not the core of this paper, we suggest that future deployments integrate our model with cloud-based or decentralized storage systems (e.g., IPFS or institutional repositories) to enhance the resilience and accessibility of cultural data. We selected T3DRM, CVRM, and SGM as baselines due to their compatibility with our data and relevance to cultural digitization tasks. While advanced models like StyleGAN, Pix2PixHD, and VideoGAN offer stronger general performance, they require significant domain adaptation and resource overhead. We plan to include these models in future extended comparisons.

Funding

This work was supported by 2023(No.12) Project "The Integration of Ideological and Political Education in Primary, Secondary, and Higher Education in Guangdong Province".

References

[1] Wang TX, Ma ZQ, Zhang F, Yang L. Research on wickerwork patterns creative design and development based on style transfer technology. Applied Sciences-Basel. 2023; 13(3). DOI: 10.3390/app13031553

[2] Yang WY, Wen C, Lin BG. Research on the design and governance of new rural public environment based on regional culture. Journal of King Saud University Science. 2023; 35(2). DOI: 10.1016/j.jksus.2022.102425

[3] Zhang SL, Huang Z, Lang YL. Construction of evaluation index system for the integrated development of rural music culture and rural economy based on fuzzy neural network. Pakistan Journal of Agricultural Sciences. 2024; 61(4):1179-1190. DOI: 10.21162/pakjas/24.9536

[4] Zhao X. Study on the protection path of rural tourism characteristic villages under the background of regional ecological culture. Fresenius Environmental Bulletin. 2021; 30(3):3070-3076.

[5] Zeng SH, Pan ZL. An unsupervised font style transfer model based on generative adversarial networks. Multimedia Tools and Applications. 2022; 81(4):5305-5324. DOI: 10.1007/s11042-021-11777-0

[6] Yang LSQ, Li JH. Research on the creation of Chinese national cultural identity symbols based on visual images. Mathematical Problems in Engineering. 2022; 2022. DOI: 10.1155/2022/8848307

[7] Fanea-Ivanovici M, Pana MC. From culture to smart culture. How digital transformations enhance citizens well-being through better cultural accessibility and inclusion. IEEE Access. 2020; 8: 37988-8000. DOI: 10.1109/access.2020.2975542

[8] Della Lucia M, Dore G, Umar RM. Handling the open culture dilemma in museum management: an exploratory interdisciplinary study. Scientometrics. 2024; 129(12):7699-7733. DOI: 10.1007/s11192-024-05164-3

[9] Kumar P, Gupta V. Restoration of damaged artworks based on a generative adversarial network. Multimedia Tools and Applications. 2023; 82(26):40967-40985. DOI: 10.1007/s11042-023-15222-2

[10] Tang CC, Liu YR, Wan ZW, Liang WQ. Evaluation system and influencing paths for the integration of culture and tourism in traditional villages. Journal of Geographical Sciences. 2023; 33(12):2489-2510. DOI: 10.1007/s11442-023-2186-7

[11] Yang WL, Fan B, Tan JB, Lin J, Shao T. The spatial perception and spatial feature of rural cultural landscape in the context of rural tourism. Sustainability. 2022; 14(7). DOI: 10.3390/su14074370

[12] de Castro GZ, Guerra RR, Guimaraes FG. Automatic translation of sign language with multi-stream 3D CNN and generation of artificial depth maps. Expert Systems with Applications. 2023; 215. DOI: 10.1016/j.eswa.2022.119394

[13] Shi YT, Cheng QW, Wu YZ, Lin QH, Xu AX, Zheng QJ. Promoting or inhibiting? Digital inclusive finance and cultural consumption of rural residents. Sustainability. 2023; 15(3). DOI: 10.3390/su15032719

[14] Wang YJ, Chen YJ, Zhang W, Chao IC, Li H. The impact of rural tourism on rural culture evidence from China. Agriculture-Basel. 2024; 14(12). DOI: 10.3390/agriculture14122116

[15] Chen ZW, Lyu DS. Procedural generation of virtual pavilions via a deep convolutional generative adversarial network. Computer Animation and Virtual Worlds. 2022; 33(3-4). DOI: 10.1002/cav.2063

[16] Li LL, Chung WJ. Application of artificial intelligence in visual communication of green urban rural integration landscape design. Ecological Chemistry and Engineering S-chemia I Inzynieria Ekologiczna S. 2024; 31(4):583-597. DOI: 10.2478/eces-2024-0038

[17] Li X, Liu H. Analysis on the Three-dimensional intervention mode of public art in rural culture from the perspective of 3D video. Scientific Programming. 2022; 2022. DOI: 10.1155/2022/5090023

[18] Fu SB, Lv F. Farmer's Eco-culture for sustainable development: A qualitative approach of understanding the rural art design. Pakistan Journal of Agricultural Sciences. 2023; 60(1):235-244. DOI: 10.21162/pakjas/23.44

[19] Wu Q, Zhu BX, Yong BB, Wei YQ, Jiang XT, Zhou R, et al. ClothGAN: generation of fashionable Dunhuang clothes using generative adversarial networks. Connection Science. 2021; 33(2):341-358. DOI: 10.1080/09540091.2020.1822780

[20] Zhou YG, Hu X. Design and implementation of rural community elderly culture platform based on real-time social media data mining. Wireless Communications & Mobile Computing. 2021; 2021. DOI: 10.1155/2021/3927773

[21] Han, L. H. N., Hien, N. L. H., Huy, L. V., & Hieu, N. V. (2024). A deep learning model for multi-domain MRI synthesis using generative adversarial networks. Informatica, 35(2), 283–309. https://doi.org/10.15388/24-INFOR556

[22] Le Huy, H. N., Minh, H. H., Van, T. N., & Van, H. N. (2021). Keyphrase extraction model: A new design and application on tourism information. Informatica, 45(4). https://www.mii.lt/informatica/