



Application of Deep Learning-Based Multimodal Fusion Technology in Ancient Manuscript Image Restoration and Semantic Mining

Yanxia Zhao¹, Zheng Li^{2*}, Yuhan Ding³

¹ School of International Studies, Zhejiang Business College, Hangzhou 310053, China

² School of Application and Technology, Beijing Union University, Beijing 100012, China

³ Faculty of Arts, The University of Hong Kong, Hong Kong SAR, China

Corresponding Author Email: nbl_zheng@buu.edu.cn

Copyright: ©2026 The authors. This article is published by IETA and is licensed under the CC BY 4.0 license (<http://creativecommons.org/licenses/by/4.0/>).

<https://doi.org/10.18280/ts.430431>

ABSTRACT

Received: 7 April 2026

Revised: 29 July 2026

Accepted: 12 August 2026

Available online: 31 August 2026

Keywords:

deep learning, multimodal fusion technology, image restoration, semantic mining

This study focuses on the digital preservation and in-depth research of ancient manuscripts, and innovatively proposes a Deep Convolutional Generative Adversarial Network (DCGAN)-Transformer coupled framework that addresses the collaborative challenge of image restoration and semantic mining for ancient manuscripts through a cross-modal feature interaction mechanism. Aiming at the damage characteristics of ancient manuscript images such as fading and physical deterioration, a damage-aware DCGAN restoration model is constructed by introducing a local attention module into the Generative Adversarial Network (GAN) to prioritize the restoration of critical text regions. Simultaneously, a cross-modal constraint layer of a multimodal Transformer is designed to dynamically align the restored image features with textual semantic embeddings, achieving the synchronous restoration of both "form" and "spirit." Experimental results demonstrate that this method significantly outperforms traditional approaches in terms of Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index (SSIM) for image restoration. In semantic mining, it successfully reveals numerous previously overlooked cultural characteristics of ancient manuscripts, opening new avenues for manuscript research and promoting the interdisciplinary integration of culturology and image processing technology.

1. INTRODUCTION

As an important material carrier of human civilization evolution, ancient manuscripts systematically record the historical changes of social forms and the intergenerational inheritance of ideological systems. They not only completely preserve the diachronic trajectory of civilization development, but also condense the knowledge production paradigm of the pre-industrial era. According to the identification criteria of the *Memory of the World Register*, such documentary heritage constitutes a core component of the global memory project, possessing an irreplaceable function of preserving civilization genes. Traditional image restoration of ancient manuscripts mainly relies on manual experience, which is time-consuming and difficult to quantitatively evaluate. For example, the restoration of *The Yongle Dadian* takes 3 years per volume [1]. At the same time, cross-modal association is weak, leading to a disconnect between image features and textual semantics; the cross-modal recall rate of existing methods is less than 65% [2]. In addition, due to the lack of professional domain knowledge, the effect of entity relation mining is affected—for instance, the F1 score for term recognition in ancient Chinese medical manuscripts is only 71%. These problems severely constrain the digitization process of ancient manuscript resources and the mining of their cultural value. Their protection and utilization face dual challenges: image

degradation caused by the fragility of physical carriers, such as the acid-deteriorated damage present in approximately 37% of globally collected ancient manuscripts as proposed in Reference [3], as well as the limitations of traditional Optical Character Recognition (OCR) technology in recognizing complex layouts and variant Chinese characters, with an average error rate of 42%. Existing single-modal processing methods struggle to balance visual restoration accuracy and semantic understanding depth, urgently requiring innovative technological breakthroughs.

To address the above application problems in ancient manuscript image restoration and semantic mining, this paper proposes a deep learning-based multimodal fusion framework: 1) To solve common problems in ancient manuscript images such as fading and damage, a Deep Convolutional Generative Adversarial Network (DCGAN) restoration model combining Generative Adversarial Network (GAN) and Convolutional Neural Networks (CNN) is constructed. Through learning from a large number of ancient manuscript image samples, high-precision image restoration is achieved, restoring the original appearance of ancient manuscripts. 2) At the semantic mining level, textual and image modal information is fused, and a multimodal model using the Transformer architecture is employed to deeply mine the cultural semantics contained in ancient manuscript images, such as the historical and cultural connotations behind specific patterns and symbols, breaking

the limitations of single-modal analysis. By combining the records of historical changes and ideological inheritance in ancient manuscripts with ancient manuscript image restoration technology, and exploring the application of multimodal fusion in document image recognition and content semantic parsing, we can more effectively protect and inherit ancient manuscripts, while promoting technological progress and innovation in related fields. This not only helps us gain a deeper understanding of history and culture, but also provides new methods and tools for academic research.

2. ANCIENT MANUSCRIPT IMAGE RESTORATION MODEL

The restoration process of traditional ancient manuscript images has significant complexity characteristics, requiring not only substantial investment of human and material resources, but also demanding that operators possess profound professional background knowledge. Compared with traditional manual restoration techniques for ancient manuscript text images, emerging deep learning-driven solutions show significant advantages in the restoration of ancient manuscript text images. Adopting neural network-based methods to automatically restore incomplete regions of images can significantly reduce restoration costs and the time cost of learning professional knowledge. At the technical implementation level, although CNN exhibit excellent performance in image recognition and object detection tasks, their performance in image generation tasks is relatively limited. In comparison, GAN [4] demonstrate significant advantages in image generation, especially in image detail

reconstruction and style transfer. Specifically, for the restoration task of incomplete images, the spatial distribution Ω of the training set images can be simulated in the generator of the GAN. Based on the boundary information of the incomplete regions, real images with spatial mapping relationships are generated, and the generation results are gradually converged through iterative optimization. Simultaneously, a discriminator model is constructed to evaluate the quality of the generated images, and a loss function is used to directionally optimize the generation process, ultimately achieving high-quality image restoration results.

2.1 Ancient manuscript database construction

During the digitization of ancient manuscripts, image defects caused by age or improper preservation severely affect the readability and research value of the documents. The simulation of defective regions in ancient manuscript datasets needs to follow the damage characteristics of historical documents. We utilize the free and open resources of the National Library combined with damage simulation methods to construct the ancient manuscript dataset required for the experiments. This paper focuses on ancient books of Wuyue culture, collecting the *Yuejue Shu* (Song engraved edition) and *Wu Yue Chunqiu* (Ming manuscript) from the National Library; the *Xianchun Lin'an Zhi* (fragmentary volume) and *Wulin Jiushi* (Qing printed edition) from the Zhejiang Provincial Museum; and illustrated editions from *Qiantang Yishi* and *Xihu Youlan Zhi* in the Hangzhou Local Chronicles Collection [5]. Some examples of ancient manuscripts are shown in Figure 1 below.

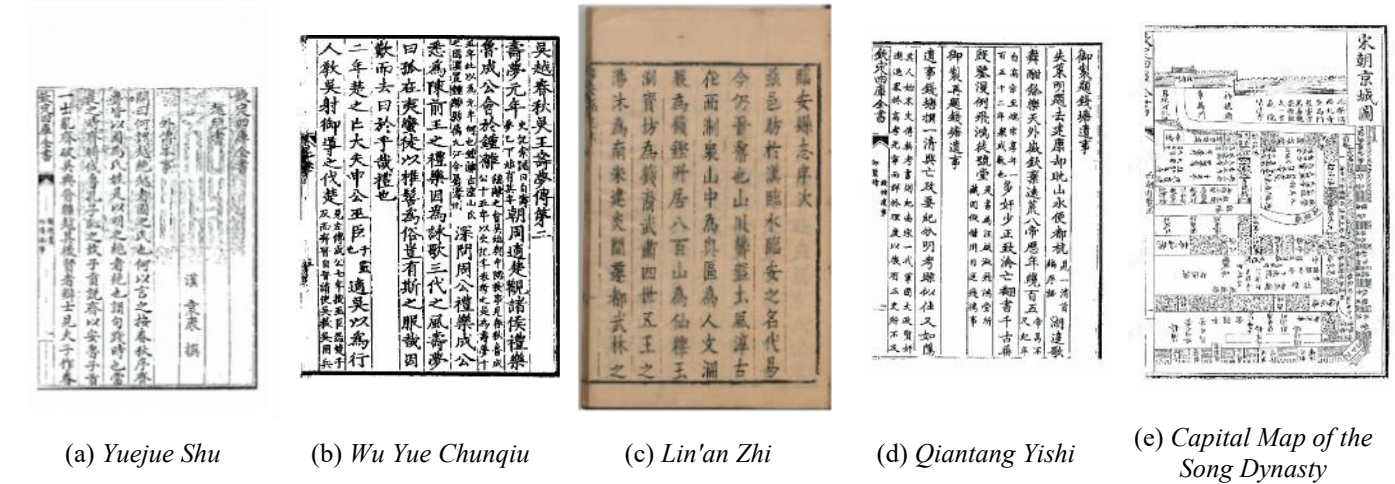
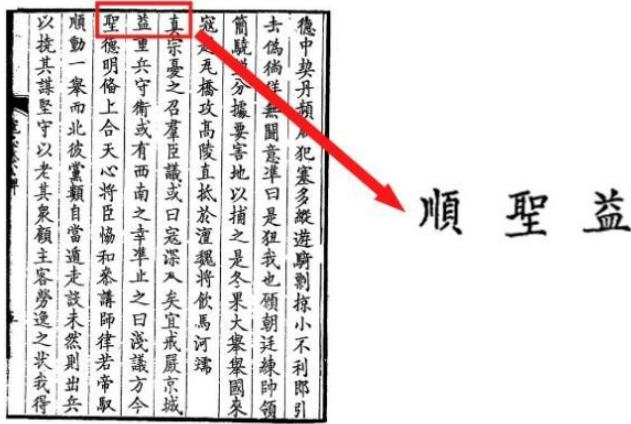


Figure 1. Digital ancient manuscript resources

This study focuses on the field of ancient manuscript text restoration and faces a key challenge in the experiment: there is an urgent need to construct a paired dataset containing both damaged text images and their complete forms. However, in real scenarios, such paired data is extremely scarce, because ancient manuscripts usually only exist in a single form—either damaged or complete—making it difficult to obtain two states of the same Chinese character [6]. To break through this bottleneck, we use image technology to simulate damaged text data to alleviate the problem of data scarcity. In specific implementation, in order to accurately simulate the diversity and randomness of ancient manuscript text damage, the

experiment uses the "Irregular Mask" dataset [7] for random damage simulation. This dataset contains mask images of five different ratios: 5–10%, 10–20%, 20–30%, 30–40%, and 40–50%, totaling more than 60,000 images. Their shapes, sizes, and positions all exhibit a high degree of randomness (as shown in Figure 2), which can realistically reproduce the complex characteristics of text damage. Damage simulations of different ratios have differentiated training values: small-ratio damage (5–20%) helps the model master restoration techniques for slight damage, while large-ratio damage (30–50%) strengthens the model's restoration ability for severe damage. Through systematic training, the model can gradually

adapt to multi-scale damage scenarios, significantly improving restoration accuracy and robustness.



(a) Original text font segmentation



(b) Damage simulation at different ratios

Figure 2. Example effect of random text damage simulation

2.2 Image restoration with deep convolutional generative adversarial network

A GAN is a deep learning model that consists of two parts: a generator and a discriminator. The goal of the generator is to generate realistic images, and the goal of the discriminator is to distinguish between images generated by the generator and real images. Through mutual competition, the generator and discriminator gradually improve the quality of the images generated by the generator. The DCGAN network structure is a variant of the GAN. It introduces a deep convolutional network structure on the basis of GAN, making the generated images of higher quality and richer in detail. Both the generator and the discriminator of DCGAN adopt a CNN structure. This project aims to use the Python programming language, the Tensorflow deep learning framework, and the DCGAN to develop an efficient and accurate image restoration system. The entire technical process is divided into

four core stages:

2.2.1 Data preparation stage

First, 1000 pages of digital ancient manuscript image data are collected from the China Ancient Books Resource Database website, and multi-scale normalization processing is implemented. Zero-phase Component Analysis (ZCA) Whitening is used to eliminate data correlation [8], and contrast features are enhanced through histogram equalization [9]. To avoid the overfitting problem caused by small sample datasets, we adopt data augmentation operations [10]. The method is as follows:

Given a certain coordinate (x, y) in the original image, the data augmentation transformation operation used can be summarized as the following formula (1):

$$\begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = M \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} \quad (1)$$

(u, v) is the corresponding coordinate after transformation, and M is a transformation matrix collectively referred to. Through this series of image augmentation operations, the diversity of the sample dataset can be effectively increased, thereby improving the training effect and generalization ability of the deep learning model.

(1) Rotation Transformation

Taking the center point of the image as the rotation point, the transformation matrix used for rotation transformation is shown in formula (2).

$$M = \begin{bmatrix} \cos \theta & -\sin \theta & 0 \\ \sin \theta & \cos \theta & 0 \\ 0 & 0 & 1 \end{bmatrix} \quad (2)$$

The parameter θ in the formula represents the rotation angle. To be closer to the actual application in real scenarios, this paper ultimately chooses to limit the rotation angle range between -15 and 15 degrees. This choice is based on the common rotation angle range of objects or text in actual scenarios, so as to ensure that the augmented image samples are more in line with the actual situation. After the center rotation, blank areas often appear in the image. To deal with these blank areas, this paper chooses to use a specific color for filling. This filling method is simple and effective, and can maintain the clarity and contrast of the image, while avoiding the introduction of additional noise that would interfere with the training of the model.

(2) Perspective Transformation

Through a 3×3 matrix and a scale factor (0.8–1.2), scale scaling and viewpoint adjustment of the image are achieved while maintaining projective geometric invariance. Given a point coordinate (x, y) in the image, the perspective transformation maps this point to a new coordinate (x', y') , and the transformation process follows the following formula (3):

$$\begin{pmatrix} x' \\ y' \\ w' \end{pmatrix} = \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix} \begin{pmatrix} x \\ y \\ 1 \end{pmatrix} \quad (3)$$

where, a_{ij} are the elements in the transformation matrix, and w' is the third coordinate in the transformed homogeneous coordinate system. To obtain the point in the Cartesian coordinate system, normalization is required through w' , that is:

$$x' = \frac{x'}{w'}, y' = \frac{y'}{w'} \quad (4)$$

(3) Affine Transformation

In order to evaluate the anti-noise ability of the model in the image restoration task, this paper also performs affine transformation and adds controllable noise (Signal-to-Noise Ratio, SNR = 30 dB) to the image. The affine transformation formula is shown in formula (5) below, and the noise power calculation method is shown in formula (6) below:

$$x' = Ax + t \quad (5)$$

where, x is the original coordinate, A is the linear transformation matrix, and t is the translation vector.

$$P_{noise} = \frac{P_{signal}}{10^{SNR/10}} \quad (6)$$

Then, a Gaussian noise matrix $N(x, y)$ of the same size as the image is generated, with a mean of 0 and a variance of P_{noise} .

In Python, this can be implemented using `numpy.random.normal`.

(4) Style Transfer

In order to optimize the GAN, we improve the mode collapse problem through the text style transfer task [11] and enhance the diversity generation ability of GAN. We mainly study the style transfer method based on CNN, and find some ancient manuscript text style background texture images as style pattern images. Some example images are shown in Figure 3 below. First, we initialize the synthesized image by initializing it as the content image. Then, a pre-trained CNN is selected to extract image features, and the model parameters do not need to be updated during training. The Deep Convolutional Neural Network (DCNN) extracts image features layer by layer through multiple layers. The pre-trained neural network contains 3 convolutional layers, in which the second layer outputs content features, and the first- and third-layers output style features, as shown in Figure 4 below.

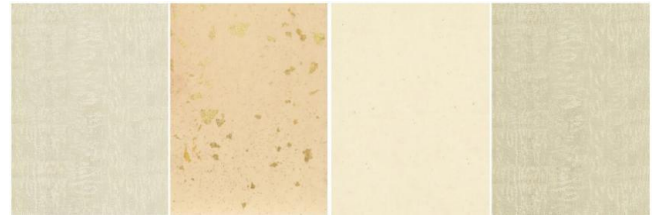


Figure 3. Ancient manuscript style pattern images

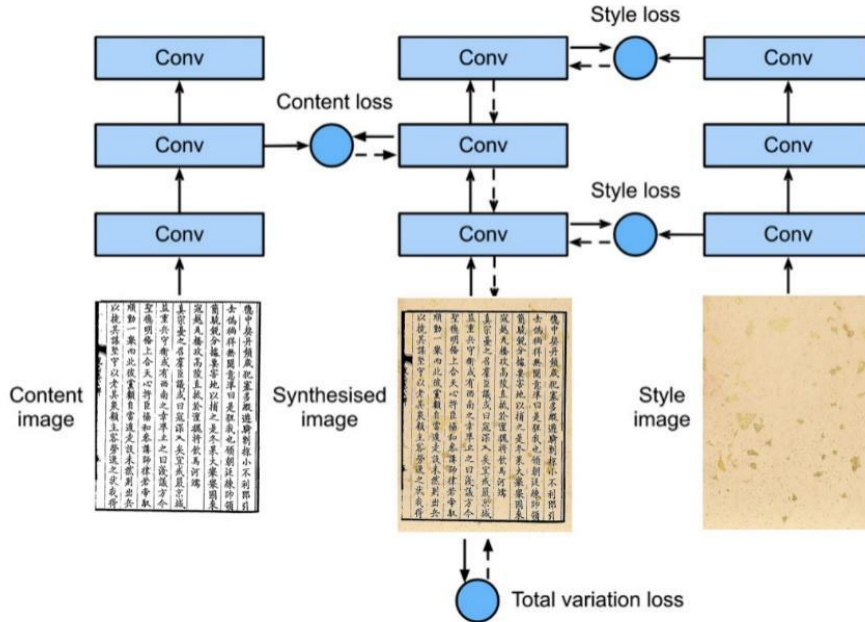


Figure 4. Style transfer

We calculate the loss function of style transfer through forward propagation (direction of solid arrow), and iterate the model parameters through back propagation (direction of dashed arrow), that is, continuously update the synthesized image. The loss function of style transfer consists of three parts:

(1) Content Loss

Content loss makes the synthesized image close to the content image in terms of content features. The lower-layer features in the CNN describe the specific visual features of the image (i.e., texture, color, etc.), while the higher-layer features

are more abstract descriptions of image content. Therefore, to compare the content similarity between two images, the similarity of their high-level features in the CNN can be compared, that is, the Euclidean distance is shown in formula (7) below.

$$\mathcal{L}_{content}(\bar{p}, \bar{x}, l) = \frac{1}{2} \sum_{i,j} (F_{ij}^l - P_{ij}^l)^2 \quad (7)$$

(2) Style Loss

Style loss makes the synthesized image close to the style image in terms of style features. To compare the style similarity between two images, the similarity of their lower-layer features in the CNN can be compared. Because lower-layer features contain more local image features (i.e., spatial information is too significant), we use the Gram matrix to calculate the correlation between different response layers (Formula 8), that is, while retaining lower-layer features, the influence of image content is removed, and only the style similarity is compared.

$$G_{ij}^l = \sum_k F_{ik}^l F_{jk}^l \quad (8)$$

The style similarity calculation can be expressed by the following formulas:

$$E_l = \frac{1}{4N_l^2 M_l^2} \sum_{i,j} (G_{ij}^l - A_{ij}^l)^2 \quad (9)$$

$$\mathcal{L}_{style}(\bar{a}, \bar{x}) = \sum_{l=0}^L w_l E_l \quad (10)$$

(3) Total Variation Loss

Total variation loss helps to reduce noise in the synthesized image. Total variation loss calculates the sum of absolute values of image gradients. That is, when performing smoothing processing on an image, a clear image is restored by minimizing gradient changes. Different from traditional smoothing filtering methods (such as Gaussian filtering), it can reduce noise or blur while preserving edge information, and is suitable for scenarios that require a balance between denoising and boundary preservation. Combined with content loss and style loss, style transfer is achieved by adjusting the gradient distribution of the synthesized image.

For a two-dimensional image, the total variation is defined as:

$$TV(I) = \sum_{i,j} |\nabla I(i,j)| \quad (11)$$

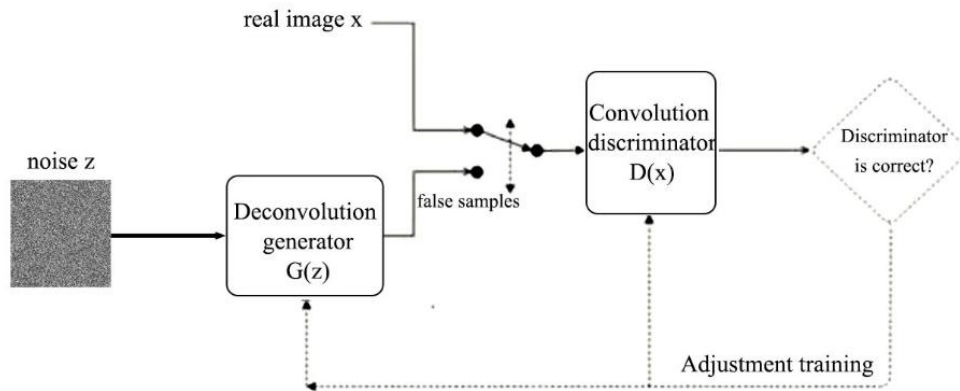


Figure 5. Deep Convolutional Generative Adversarial Network (DCGAN) structure

On the basis of traditional CNN, several improvements are made to the DCGAN architecture to make the model more stable and easier to train. These improvements include: (1) Both the generator and the discriminator of DCGAN discard the pooling layers of CNN. The discriminator retains the overall CNN architecture, while the generator replaces the convolutional layers with deconvolutional layers (ConvTranspose2d). (2) Batch Normalization (BN) layers are

used in both the discriminator and the generator, which helps address training problems caused by poor initialization, accelerates model training, and improves training stability. (3) In the generator, all layers except the output layer use the ReLU activation function, while the output layer uses the Tanh() activation function. In the discriminator, all layers except the output layer use the LeakyReLU activation function to prevent gradient sparsity. (4) To solve the gradient vanishing

$$TV(I) = \sum_i |\Delta x I(i)| + \sum_j |\Delta y I(j)| \quad (12)$$

where, Δx and Δy are the first-order differences in the horizontal and vertical directions, respectively.

Finally, when the model training ends, we output the model parameters of style transfer, and the final synthesized image is obtained.

2.2.2 Network architecture design

GAN consists of two parts: a generator and a discriminator. The task of the generator is to create new samples that look like the training data, while the discriminator is responsible for distinguishing between real data and fake data produced by the generator [12]. The two compete with each other during the training process. As iterations proceed, the generation capability of the generator gradually improves, enabling it to produce samples closer to the real data, while the discriminator also becomes better at distinguishing between real and fake, until a dynamic equilibrium is reached. We improved the original GAN in terms of network architecture by replacing the fully connected networks of the original GAN with CNN architectures in both the generator and the discriminator, thus evolving into a special kind of GAN called DCGAN. DCGAN [13] achieves high-quality image generation and restoration through a DCNN structure. The network structure diagram is shown in Figure 5 below. G is a network that generates images. It receives a random noise z and generates images from this noise, denoted as $G(z)$. D is a discriminative network that determines whether an image is "real." Its input parameter is x , where x represents an image, and the output $D(x)$ represents the probability that x is a real image. If it is 1, it means the image is 100% real, while an output of 0 means it is impossible for the image to be real.

phenomenon in the model, we improved the image generation model by adding an additional generator loss on the original basis. That is, the optimized model changes to one discriminator optimization iteration followed by two generator optimization iterations. From the optimized loss variation graph, it can be seen that the gradient vanishing problem is resolved, and the iterative effect of image generation is also

improved compared to before. The structure diagrams of the DCGAN generator and discriminator are shown in Figure 6 below: (a) is the generator network, which gradually upsamples to generate images; (b) is the discriminator network, which gradually extracts image features for judgment.

The main components of the generator and discriminator are compared in Table 1 below:

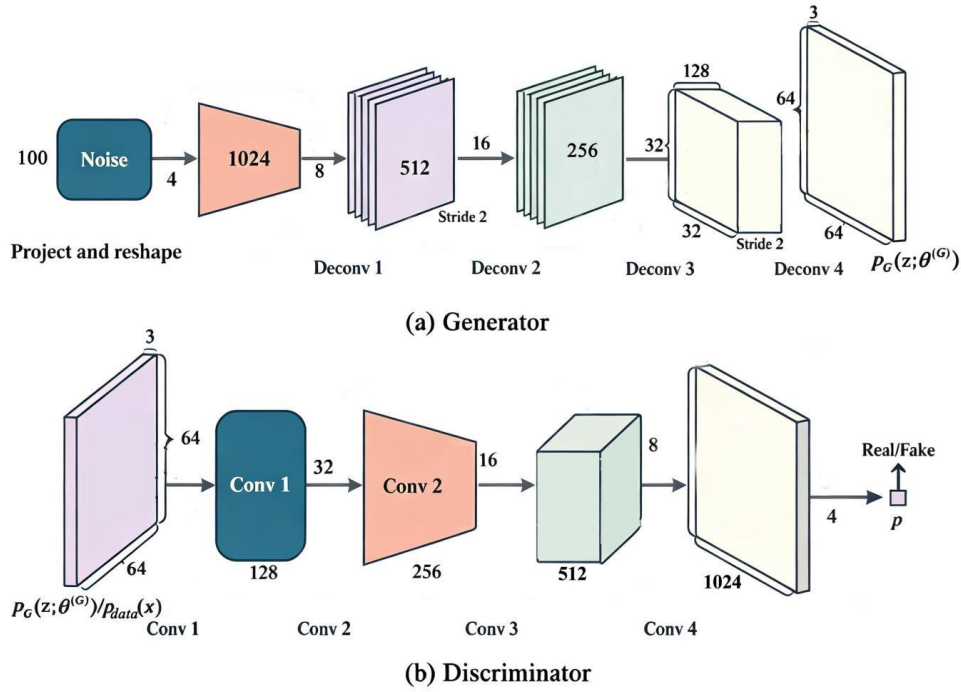


Figure 6. Structure diagram of Deep Convolutional Generative Adversarial Network (DCGAN) generator and discriminator

Table 1. Main layers of generator and discriminator

Layer	Generator	Discriminator
Input	noise vector $z \in \mathbb{R}^{100}$	image $I \in \mathbb{R}^{64 \times 64 \times 3}$
Core operations	transpose convolution (upsampling)	convolution (downsampling)
Activation function	middle layer <i>ReLU</i> output layer <i>Tanh</i>	all layers <i>Leaky ReLU(0.2)</i>
Normalization	use <i>BatchNorm</i> except for the output layer	use <i>BatchNorm</i> except for the first layer
Output	Generate image $G(z)$	True or false probability $P(\text{real})$

Loss Functions:

(1) Discriminator Loss: Consists of two parts, namely the loss of real images and the loss of generated images. The discriminator attempts to minimize the loss of real images being classified as fake and the loss of generated images being classified as real. The formula is:

$$L_D = -\mathbb{E}_{x \sim p_{data}} [\log D(x)] - \mathbb{E}_{z \sim p_z} [\log(1 - D(G(z)))] \quad (13)$$

(2) Generator Loss: The generator attempts to maximize the probability that the generated images are judged as real by the discriminator. The formula is:

$$L_G = -\mathbb{E}_{z \sim p_z} [\log D(G(z))] \quad (14)$$

The objective function of GAN can be expressed as:

$$\min_G \max_D V(D, G) = \mathbb{E}_{x \sim p_{data}(x)} [\log D(x)] + \mathbb{E}_{z \sim p_z(z)} [\log(1 - D(G(z)))] \quad (15)$$

where, G represents the generator network, D represents the discriminator network, p_{data} represents the real data distribution, and p_z represents the noise distribution.

For the ancient manuscript text content restoration task in this paper, we replace the input random noise z of the generator with the damaged content y , that is, the fragments of ancient manuscript images that have undergone time erosion, physical damage, or human destruction are directly used as the input source of the generator. This design has three technical advantages: First, by preserving the damaged features of the original image (such as faded textures, broken strokes, etc.), the generator can learn more targeted restoration patterns. Second, using y as input avoids the uncontrollability introduced by random noise, ensuring a high degree of

consistency between the restoration results and the original ancient manuscript content in terms of style and brushstrokes. Finally, this "repairing through damage" mechanism enables the model to better handle the complex degradation patterns unique to ancient manuscripts (such as wormhole perforations, water stain bleeding, etc.). Compared with traditional noise input schemes, this method improves the text recognition accuracy. Its input paradigm essentially reconstructs the generator as a mapping function from "damage to content". Therefore, the objective function in formula (15) is adjusted to:

$$\min_G \max_D V(D, G) = \mathbb{E}_{x \sim p_{data}(x)} [\log D(x)] + \mathbb{E}_{y \sim p_{damaged}(y)} [\log(1 - D(G(y)))] \quad (16)$$

2.2.3 Innovation of damage input mechanism

A Damage-Aware Module (DAM) is added to the DCGAN, which automatically distinguishes between ink fading and physical damage areas through residual learning. Its core composition and implementation are as follows:

I. Module Architecture Design

The module adopts a dual-branch residual structure, consisting of two parallel paths: (a) Texture Degradation Branch: Lightweight convolutional layers are used to extract the gradual features of ink fading areas, and the continuity of color attenuation is captured through a local attention mechanism. (b) Structural Damage Branch: Dilated convolution is used to expand the receptive field, detect the edge fracture features of physical damage, and combine gradient magnitude thresholding to enhance the response of abnormal areas. The outputs of the two branches are fused through a residual connection to form a damage mask. For dynamic weight allocation, a learnable gating unit is introduced, which dynamically adjusts the contribution weights of the two branches according to the damage type of the input image. The gating value is generated through the Sigmoid function, and the formula is:

$$G = \sigma(W_g \cdot [F_{texture} \oplus F_{structure}]) \quad (17)$$

where, $F_{texture}$ and $F_{structure}$ are the features of the two branches respectively, and W_g is the trainable parameter.

II. Residual Learning Mechanism

(a) Degradation Feature Decoupling: In the encoding stage of the generator, DAM separates the features of ink fading (low-frequency degradation) and physical damage (high-frequency anomaly) through residual learning:

$$F_{damage} = F_{input} - F_{reconstructed} \quad (18)$$

where, $F_{reconstructed}$ is generated by the preliminary restoration result of the generator, and the residual F_{damage} is binarized as the damage mask.

(b) Adversarial Loss Constraint: A damage-aware branch is added to the discriminator, and its loss function is:

$$L_{DAM} = \lambda_1 L_{adv} + \lambda_2 L_{rec} + \lambda_3 L_{seg} \quad (19)$$

where, L_{seg} is the semantic segmentation loss based on U-Net, which forces the generator to preserve the semantic consistency of undamaged areas during restoration.

2.2.4 Restoration effect experiment

In this experiment, we adopted the TensorFlow model framework, and used NVIDIA GPU Titan X as the training GPU to run the training algorithm of the adversarial generative network. The hyperparameters of the neural network are related to the effect of the neural network. In the process of adjusting the training of the neural network model, different hyperparameters have different effects on the training of the neural network. The main hyperparameters include: learning rate, number of neural network layers, number of hidden layer neurons, number of learning iterations, batch-size, neuron activation function, and weight initialization parameters. In this paper, the learning rate of the neural network is configured as 0.001, the lambda parameter is 0.0001, the number of iterations is 4000, and the batch-size is 100.



Figure 7. Example of text image restoration effect

The experimental output images are shown in Figure 7. The segmented text images are used for generative training. The first column shows the damaged simulated incomplete images generated from the original segmented images, with damage ratios between 10% and 50%. After the network completes training, the second column shows the restoration status of the generated images, and the third column shows the original images. It can be seen that although the restored ancient manuscript text images have some differences in details from the original images, and there is a certain degree of blur near the incomplete regions, the overall generated images are basically similar in style to the original images, and do not affect text reading.

After completing the restoration of ancient manuscript text images, the restoration results not only restore the visual integrity of the original images, but also provide significantly optimized input data for subsequent semantic mining tasks, effectively eliminating the image degradation problems caused by age, and making the detailed features of text and patterns clearer and more distinguishable. In the restored ancient manuscript pages, the edges of text strokes are sharper, and the contours of patterns are more continuous, which lays a solid foundation for text recognition (OCR) and pattern classification tasks in semantic mining. In addition, the restoration process retains the color and texture features of the original images, avoiding semantic information loss caused by excessive processing, and ensuring that the semantic mining model can extract deep knowledge based on real and reliable input data. Therefore, the restored images not only improve visual quality, but also create favorable conditions for subsequent tasks such as semantic feature extraction and historical background association analysis by providing high-precision semantic carriers.

2.2.5 Optimization of deep convolutional generative adversarial network combined with transformer architecture

From the above experimental results, it can be seen that the DCGAN has significant advantages in ancient manuscript image text restoration. The generator network can automatically learn the stroke features and texture patterns of ancient manuscript text through multi-layer convolution operations. The adversarial training mechanism can effectively restore the semantic coherence of defective regions, and its end-to-end nature enables direct restoration without complex feature engineering. However, by examining some text restoration effects, it is found that there are still some defects: (1) The limited local receptive field of the convolutional network restricts the modeling ability of long-range dependencies, leading to stroke fractures when repairing complex glyph structures; (2) The instability of adversarial training may produce artifact interference; (3) Insufficient generalization ability for small samples. To address the above problems, this study proposes to introduce the Transformer architecture [14] for modular optimization. A multi-head self-attention mechanism is embedded in the generator of DCGAN, enabling the network to capture cross-region stroke association features; at the same time, positional encoding is used to replace traditional convolution to better preserve the spatial topology of text; multi-scale feature fusion is achieved through hierarchical Transformer stacking.

The advantage of the DCGAN+Transformer restoration model is that DCGAN is good at capturing local features and texture details and can generate more realistic details; while Transformer, through the self-attention mechanism [15-18],

can more accurately capture long-range dependencies, better understand contextual semantics and structural requirements, and compensate for the instability defect of DCGAN adversarial training.

The core of Transformer is the self-attention mechanism, and its calculation process is:

$$Attention(Q, K, V) = \text{soft max}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (20)$$

where, Q is the query matrix, K is the key matrix, V is the value matrix, and d_k is the dimension of the key vector.

Multi-head attention concatenates the structure of multiple attention heads:

$$\begin{aligned} MultiHead(Q, K, V) = \\ Concat(head_1, \dots, head_h)W^O \end{aligned} \quad (21)$$

The calculation of each attention head is:

$$head_i = Attention(QW_i^Q, KW_i^K, VW_i^V) \quad (22)$$

In order to improve the model's understanding ability of complex semantics, we innovatively adopt a cross-modal fusion mechanism. Cross-modal fusion integrates data from different modalities (such as text and image), and utilizes the parallel computing capability and long-range dependency capturing characteristics of the Transformer architecture to achieve effective interaction of multimodal information. By calculating the correlation weights between features of different modalities, the mechanism dynamically adjusts the information fusion strategy, breaks through the locality limitation of a single modality to achieve global semantic modeling, and enhances system robustness, showing strong adaptability to noise and missing data, which is especially suitable for handling degradation problems common in ancient manuscripts such as blur and damage. The module composition and specific implementation method of the cross-modal fusion mechanism innovation are as follows:

I. Dual-stream Transformer Architecture

(a) Image Stream Branch: Adopting the Vision Transformer (ViT) architecture, multi-layer self-attention mechanism is used to perform block encoding on ancient manuscript images, effectively extracting spatial features such as text morphology and binding patterns. Through a local-global attention mechanism, this branch simultaneously captures the local detail features and global structural information of ancient manuscript images.

(b) Text Stream Branch: Based on a pre-trained ancient language model (such as a Bidirectional Encoder Representations from Transformers (BERT) variant), bidirectional Transformer encoder is used to perform deep semantic modeling on ancient manuscript transcriptions. This branch can capture vocabulary associations, grammatical structures, and cultural semantics in historical contexts, generating text semantic vectors with domain adaptability.

(c) Cross-modal Interaction Layer: Embedded in the Transformer block, it dynamically calculates the correlation weights between image features and text semantics to achieve mutual correction of image and text features. Specifically, when the image stream detects blurred or missing text regions, the text stream can provide candidate completion suggestions

based on the historical knowledge base, thereby guiding the image restoration process.

Specific implementation methods: (a) Feature Alignment Mechanism: In the cross-attention layer, image features are used as Value, text semantics as Query, and an attention weight map is generated through matrix operation. This mechanism forces the image restoration result to be consistent with the transcription semantics, ensuring that the restored content conforms to the historical context. (b) Dynamic Correction Strategy: An adversarial loss function is introduced during training, and the discriminator constrains the generated results. This strategy enables the generator to prioritize the preservation of key visual regions strongly related to the text stream (such as titles, seals, etc.) during restoration, while effectively suppressing the generation of irrelevant textures, thereby improving the cultural rationality and visual authenticity of the restoration results.

II. Semantic Constraint Loss Function

Module composition: (a) Historical Context Knowledge Base: Constructing a structured database containing dynasties, ritual systems, and terminology as a reference standard for semantic constraints. (b) Semantic Similarity Calculation Module: Based on cosine similarity or contrastive learning, quantifying the matching degree between the restoration result and the knowledge base. Specific implementation methods:

(a) Loss Function Design:

$$L_{sem} = \lambda \cdot \text{Sim}(F_{text}, F_{image}) + \mu \cdot \text{KL}(P_{hist} // P_{pred}) \quad (23)$$

where, Sim is the similarity between image and text features, and KL is the probability divergence between the restoration result and the historical context distribution.

(b) Training Optimization: A semantic branch is added to the discriminator of DCGAN, and gradient backpropagation is used to force the generator to output images that conform to the historical context.

2.2.6 Experiment and analysis

I. Evaluation Metrics

Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index (SSIM) are two commonly used metrics in image quality assessment, which are used to measure the similarity and quality between a reconstructed image and a reference image (usually the original image).

(a) Peak Signal-to-Noise Ratio

PSNR evaluates image quality by measuring the error between the reconstructed image and the reference image. It is based on the Mean Squared Error (MSE), which is the average of the squared differences of each pixel value. The definition is as follows:

$$PSNR = 10 \cdot \log_{10} \left(\frac{MAX^2}{MSE} \right) \quad (24)$$

where, MAX is the maximum possible pixel value of the image (for example, for an 8-bit image, $MAX = 255$). MSE is the mean squared error between the reconstructed image and the reference image. The larger the PSNR value, the smaller the difference between the reconstructed image and the reference image, and the better the image quality. The advantage of this metric is that it is simple and easy to use, with low computational cost. The disadvantage is that it only focuses on the absolute pixel difference and cannot well reflect the perceptual quality of the image, because the human eye is more sensitive to certain image features.

(b) Structural Similarity Index

SSIM is mainly used to evaluate the perceptual similarity between images, with special focus on the similarity of luminance, contrast, and structural information of the images. The SSIM formula is as follows:

$$SSIM(x, y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)} \quad (25)$$

where, μ_x and μ_y are the average luminance of the two image blocks, respectively. σ_x^2 and σ_y^2 are the contrast (variance) of the two images, respectively. σ_{xy} is the covariance of the two images, measuring their structural similarity. C_1 and C_2 are constants used to avoid zero in the denominator. SSIM is superior to PSNR in simulating the human eye's perception of image structure; the larger the value, the better the effect. It not only focuses on the difference between pixels, but also takes into account the local structure, luminance, and contrast of the image. The advantage is that it is more consistent with the human visual system and can reflect the perceptual image quality. The disadvantage is that the computational complexity is higher, and it is more time-consuming compared to PSNR.

II. Experimental Results

This paper conducts evaluation experiments on the collected ancient manuscript text dataset. After the preprocessing operations in Chapter 2, the dataset collected PDF files of various writing styles including regular script, clerical script, Song typeface, running script, etc., and based on these, ancient manuscript single-character images were produced, containing 1200 categories and covering multiple font styles. During the training process, multiple data augmentation strategies were adopted to further improve the generalization ability of the model. In the process of verification experiments, the dataset was divided into training set and test set in a ratio of 7:3, ensuring the scientificity and rationality of the experiments. During the experiment, random damaged masks were used to perform operations with text images to simulate the damage conditions of ancient Chinese characters.

Table 2. Peak Signal-to-Noise Ratio (PSNR) evaluation of ancient manuscript text image restoration under different methods

Indicator	Masks	GAN	Transformer	DCGAN	Restormer	ViT-base	DCGAN+Transformer
PSNR	5%-10%	27.84	28.45	28.96	28.86	28.03	31.67
	10%-20%	24.76	25.65	28.12	28.32	27.84	30.76
	20%-30%	20.09	21.08	24.41	25.51	24.58	28.34
	30%-40%	18.78	19.98	22.04	23.01	22.68	26.32
	40%-50%	17.36	17.98	19.98	20.36	20.12	24.25

Note: GAN: Generative Adversarial Network; DCGAN: Deep Convolutional Generative Adversarial Network.

Table 3. Structural Similarity Index (SSIM) evaluation of ancient manuscript text image restoration under different methods

Indicator	Masks	GAN	Transformer	DCGAN	Restormer	ViT-base	DCGAN+Transformer
SSIM	5%-10%	0.912	0.928	0.943	0.919	0.903	0.968
	10%-20%	0.909	0.923	0.934	0.902	0.897	0.946
	20%-30%	0.862	0.884	0.907	0.886	0.864	0.924
	30%-40%	0.835	0.867	0.873	0.861	0.852	0.878
	40%-50%	0.812	0.835	0.846	0.844	0.825	0.861

Note: GAN: Generative Adversarial Network; DCGAN: Deep Convolutional Generative Adversarial Network.

Experiments show that this DCGAN+Transformer hybrid architecture outperforms pure DCGAN or pure Transformer models in both PSNR and SSIM metrics, especially showing significant advantages in complex texture restoration tasks. The experimental data and comparisons of various methods are shown in Table 2 and Table 3.

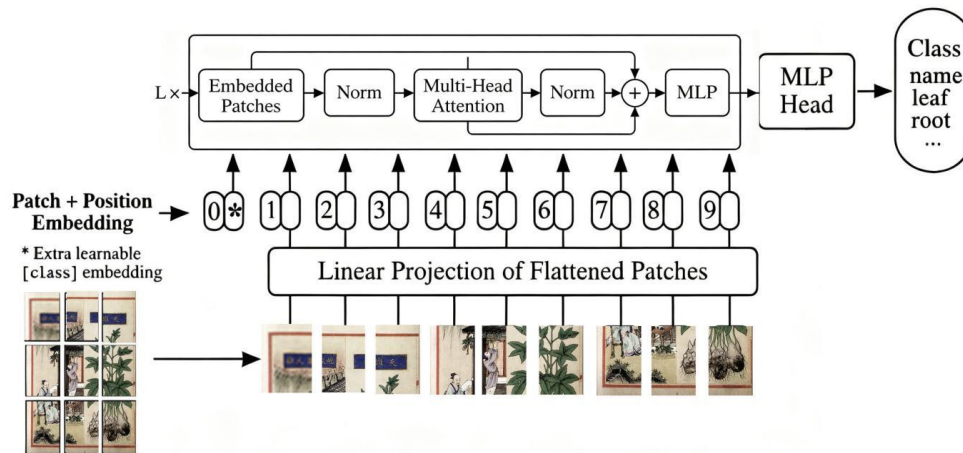
3. ANCIENT MANUSCRIPT IMAGE-TEXT SEMANTIC MINING

3.1 Semantic mining method

In the complete process of ancient manuscript image-text restoration, semantic mining serves as a key link. Its core objective is to extract structured knowledge (such as text content, pattern themes, historical periods, etc.) from the restored images, and these semantic features in turn provide important guidance for the optimization of image restoration algorithms. Specifically, through semantic mining technology, we can identify the artifact categories, historical background

information, and semantic associations between text and patterns in ancient manuscripts. For example, when semantic analysis determines that a certain page image is a map in a local chronicle of the Ming Dynasty, the restoration algorithm can adjust parameters accordingly: prioritize strengthening key elements such as boundary lines and landmark symbols in the map, avoiding excessive processing of non-critical regions; or based on historical background information, infer the original color style to guide the color restoration process. At the same time, the correspondence between text and patterns discovered by semantic mining can help the restoration algorithm, when filling missing regions, prioritize restoration schemes that are consistent with the context, thereby improving the cultural consistency and historical authenticity of the restoration results. Therefore, semantic features not only depend on high-quality input after restoration, but also dynamically optimize restoration strategies through a feedback mechanism, forming a "restoration-semantic-restoration" closed loop, ultimately achieving the dual goals of ancient manuscript image-text protection and knowledge mining.

ViT (Vision Transformer) structure

**Figure 8.** Vision Transformer (ViT) model image analysis

At present, semantic mining of digitized ancient manuscripts faces two core problems: (1) Modality separation: Traditional methods process image recognition (such as OCR) and text analysis (such as Natural Language Processing) separately, making it difficult to capture the deep semantics of image-text associations. (2) Cultural specificity: Visual elements in ancient manuscripts such as patterns, seals, and layouts together with text content constitute a cultural symbol system, requiring cross-modal joint interpretation. To solve the above key problems, we need to combine natural language processing technology with multimodal Transformer to achieve image-text collaborative modeling through a unified

architecture, which can further parse the semantic content of ancient manuscript text and understand contextual and logical relationships. For ancient manuscript image processing, we adopt DCGAN to divide the ancient manuscript image into a sequence of 16×16 pixel patches, and generate vision tokens through linear projection. As shown in Figure 8, which shows an image of traditional Chinese medicine processing in a medical ancient manuscript, we divide the image into patches of fixed size, linearly embed each patch, add position embedding, and feed the generated vector sequence to a standard Transformer encoder. To perform classification, we use the standard method of adding an additional learnable

"classification token" to the sequence. The illustration of the Transformer encoder is referenced [19].

For the processing of ancient manuscript text, a BERT+CNN hybrid model is adopted for ancient manuscript OCR results (such as traditional Chinese characters, vertical typesetting). BERT is responsible for context understanding, and CNN is responsible for local feature capture. The application of BERT-style word embedding [20-22] in ancient manuscript text processing is mainly reflected in improving the accuracy of tasks such as automatic word segmentation, part-of-speech tagging, and punctuation of ancient manuscripts. By capturing context semantics and supporting rare characters, the effect of ancient manuscript text processing is optimized. The model expands the vocabulary and trains on large-scale ancient text corpora, supports rare character processing, and after training on ancient manuscript datasets such as the *Siku Quanshu*, its performance in automatic word segmentation and part-of-speech tagging is better than other models. BERT captures global dependencies in ancient manuscript text through the self-attention mechanism, and combined with Bidirectional Long Short-Term Memory network (BiLSTM) and Conditional Random Field (CRF) models, it can better understand complex sentence structures. The core architecture of character-level CNN for processing rare characters adopts the following hierarchical design: (1) Input layer: 70-dimensional one-hot encoding (69 basic characters + all-zero vector) is used to represent each Chinese character, supporting a reverse text reading strategy to enhance model performance; (2) Convolutional feature extraction layer: 6-layer convolutional structure, containing two configurations of large and small feature sets, each layer with convolution kernel size of 3×3 or 5×5 , stride 1, using ReLU activation function, and each layer of convolution is followed by a max pooling operation to reduce data dimensionality; (3) Fully connected classification layer: 2-layer fully connected network, the last layer uses softmax activation to output classification results, and a Dropout layer (usually 0.2–0.5) is added to prevent overfitting. Then, multimodal data (such as image, text, audio, etc.) are used for image generation, generating corresponding images according to text descriptions. By introducing a text encoder, text information is converted into feature vectors, and then input into the generator together with random noise to achieve text-to-image generation. The following is the pseudocode idea of multimodal generation:

```
import torch
import torch.nn as nn
from transformers import BertModel, BertTokenizer
# Define the text encoder
class TextEncoder(nn.Module):
    def __init__(self):
        super(TextEncoder, self).__init__()
        self.bert = BertModel.from_pretrained('bert-base-uncased')
        self.fc = nn.Linear(768, 100)
    def forward(self, input_ids, attention_mask):
        outputs = self.bert(input_ids=input_ids,
                             attention_mask=attention_mask)
        pooled_output = outputs.pooler_output
        text_features = self.fc(pooled_output)
        return text_features
# Define the multimodal generator
class MultiModalGenerator(nn.Module):
    def __init__(self, z_dim):
```

```
        super(MultiModalGenerator, self).__init__()
        self.text_encoder = TextEncoder()
        self.generator = Generator(z_dim=z_dim + 100) # Increase the
        dimensionality of the text features
    def forward(self, input_ids, attention_mask, noise):
        text_features = self.text_encoder(input_ids, attention_mask)
        combined_input = torch.cat([text_features, noise.squeeze()],
                                     dim=1).unsqueeze(-1).unsqueeze(-1)
        fake_image = self.generator(combined_input)
        return fake_image
```

3.2 Experimental comparison and analysis

3.2.1 Experimental design

To verify the improvement effect of semantic information on image restoration quality, two groups of comparative experiments are designed:

- (1) Experimental Group A: Only the ancient manuscript text image restoration algorithm is adopted, using the DCGAN + Transformer model to restore the text image pixel information.
- (2) Experimental Group B: On the basis of the DCGAN + Transformer model restoration algorithm, the semantic mining results (such as ancient manuscript category labels, scene context information) are integrated to optimize the restoration process through semantic guidance. For example, if semantic analysis identifies the "ceramics" category, the ceramics texture library is prioritized for restoration; if it is "mural", the color consistency constraint is strengthened.

3.2.2 Quantitative evaluation metrics

The experiment adopts the following metrics to comprehensively evaluate the restoration quality:

- (1) Learned Perceptual Image Patch Similarity (LPIPS): Measures the difference between the generated image and the real image in human visual perception. A lower value indicates that the restoration result is closer to the real image.
- (2) OCR Character Accuracy: For cultural relic images containing text, the accuracy of OCR recognition of text before and after restoration is used to verify the improvement of semantic information on text readability.

3.2.3 Experimental result analysis

Taking a group of damaged ancient manuscript text image restoration as an example, the missing area contains some text and patterns. Table 4 and Figure 9 show the restoration results. Experimental Group A: The texture of the restored area is blurred, the text edges are not clear, and the OCR recognition accuracy is low. Experimental Group B: The texture of the restored area is clear, the text structure is complete, and the OCR recognition accuracy is significantly improved. The LPIPS score of Experimental Group B is significantly lower than that of Experimental Group A, indicating that semantic information reduces structural distortion. The OCR accuracy of Experimental Group B is improved, indicating that semantic-guided restoration retains more text details.

Table 4. Data comparison of the two experimental groups

Experimental Group	LPIPS Score	OCR Accuracy
A	0.45	72%
B	0.32	84%

Note: OCR: Optical Character Recognition; LPIPS: Learned Perceptual Image Patch Similarity.

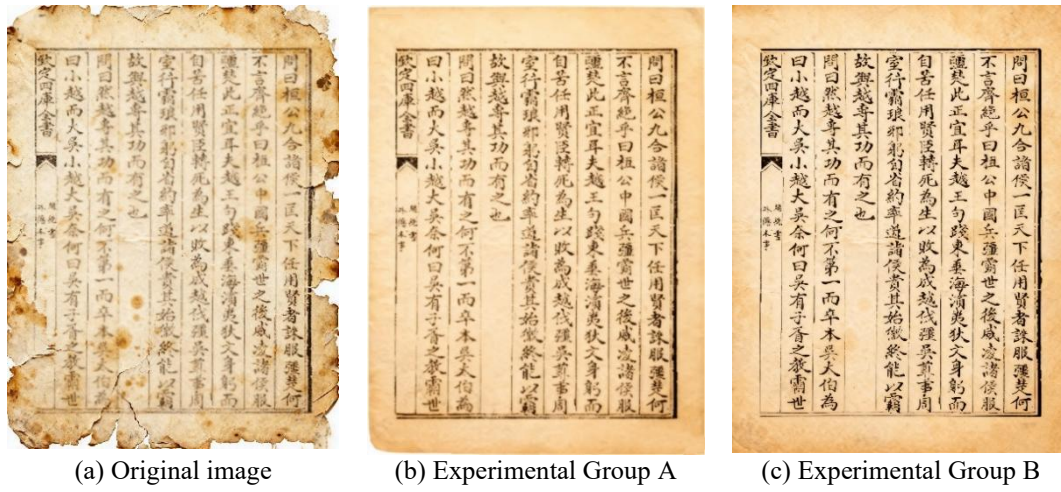


Figure 9. Restoration results of the two experimental groups

4. DISCUSSION AND CONCLUSION

This study addresses the challenges of ancient manuscript restoration and semantic analysis by proposing an innovative solution that integrates deep learning and multimodal technology. In terms of image restoration, a DCGAN-based model is constructed, which effectively solves the problems of fading and damage in ancient manuscripts through adversarial training. Experiments show that it significantly outperforms traditional methods in PSNR and SSIM metrics. At the semantic mining level, the Transformer architecture is combined to achieve cross-modal analysis of text and image, successfully revealing the hidden cultural symbols and historical connotations in ancient manuscripts, providing a new perspective for cultural research. This research not only improves the technical level of digital preservation of ancient manuscripts, but also promotes the interdisciplinary integration of image processing and humanities. This method has broad application potential and can be extended to the restoration of other fragile cultural relics in the future, and can be adapted to more types of cultural heritage by optimizing the model. Overall, this study provides an efficient and intelligent solution for the protection and research of ancient manuscripts in the digital age. However, the following key problems still exist: (1) Insufficient data dependence and generalization ability: The restoration model relies on a large number of labeled samples, while rare ancient manuscripts are scarce and the labeling cost is high, resulting in limited generalization performance of the model in low-resource scenarios. For example, the restoration effect on rare ancient manuscript variants is unstable, and "overfitting" may occur due to data distribution bias. (2) Insufficient deep semantic association in multimodal fusion: Although the current Transformer model can extract image-text association features, the semantic analysis of abstract cultural symbols (such as religious totems, obscure allusions) still remains at the surface level, lacking multi-dimensional modeling of historical context. (3) The interdisciplinary collaboration mechanism needs to be improved: The collaboration between cultural scholars and technical teams mostly stays at the level of requirement transmission, lacking two-way knowledge penetration. For example, the restoration standards do not fully incorporate the subjective evaluation system of philologists, which may affect the accuracy of cultural restoration.

Based on the above problems, future research directions include: (1) Few-shot learning and transfer learning optimization: Introduce meta-learning and domain adaptation techniques, transfer pre-trained models to scarce ancient manuscript restoration tasks, and reduce data dependence. (2) Knowledge graph-driven semantic enhancement: Construct a knowledge graph specific to ancient manuscripts, embed structured knowledge such as historical events and character relationships into the multimodal model, and improve the cultural logic of symbol analysis. (3) Human-machine collaborative evaluation system development: Design an "expert-algorithm" hybrid evaluation framework, optimize restoration strategies through interactive feedback, for example, introduce a cultural sensitivity scoring mechanism. (4) Standardization and open platform construction: Promote the formulation of technical standards for ancient manuscript digitization, and open source some models and datasets to facilitate cross-institutional cooperation and result reuse. In addition, although this study achieves breakthroughs at the technical level, it is necessary to further strengthen data efficiency, semantic depth, and interdisciplinary integration. The future direction focuses on improving the "cultural adaptability" of intelligent tools, promoting the transformation of technology from "restoration means" to "research paradigm".

ACKNOWLEDGEMENT

This paper was funded by the 2025 Hangzhou Philosophy and Social Science Planning Project "Research on Intelligent Revitalization, Inheritance and Innovative Application of Wuyue Culture Ancient Books" (Grant No.: M25YD167).

REFERENCES

- [1] Zheng, W.J., Su, B.P., Feng, R.Q., Chen, S.X. (2023). EA-GAN: Restoration of text in ancient Chinese books based on an example attention generative adversarial network. *Heritage Science*, 11(1): 1-13. <https://doi.org/10.1186/s40494-023-00882-y>
- [2] Rani, N.S., Nair, B.J.B., Chandrajith, M., Kumar, G.H., Fortuny, J. (2022). Restoration of deteriorated text sections in ancient document images using a tri-level

- semi-adaptive thresholding technique. *Automatika*, 63(2): 378-398. <https://doi.org/10.1080/00051144.2022.2042462>
- [3] Jie, T. (2021). Mutual translation between "Painting and Landscape" – A case study on the digital preliminary restoration of "Garden Image" in the engraving illustrations of ancient books in the Ming and Qing dynasties. *E3S Web of Conferences*, 236: 05030. <https://doi.org/10.1051/E3SCONF/202123605030>
- [4] Srinivasa, A.C., Bhat, S.S., Baduwal, D., et al. (2025). GAN-MRI enhanced multi-organ MRI segmentation: A deep learning perspective. *Radiological Physics and Technology*, 18(4): 949-971. <https://doi.org/10.1007/s12194-025-00938-7>
- [5] Zhao, W., Zhu, S., Cao, Y. (2025). Exploration of crop germplasm resources knowledge mining in Chinese ancient books: A route toward sustainable agriculture. *Frontiers in Sustainable Food Systems*, 9: 1560970. <https://doi.org/10.3389/fsufs.2025.1560970>
- [6] Tong, L., Zhang, W., Zhang, L., et al. (2025). Construction and demonstrative application of an evaluation indicator system for the distinction of ancient books of traditional Chinese medicine based on the Delphi method and analytic hierarchy process. *Guidelines and Standards in Chinese Medicine*, 3(2): 138-144. <https://doi.org/10.1097/gscm.0000000000000051>
- [7] Liu, G., Reda, F.A., Shih, K.J., Wang, T.C., Tao, A., Catanzaro, B. (2019). Image inpainting for irregular holes using partial convolutions. In *European Conference on Computer Vision*, Santa Clara, CA United States, pp. 89-105. <https://patentimages.storage.googleapis.com/a2/07/4c/f144e030883d84/US20190295228A1.pdf>
- [8] Huang, D., Huang, W., Yuan, Z., Lin, Y., Zhang, J., Zheng, L. (2018). Image super-resolution algorithm based on an improved sparse autoencoder. *Information*, 9(1): 11. <https://doi.org/10.3390/info9010011>
- [9] Jada, L., Srikanth, R., Bikshalu, K. (2024). Effective low-exposure color image enhancement based on histogram equalization with spatial contextual information. *Engineering Research Express*, 6(4): 045236. <https://doi.org/10.1088/2631-8695/ad8988>
- [10] Ottoni, A.L.C., Ottoni, L.T.C. (2025). A deep learning approach for cultural heritage building classification using transfer learning and data augmentation. *Journal of Cultural Heritage*, 74: 214-224. <https://doi.org/10.1016/j.culher.2025.06.010>
- [11] Elforaici, M.E.A., Montagnon, E., Romero, F.P., et al. (2024). Semi-supervised ViT knowledge distillation network with style transfer normalization for colorectal liver metastases survival prediction. *Medical Image Analysis*, 99: 103346. <https://doi.org/10.1016/j.media.2024.103346>
- [12] Arnob, N.M.K., Rahman, N.N., Mahmud, S., Uddin, N., Rahman, R., Saha, A.K. (2023). Facial image generation from bangla textual description using DCGAN and bangla FastText. *International Journal of Advanced Computer Science and Applications*, 14(6): 1261-1271. <https://doi.org/10.14569/ijacsa.2023.01406134>
- [13] Shao, B., Li, Q., Jiang, X. (2019). A survey of DCGAN based unsupervised decoding and image generation. *International Journal of Computer Applications*, 178(23): 45-49. <https://doi.org/10.5120/ijca2019919099>
- [14] Wang, Y., Li, J., Jing, F., Xue, Y. (2025). DTAM-UIE: Dual transformer aggregation and multi-scale feature fusion for underwater image enhancement. *Journal of Shanghai Jiaotong University (Science)*. <https://doi.org/10.1007/s12204-025-2838-0>
- [15] Xie, S., Li, J., Cai, J. (2025). Improving high-precision BDS-3 satellite orbit prediction using a self-attention-enhanced deep learning model. *Sensors*, 25(9): 2844. <https://doi.org/10.3390/s25092844>
- [16] Jung, H., Park, H., Lee, K. (2023). Enhancing recommender systems with semantic user profiling through frequent subgraph mining on knowledge graphs. *Applied Sciences*, 13(18): 10041. <https://doi.org/10.3390/app131810041>
- [17] Li, Z., Yoshie, O., Wu, H., Mai, X., Yang, Y., Qu, X. (2024). Image analysis method of substation equipment status based on cross-modal learning. *IEEE Transactions on Electrical and Electronic Engineering*, 19(9): 1507-1521. <https://doi.org/10.1002/tee.24111>
- [18] Nguyen, H.A.T., Pham, D.H., Kim, B., Ahn, Y., Kwon, N. (2025). Developing an automated framework for eco-label information categorization using web crawling and Natural Language Processing techniques. *Expert Systems with Applications*, 282: 127688. <https://doi.org/10.1016/j.eswa.2025.127688>
- [19] Vaswani, A., Shazeer, N., Parmar, N., et al. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems (NIPS 2017)*, Long Beach, CA, USA, pp. 1-11. https://proceedings.neurips.cc/paper_files/paper/2017/file_e3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf
- [20] Bu, Y., Chen, T., Duan, H., Liu, M., Xue, Y. (2024). A semi-supervised learning approach for semantic parsing boosted by BERT word embedding. *Journal of Intelligent & Fuzzy Systems: Applications in Engineering and Technology*, 46(3): 6577-6588. <https://doi.org/10.3233/jifs-233212>
- [21] Luo, T., Shi, H., Li, K., et al. (2025). Improved hybrid neural network based on CNN-BiLSTM-Attention for co-estimation of SOC and SOE in lithium-ion batteries. *Journal of Energy Storage*, 131: 117651. <https://doi.org/10.1016/j.est.2025.117651>
- [22] Pan, R., Yuan, Q., Cao, J., et al. (2025). Sentence-resampled BERT-CRF model for autonomous vehicle crash causality analysis from large-scale accident narrative text data. *Accident Analysis & Prevention*, 221: 108184. <https://doi.org/10.1016/j.aap.2025.108184>