Digital Intelligence PRESS

Journal of Discovery Core

https://doi.org/10.67541/jdc2613

Research Article **Volume 1, Issue 3 (Nov 2026)** *Open Access*

Research on Restoration and Completion Algorithm of Cultural Relics Based on Generative Adversarial Network and Diffusion Model

Min Ma *

Academy of Arts and Design, Beijing City University, Shunyi District, Beijing, China

***Corresponding author:** Min Ma. Email: zhangruishengZZ@163.com

Abstract

A decision-level fusion framework was proposed to solve the problem that the texture generated by the generative adversarial network (GAN) is sharp, but the large-scale loss is easy to drift, and the complementarity of the diffusion model is coherent but the fine-grained texture is easy to over-smooth in the restoration of cultural relics. An asymmetric dual-stream generator with feature representation isolation is constructed, and a defect-aware routing controller is designed. The local edge density, fragment compactness, and neighborhood semantic gradient are taken as inputs, and the path is generated by pixelwise allocation through online preference mining. The routing weight modulates the cross-attention of the diffusion stream to the adversarial stream texture features by multiplicative gating, and reverse-infuses the low-level structure prior of the diffusion stream. An adaptive fusion operator guided by routing gradient is proposed to generate the bandwidth point by point according to the weighted gradient tensor and stretch it anisotropic along the main direction of the structure tensor. The three style anchoring constraints of stroke direction, pigment color moment and texture energy spectrum slope are introduced. On the hierarchical incomplete test set, the PSNR of the method reaches 28.94 dB and 25.31 dB in the transition interval from L2 to L3, which is 1.48 dB and 1.29 dB higher than the optimal shallow fusion baseline. The boundary LPIPS is reduced to 0.087, and the style distance is reduced to 0.108. Ablation experiments verify the independent contribution of routing mechanism and feature-level interaction and show that decision-level fusion can effectively realize the complementarity of the two streams under the capability constraints of the existing generation model.

Keywords

Restoration of cultural relics; Generative adversarial network; Diffusion model; Decision level integration; Dynamic routing

Received: 01 Sep 2026; Revised: 21 Sep 2026; Accepted: 30 Oct 2026; Published: 19 Nov 2026
Copyright: ©2026 The Author(s). Published by Digital Intelligence Press Limited. This is an open access article under the CC BY 4.0 license.

1. INTRODUCTION

Over the past decade, the technical route of digital restoration of cultural relics has undergone a profound migration from sample-driven to generation-driven approaches. Early methods centered on sample-block matching and sparse representation relied on local self-similarity between intact and missing regions for texture transfer, and could achieve acceptable completion on regular patterns and smooth color blocks. Their essence, however, is the reorganization of existing pixels rather than the generation of new content: once the semantic structure of the missing region cannot find a transferable counterpart in the neighborhood, the result degrades into a blurred average color or an obvious copy mark. The introduction of deep generative models changed this pattern by reformulating restoration as

a conditional generation problem given incomplete observations and intact regions [1]. The prior learned by the network from large-scale data enables it to infer plausible missing content in the absence of direct sample counterparts, shifting cultural-relic restoration from passive pixel filling to active content inference, and making generation quality itself an object that can be evaluated hierarchically and improved continuously [2].

In current generative restoration practice, two classes of models dominate, and their respective advantages and failure modes exhibit highly complementary spatial distributions. Generative adversarial networks drive the generator to approximate the real data distribution through adversarial game, and the local receptive-field design of their discriminators gives the generated results sharp expressive capability on medium- and high-frequency textures such as stroke direction, fabric pattern, and pigment peeling boundaries, producing details at the junction between missing and intact regions that are highly consistent with surrounding texture statistics [3],[4],[5]. However, the generative decision of this paradigm is local in nature, and the value of each missing pixel is mainly determined by its neighborhood characteristics; when the area of the missing connected domain grows beyond the effective coverage of the receptive field, the cumulative error of local decisions manifests as overall drift of structural trends and collapse of semantic content, so that building contours may curve, figure poses may become unbalanced, and composition axes may skew. Diffusion models generate content by gradual denoising in a low-dimensional latent space, with each denoising step carried out under global semantic constraints; their ability to model long-range structural relations makes content prediction for large missing regions significantly better than the adversarial paradigm in semantic rationality and structural coherence. Yet the randomness of successive denoising tends to produce over-smoothed results in fine-grained textures, or to introduce details inconsistent with the original style of cultural relics in the name of semantic rationality, forming hallucinated content that appears reasonable but actually deviates [6],[7]. Attempts that combine convolutional networks, graph attention, and gradient boosting, and that introduce physical constraints into temporal modeling, further demonstrate the potential of heterogeneous-model collaboration and interpretability enhancement in complex prediction tasks [8].

This pattern of complementarity constitutes the starting point of the present problem. The texture sharpness of generative adversarial networks and the structural fidelity of diffusion models are not simply a matter of quality levels, but a difference in how the two generation behaviors divide labor across spatial scales [9],[10]. Existing studies have recognized that combining the two may bring benefits and have proposed various joint schemes, but these schemes generally remain at the output-end fusion level and fail to touch the generation process itself. In the serial scheme, coarse restoration is first completed by the adversarial network and the result is then sent to the diffusion model for refinement; its fundamental defect is that once the structural drift produced by the adversarial network in a large missing area is formed, it is fed into the diffusion stage as an established fact, so the diffusion model can only add reasonable details on a wrong skeleton and the error is irreversibly transmitted and amplified [11],[12]. The parallel scheme runs two generation paths separately and then fuses the two outputs with fixed weights; its drawback is that the fusion weights are globally set hyperparameters, whereas the scale, semantic density, and style-constraint strength of cultural-relic fragment regions vary extremely drastically, and the same image may contain both fabric-pattern areas requiring sharp texture and figure-contour areas requiring structural inference, so fixed weights cannot adapt to this spatial heterogeneity [13],[14]. The distillation scheme uses the diffusion model as a teacher to guide the adversarial network, attempting to transfer the structural prior of diffusion to the adversarial student; the cost is sacrificing the controllable generation capability of the diffusion model at inference time,

since once training is complete the student network is fixed and cannot dynamically invoke the teacher's generation behavior according to local incomplete characteristics in a specific restoration task [15],[16]. Meanwhile, recent work on discrete diffusion and constraint-guided generation further shows that the controllability of the diffusion process depends heavily on the way conditions are injected and constraints are designed; diffusion-based generation practice for structured objects reveals a tension between semantic constraints and detail fidelity that must be explicitly reconciled; edge-AI-based defect detection practice for design materials shows that combining local structural priors with lightweight decision-making can preserve discriminative power under resource constraints; digital-twin construction with multi-modal sensing fusion demonstrates that heterogeneous information can be coordinated with real-time status monitoring within a unified framework; and attention-based sequence modeling of intangible cultural heritage shadow play suggests that, when properly designed, cross-branch conditional interaction can achieve selective feature fusion without significantly increasing parameter scale [17],[18],[19],[20],[21].

The common crux of the above three schemes is that fusion is understood as an image-level operation after generation is completed, rather than decision-level collaboration within the generation process. This paper therefore establishes the core academic proposition that the complementarity of generative adversarial networks and diffusion models should be realized at the decision-making level of the generation process, that is, a set of incompleteness-aware mechanisms decides pixel by pixel in the feature-generation stage where to rely on the texture expression of the adversarial path, where to rely on the structural inference of the diffusion path, and where the conditional interaction of the two is needed. From this perspective, the fusion problem is transformed from a weighted average of outputs into path allocation and conditional injection within the generation process, so that the two streams constrain each other in the intermediate stage of denoising and decoding, and errors are partially offset by another conditional signal during formation rather than compromised after formation. Around this proposition, the technical line of this paper unfolds along three progressive questions. The first is when to use which: how to extract low-level structural cues from incomplete masks and known regions, and then allocate the dominant weights of the two generation paths pixel by pixel in the feature-generation stage. The second is how to connect: when routing weights have steep spatial gradients, how to design a differentiable edge-consistency fusion operator so that the two stream outputs are continuous in the transition band across the three dimensions of chrominance, texture frequency, and structural direction, with fusion parameters dynamically generated by routing confidence rather than globally set. The third is how much to use: how to take stroke-direction statistics, pigment-aging color moments, and texture energy-spectrum slope as style anchors to constrain texture repetition in the adversarial path and over-creation in the diffusion path, so that the fusion result matches the original cultural-relic region in style statistics. These three designs correspond respectively to the incompleteness-aware routing controller, the edge-consistency fusion operator, and the style-anchoring constraints, which together constitute a complete closed loop of decision-level fusion. The validation strategy centers on hierarchical evidence: on public cultural-relic images and a self-built high-damage subset, fidelity is measured by peak signal-to-noise ratio and structural similarity, perceptual quality by learned perceptual image patch similarity and Fréchet Inception Distance, and style consistency by the distance of three style statistics; test samples are divided by defect scale into small-area peeling, strip-shaped loss, and large-area structural fracture, so as to reveal the applicable boundary of the fusion strategy, and fair comparison is conducted with serial, parallel, and distillation shallow-fusion baselines under the same data and the same backbone, supplemented by module-by-module ablation attribution, interpretable diagnosis of routing behavior, and robustness stress tests under mask morphological shift and cross-material generalization, thereby fully characterizing the effective interval and known limitations of the method.

2. DUAL-STREAM GENERATION ARCHITECTURE AND DYNAMIC ROUTING MECHANISM OF HANDICAP AWARENESS

2.1 Asymmetric design principle of dual-stream generator

The design starting point of the dual-flow architecture is a fact that has been ignored by the existing fusion schemes: the advantage intervals of the generative adversarial network and the diffusion model do not coincide in spatial distribution, and there are structural differences in the induction bias of the two models. If the two streams are designed as symmetric structures and share encoders or intermediate features, the early feature coupling will gradually homogenize the two streams in training and eventually degenerate into two approximately identical generators, and the fusion will lose significance. Therefore, this paper adheres to the asymmetric design. The two streams only share the same incomplete mask code, and the other features represent complete isolation.

Generative adversarial stream undertakes sharp reconstruction of high-frequency textures. The generator adopts the encoder decoder structure with local receptive field bias, the encoder is composed of stacked residual convolution blocks, and the decoder returns the spatial resolution step by step through jump connection, and introduces gated convolution at each level, so that the convolution response of the missing region is driven only by effective pixels. Let the gated convolution output of layer l be F^l, and its calculation can be written as:

$$F^l = \phi\left(W^l \cdot \left(F^{l-1} \odot M^l\right) + b^l\right) \tag{1}$$

Where W^l and b^l are the convolution kernel and bias of this layer, M^l is the layer-by-layer gating map obtained by the undersampling of the incomplete mask, $\odot$ is the element-by-element product, and ϕ is the activation function. The meaning of this equation is that the response at the missing position is zeroed by the gated map, and the convolution only aggregates information from valid pixels, thus avoiding the pollution of texture estimation by invalid values. The receptive field of this flow is deliberately restricted to a medium local scale to prevent the network from making unfounded global semantic guesses in large missing areas. The training objective consists of adversarial loss and perceptual loss. The adversarial branch adopts a local Patch discriminator, forcing the generated texture to be statistically indistinguishable from real artifacts at the stroke scale. The flow has a clear generation advantage on regular texture, repeated pattern and pigment peeling-off boundary, but when the scale of missing connected domain increases, the cumulative error of local decision will lead to structural trend drift.

The diffusive flow undertakes the consistent maintenance of the global structure. The backbone of the diffusion stream is a latent-space conditional denoising U-Net [22]. The input image is compressed to a low-dimensional latent representation by a pre-trained autoencoder, and the denoising U-Net performs conditional denoising on the manifold. Let z_t be the noisy potential representation at step t, and the condition term is composed of the incomplete mask coding and the potential features of the known region. The denoising target can be written as:

$$L_d = \mathbb{E}_{z_0,\epsilon,t}[\| \epsilon - \epsilon_\theta(z_t, t, c_m, c_k) \|_2^2] \tag{2}$$

Where z_0 is the latent representation of the intact image, ϵ is the injected noise, ϵ_θ is the de-noised U-Net, c_m is the mask coding condition, and c_k is the latent feature condition of the known region. This objective makes each step of denoising be carried out under the semantic constraints of the

intact region. The flow operates on a low-dimensional manifold and naturally has the ability to model long-range structural relationships. It can maintain the rationality of building contours, character postures and composition axes under the condition of strip and large area missing. However, the randomness of its successive denoising is prone to produce excessive smoothness or details inconsistent with the original stroke statistics in fine-grained textures.

The training strategies of the two streams are independent of each other. The generative adversarial stream is trained to convergence with the weighted sum of adversarial loss and perceptual loss as the target, while the diffusive flow is trained to convergence independently with the standard denoising target. The two do not have any gradient interaction before entering the routing stage. The significance of this warm-up strategy is to ensure that each flow reaches the upper bound of its own generation capacity within its advantage interval, so that the subsequent routing decision is faced with two experts who have a clear division of labor, rather than two semi-finished products that have not been finalized. Table 1 summarizes the asymmetric Settings of the two streams in structural configuration and training objectives.

Table 1. Asymmetric configuration of dual-stream generators

Dimension of configuration	Generate adversarial stream	Flow of diffusion
Backbone structure	Gated convolutional codec	Latent spatial conditional denoising U-Net
Receptive field bias	Medium local scale	Global scale
Representation of features	Completely isolated from diffusion stream	Completely isolated from adversarial streams
Item of sharing	Incomplete mask coding	Incomplete mask coding
Main training objectives	Adversarial loss and perceived loss	Standard denoising target
Interval of advantage	High frequency texture and regular texture	Global structure and long-range semantics
Typical failure mode	Structural drift under large-scale missing	Fine grain texture is excessively smooth

The configuration difference in Table 1 is not an arbitrary choice in engineering, but an explicit respect for the inductive bias of the two generative paradigms. It is this deliberate separation at the structural level that enables the two streams to have distinguishable generation behaviors in the subsequent routing stage, so that the routing controller has assignable objects.

2.2 Impartially aware routing controller

The routing controller is the core innovation module of this section, and its design philosophy is to transform path selection from manual heuristics to learnable decision behavior. The input of the controller is not the original pixel, but three types of low-level structural cues extracted jointly from the incomplete mask and known regions, which jointly describe the relative credibility of the two generated paths at each spatial location.

The first type of cue is the local edge density, which counts the structural tensor response intensity of known regions within a window centered on the target pixel to reflect the strength of structural constraints near the missing boundary. Let the two eigenvalues of the structure tensor in the window be

λ_1 and λ_2, and the edge density is defined as:

$$e_p = \frac{1}{|\mathcal{N}(p)|}\sum_{q\in\mathcal{N}(p)}\frac{\lambda_1(q)-\lambda_2(q)}{\lambda_1(q)+\lambda_2(q)+\varepsilon} \tag{3}$$

Where $\mathcal{N}(p)$ is the neighborhood centered at position p, and ε is the small quantity that prevents division by zero. A high edge density means that the location is adjacent to a clear structural trend, and the global structural prior of the diffusion stream is more valuable. The second type of cue is the compactness of the incomplete region, which is measured by the area of the missing connected domain and the shape complexity. Low compactness means that the missing region is narrow or broken, and it is more appropriate to generate local texture completion against flow. The third type of cue is neighborhood semantic gradient, which estimates the semantic richness of the location based on the local statistical dispersion of known regional features. High semantic gradient indicates that the content categories of the neighborhood are mixed, and it is difficult to take into account a single generation path, so finer weight allocation is needed.

The controller itself is a lightweight fully convolutional network, which is alternately composed of three pointwise convolution layers and two dilated convolution layers. The number of parameters is controlled within 3% of the backbone network to ensure that routing decision does not become a new computing bottleneck. The output is a pixel-by-pixel dual-channel weight map, which after SoftMax normalization represents the allocation probability of the location dominated by generative adversarial stream and diffusive flow respectively. Let the original output of the controller to position p be $u_p = [u_p^g, u_p^d]$, then the routing weight is:

$$\omega_p^k = \frac{\exp(u_p^k/\tau)}{\sum_{j\in\{g,d\}}\exp(u_p^j/\tau)},\ k\in\{g,d\} \tag{4}$$

Where τ is the temperature coefficient, which controls the sharpness of the weight distribution. ω_p^g and ω_p^d correspond to the adversarial stream and diffusive flow respectively, which satisfy $\omega_p^g + \omega_p^d = 1$. This weight will be directly involved in subsequent feature-level conditional interactions.

The training signals of the controller come from online preference mining rather than manually annotated routing labels. Specifically, in the training stage, the perceived distance between the separate generated results of two streams and the true value of the location is calculated for each location, and the side with the smaller distance is taken as the preferred path of the location, so as to construct the routing supervision signal. Let x_p^g and x_p^d be the separate generated results of the two streams at position p, y_p be the true value, and the preference label and routing loss are written as follows:

$$\hat{k}_p = \arg\min_{k\in\{g,d\}} \|\psi(x_p^k)-\psi(y_p)\|_2^2,\ L_r = -\frac{1}{|\Omega|}\sum_{p\in\Omega}\log\omega_p^{\hat{k}_p} \tag{5}$$

Where ψ is the perceptual feature extractor and Ω is the set of missing region pixels. Since the preference label is dynamically updated with the evolution of the generation capability of the two streams, the controller learns what is the optimal path allocation in the current generator state, rather than a static prior rule. Compared with the output decision controller, this controller works in the feature generation stage, and the routing granularity is reduced from image level to pixel level, which is the fundamental reason why it can deal with the spatial heterogeneity of texture and structure requirements in the same missing region. Table 2 shows the corresponding relationship between controller input cues

and routing tendencies, which is obtained from the trained controller behavior statistics rather than preset rules.

Table 2. Statistical correspondence between routing cues and path tendencies

Density of edges	Degree of incomplete tightness	Gradient of semantics	Dominant path orientation	Proportion of statistics
High	High	Low	Flow of diffusion	31.2%
High	Low	High	Flow of diffusion	18.7%
Low	High	Low	Flow of resistance	26.5%
Low	Low	High	Two flow equilibrium	14.3%
Low	Low	Low	Flow of resistance	9.3%

Table 2 shows that edge density and incomplete compacting are the dominant cues that determine the path tendency, and semantic gradient mainly plays a regulatory role when the two signals conflict. The proportion of balanced routes is less than 15%, indicating that the path ownership of most spatial locations is clear, and the routing mechanism does not degenerate into meaningless soft average.

2.3 Feature-level conditional interactions in the generation process

If the routing weight is only used for the weighting of the final output, it is no different from the output level fusion. The key design of this paper is to make the routing weight intervene in the generation process itself, so that the two streams form bidirectional, asymmetric and spatially variable coupling in the intermediate stage of denoising and decoding.

For diffuse flows, routing weights modulate the intensity of cross-attention to the intermediate features of generated adversarial streams during its denoising process. In the middle layer of the diffused U-Net, each spatial location is allowed to query the texture features of the corresponding scale of the counterflow decoder, and the strength of the query is determined by the counterflow weight at that location. Let the feature of layer l of diffusion stream be H^l, and the corresponding scale feature of adversarial stream be G^l. The cross-attention output after modulation is written as:

$$\widetilde{H}^l = H^l + \Lambda(\omega^g) \cdot \mathrm{Attn}\left(H^l, G^l, G^l\right) \tag{6}$$

Where $Attn$ is the standard multi-head cross-attention, $\Lambda(\omega^g)$ is the diagonal gating matrix formed by downsampling the route weight graph to this scale, and its diagonal element is ω_p^g. At the position with high weight, the diffusion stream obtains sharp texture conditions, and its denoising trajectory is drawn to the result with richer details at high frequency. In locations with low weights, cross-attention is suppressed, and the diffusion stream remains autonomously generated to avoid interfering with the global structure by local decisions of the counterflow. The modulation is implemented by multiplicative gating, and the gating coefficient is directly taken from the down sampled version of the routing weight graph at this scale to ensure the consistency of spatial correspondence.

The reverse path also holds and is asymmetric. The generator that generates the counterflow

receives a structural prior injection from the low-level features of the diffusion stream in the region with high structural uncertainty, and the injection intensity is determined by the diffusion stream weights at that location. Let the l-th layer of the adversarial stream be characterized by G^l, and the feature of the lower layer of the diffusion stream be $P^l(H^s)$ after projection. The injection operation is written as:

$$\tilde{G}^l = G^l + \Lambda(\omega^d) \cdot P^l(H^s) \tag{7}$$

Where P^l is the linear projection mapping the features of the s-layer of diffusion stream to the channel dimension of the l-layer of adversarial stream, and $\Lambda(\omega^d)$ is the gating matrix composed of the route weights of diffusion stream. The low-level features of the diffusion stream carry the long-range structural relations formed in the latent space, which are projected to the intermediate scale of the adversarial stream and then spliced as additional conditions, so that the adversarial stream can avoid structural drift in the edge zone with large area missing. The interaction of the two directions is not equal in scale. The diffusion stream queries the texture features at the deeper level of the adversarial stream, while the adversarial stream receives the structural features at the lower level of the diffusion stream. This dislocation in depth is the respect for the respective information organization mode of the two streams. Table 3 compares the differences in coupling attributes between feature-level interaction and existing output-level fusion in this paper.

Table 3. Comparison of coupling attributes between feature-level interaction and output-level fusion

Property of coupling	Output stage fusion	This paper features level interaction
Location of coupling occurrence	After the generation is completed	Denoising and decoding intermediate layer
Routing action granularity	Image level or block level	Pixel level
Direction of interaction	One way weighting	Two-way asymmetry
Spatial variability	Fixed fusion nucleus	Change point by point with route weight
Propagation of error	Generation error is transmitted irreversibly	The intermediate stage can be modified by conditions

The core message of Table 3 lies in the differences in the error propagation properties. In the output level fusion, the generation error of any class has been finalized before the fusion, and the fusion can only be compromised rather than corrected. Feature-level interaction makes the two streams constraint each other in the middle of generation, and the error is partially offset by another first-class conditional signal in the formation process. This difference is the essence of the difference between decision-level fusion and output-level fusion, and its quantitative consequences will be presented as hierarchical indicators in the experimental section.

2.4 Training scheduling and convergence guarantee

The joint optimization of dual-flow and routing controllers faces two types of training dynamics conflicts. The first is the conflict of target properties, and the loss of adversarial stream comes from the game between discriminator and generator. The gradient scale fluctuates violently with the state of discriminator, while the denoising target gradient of diffusion stream is stable and small in magnitude.

If the joint optimization is direct, adversarial gradient will dominate parameter update. The second is the non-stationarity of routing supervision. The preference label depends on the current generation ability of the two streams. At the initial stage of training, the ability of the two streams is weak, and the preference signal has large noise.

Accordingly, three-stage scheduling is adopted in this paper. The first stage is independent warm-up, and the two streams are trained to convergence according to their respective objectives. In this stage, the controller does not participate, and the preference label is not collected. In the second stage, all backbone parameters of the two streams are frozen, and only the routing controller and cross-attention adaptation layer are trained. In this stage, the generation ability of the two streams is fixed, the preference label is stable, and the controller can learn path allocation based on the determined expert behavior. The third stage is end-to-end fine-tuning, which unfroze all parameters but reduces the learning rate to one-tenth of that of the warm-up stage. The controller coevolves with the two streams in a small step to avoid drastic destruction of the learned routing policy.

Aiming at the gradient scale difference, this paper introduces gradient normalization and alternating update in the joint stage. Let the gradient norm of the adversarial stream and the diffusive flow at the i-th iteration be $\| g_g^{(i)} \|$ and $\| g_d^{(i)} \|$ respectively, and the scaled adversarial stream gradient is written as:

$$\tilde{g}_g^{(i)} = g_g^{(i)} \cdot \min\left(1, \frac{\| g_d^{(i)} \|}{\| g_g^{(i)} \| + \varepsilon}\right) \tag{8}$$

In this formula, the adversarial stream gradient norm is truncated to no more than the diffuse flow gradient norm, so that the magnitude of the two is comparable, and then the parameter update is performed. At the same time, alternating iteration is used to alternately update the two stream backbones, so that only one flow's parameters change rapidly at any time, and the other flow provides a relatively stable reference for routing decisions. The above scheduling does not involve the theoretical convergence proof, and its effectiveness is demonstrated by the stationarity of loss curve and the entropy stability of routing weight distribution in the training process. Relevant quantitative results are given in the experimental section. It should be emphasized that the essence of three-stage scheduling is to decompose a non-stationary joint optimization problem into two stationary subproblems plus a weak non-stationary fine-tuning problem, so as to avoid the risk of training collapse when adversarial game is directly coupled with denoising regression in engineering.

3. EDGE CONSISTENCY INTEGRATION AND STYLE ANCHORING CONSTRAINTS

3.1 Diagnosis of perceived discontinuity at route boundaries

The routing controller outputs pixel-by-pixel weights. When the weights have steep gradients in space, the pixels generated by the two streams have three types of quantifiable discontinuities at the boundary. The first type is chrominance shift. The two streams have different estimated distributions of the pigment base color. The generative antagonistic flow tends to match the local pixel statistics sharply, while the diffusive flow tends to maintain the global hue integrity in the latent space. The second type is texture frequency fracture. The high frequency energy of the adversarial stream is significantly higher than that of the diffuse flow, and the hard switching of frequency components is formed at the boundary,

which is extremely sensitive to the human eye. The third type is the misalignment of structural cues, where the estimation of the edge direction by the two streams is inconsistent on both sides of the boundary, resulting in the folding or breaking of the structural direction that should be continuous.

Diagnosis is based on statistical indicators rather than visual observation. The magnitude of the spatial gradient of the routing weight is given as $|\nabla\omega_p|$, and it is divided into confidence region and transition band according to quantile in the missing region. For each route boundary, narrow band $\mathcal{B}$ is taken along the normal direction of the boundary, chromaticity shift is measured by the local color mean difference between the two streams on both sides of the narrow band, texture frequency break is measured by the jump variable of high frequency energy in the narrow band along the normal direction, and structural misalignment is measured by the Angle of the main direction of the structure tensor on both sides of the narrow band. The three indicators are written as follows:

$$D_c = \| \mu_g(\mathcal{B}) - \mu_d(\mathcal{B}) \|_2 \tag{9}$$

$$D_f = | E_h^g(\mathcal{B}) - E_h^d(\mathcal{B}) |, \; D_s = | \theta_g(\mathcal{B}) - \theta_d(\mathcal{B}) | \tag{10}$$

Where μ_g and μ_d are the local color mean values of the two streams in a narrow band, E_h^g and E_h^d are the high frequency energies of the two streams, and θ_g and θ_d are the main directions of the structure tensors. Table 4 shows the statistical measurement results of the three types of discontinuities under the three handicap levels. The data are aggregated according to the handicap level after the point-by-point statistics of the narrow band of the route boundary.

Table 4. Hierarchical statistics of three types of discontinuities at route boundaries

Grade of disability	Chroma shift D_c	Frequency break D_f	Structural dislocation D_s	Proportion of border pixels
L1 slight loss	3.42	0.087	4.1°	6.8%
L2 medium loss	5.76	0.154	7.3°	14.2%
L3 heavy loss	8.91	0.236	11.6°	21.5%

Table 4 shows that the three types of discontinuities rise monotonically with the level of disability, and the growth of structural dislocation is the largest, indicating that the divergence of structural estimation between the two streams under large-scale missing is the main threat to the quality of splicing. The frequency break is of considerable magnitude under L1, reflecting that hard switching cannot be ignored even in micro-loss scenarios. The proportion of boundary pixels increases with the impairment level, which means that the range of action of the fusion operator must expand adaptively with the spatial distribution of routing weights, and the fusion with fixed bandwidth cannot cover the transition band width under heavy loss. The above quantization provides a clear goal for the design of the fusion operator: narrowing in the confidence region to retain the advantages of each flow, broadening in the transition band and stretching along the direction of the structure.

3.2 Improvement of alpha fusion operator for edge perception

Poisson fusion with fixed kernel and simple alpha blending regard the fusion bandwidth as a global hyperparameter, which cannot adapt to the spatial variation of routing confidence, let alone align the hybrid region with the inherent structure of cultural relic [23],[24],[25]. In this paper, the adaptive fusion

guided by routing gradient is proposed, and its bandwidth and kernel shape are dynamically generated by the spatial gradient tensor of routing weights.

The first step of fusion is to construct a pixel-by-pixel fusion bandwidth. Let the routing weight gradient amplitude at position p be $|\nabla\omega_p|$, and the fusion bandwidth be written as:

$$\sigma_p = \sigma_0 \cdot \exp(\beta \cdot |\nabla\omega_p|) \tag{11}$$

Where σ_0 is the base bandwidth and β is the bandwidth extension coefficient. The meaning of this formula is that the steeper the routing gradient is, the location is in the transition band of fuzzy decision making, and the fusion bandwidth should be broadened to smooth the hard handover. When the routing gradient is gentle, the decision confidence is high, and the bandwidth is narrowed to minimize the dilution of the advantages of the two streams. Compared with fixed bandwidth, this construction makes the scope of fusion and routing uncertainty correspond strictly in space.

The second step of fusion is the anisotropic stretching of the shape of the nucleus. The isotropic Gaussian kernel in the region where the structure direction is obvious will cause excessive mixing across the structure direction and blunt the edge [26],[27]. In this paper, the anisotropic kernel is constructed by the principal direction of the local structure tensor, and its covariance matrix is written as:

$$\Sigma_p = R(\theta_p)\,\mathrm{diag}\left(\sigma_p, \frac{\sigma_p}{\gamma_p}\right) R(\theta_p)^{\top} \tag{12}$$

Where $R(\theta_p)$ is the rotation matrix along the main direction θ_p of the structure tensor, and γ_p is the anisotropy ratio, which is determined by the eigenvalue ratio of the structure tensor. The core is stretched along the structural strike and compressed vertically, so that the transition direction of the mixing region is aligned with the inherent edge of the cultural relic, and the edge passivation caused by cross-structure mixing is avoided.

The third step of fusion is pixel-by-pixel weighted synthesis. Let the generation results of the two streams at position p be x_p^g and x_p^d, and the fusion output be written as:

$$\hat{x}_p = \alpha_p\, x_p^g + (1-\alpha_p)\, x_p^d \tag{13}$$

$$\alpha_p = \frac{\sum_q K_{\Sigma_p}(p-q)\,\omega_q^g}{\sum_q K_{\Sigma_p}(p-q)} \tag{14}$$

Where K_{Σ_p} is the Gaussian kernel with Σ_p as the covariance, and α_p is the fusion coefficient after kernel weighting and smoothing. The operator is differentiable for both stream output and route weight and can be directly embedded in end-to-end fine-tuning. Compared with the existing edge-aware mixing, the fusion parameters in this paper are generated point by point by route confidence rather than set globally, and the kernel shape is oriented by the structure tensor, which jointly ensure the continuity of the fusion boundary in the three dimensions of chromaticity, frequency and structure. Table 5 compares the residual discontinuities of different fusion strategies in the narrow band of the routing boundary.

Table 5. Comparison of boundary residual discontinuities of different fusion strategies

Convergence strategy	Remnant D_c	Remnant D_f	Remnant D_s	Boundary LPIPS
No integration	8.91	0.236	11.6°	0.214
Fixed-kernel alpha blending	5.12	0.141	8.7°	0.163
Fixed kernel Poisson	4.38	0.128	7.9°	0.151
In this paper, adaptive fusion	1.96	0.052	3.4°	0.087

Table 5 shows that the fusion operator in this paper reduces the pressure of the three residual discontinuities to 22%, 22% and 29% of the non-fusion state respectively, and the boundary LPIPS to 41% of the non-fusion state. Although the fixed kernel method has some improvement, due to the limitation of global bandwidth and isotropic kernel, the residual in the dimension of structural dislocation is still more than twice that of the operator in this paper. The source of this gap is the point-by-point adaptation of bandwidth and kernel shape to route confidence and structure direction.

3.3 Triple statistics of style anchoring constraints

The fusion operator solves the physical continuity of spatial Mosaic, but the two streams may still deviate from the inherent style of cultural relics in their respective dominant areas. The local texture matching of generative adversarial stream is easy to produce texture repetition in the repetitive pattern region, while the latent spatial denoising of diffusive flow is easy to introduce over-creation in the name of semantic rationality. Both do not exhibit boundary discontinuity, but statistical shift within the region. In this paper, three statistics sensitive to cultural relic style and differentiable are designed as regular terms, which are estimated in real time from known intact areas and serve as dynamic anchors in the generation process rather than post-processing correction after training.

The first statistic is the KL divergence of the stroke direction histogram. The weighted histogram of local gradient direction is calculated in the known region and the generated region respectively, and the direction distribution of the generated region is constrained to be consistent with the known region, which is written as [28],[29]:

$$L_o = \sum_b h_b^{\text{gen}} \log \frac{h_b^{\text{gen}}}{h_b^{\text{known}}} \tag{15}$$

Where h_b^{gen} and h_b^{known} are the normalized histograms of the generated region and the known region on the b-th directional interval, respectively. This constraint inhibits the texture repetition of the generative antagonism flow in the repetitive pattern region, because the repetitive pattern will make the orientation histogram abnormally concentrated in a specific interval, and the KL divergence will rise accordingly.

The second statistic is the moment matching of the pigment aging color distribution. The aging of cultural heritage pigments is manifested as fading and discoloration, which is reflected in the color space as a specific shift between the first moment and the second moment. The mean and covariance of the constrained generated region and the known region in the color space are consistent, which is written as:

$$L_m = \| \mu_{\text{gen}} - \mu_{\text{known}} \|_2^2 + \| \Sigma_{\text{gen}} - \Sigma_{\text{known}} \|_F^2 \quad (16)$$

Where μ and Σ are the mean vector and covariance matrix of the color space respectively, $\|\cdot\|_F$ is the Frobenius norm. This constraint inhibits the over creation of the diffusion stream from the global tone in the latent space, so that the fading and discoloration characteristics of the generated region are the same as those of the known region.

The third statistic is the radial attenuation slope of the texture energy spectrum. The spatial frequency distribution of cultural relic texture follows a specific power law decay, and its slope reflects the roughness and stroke scale of the material. The slope of the radial energy spectrum of the constrained generated region is consistent with that of the known region, which is written as [30]:

$$L_s = \left| \frac{d \log E_{\text{gen}}(r)}{d \log r} - \frac{d \log E_{\text{known}}(r)}{d \log r} \right|^2 \quad (17)$$

Where $E(r)$ is the radial average energy spectrum and r is the radial frequency. This constraint suppresses both the high frequency overshoot of the generative adversarial stream and the high frequency loss of the diffusive flow, so that the fusion result remains unified at the spectrum level. The three statistics are added to the loss of the two streams as regular terms in training, and their weights are determined by ablation. Table 6 shows the independent contributions of the three constraints to the style and perception indicators.

Table 6. Independent contributions of style anchoring constraints

Constraint configuration	The directional KL divergence	Color moment distance	Spectrum slope deviation	LPIPS
No style constraints	0.412	6.83	0.197	0.142
Stroke direction only	0.156	6.41	0.184	0.138
Color moment only	0.387	2.94	0.176	0.131
Just the spectrum slope	0.374	6.52	0.061	0.129
The triple constraint	0.148	2.87	0.058	0.104

Table 6 shows that the three constraints are effective for different statistical dimensions respectively: the stroke-direction constraint mainly reduces the KL divergence, the color moment constrains the color moment distance of the main pressure drop, and the spectrum slope constrains the spectrum deviation of the main pressure drop. After the superposition of the three, all dimensions reach the lowest and LPIPS further decreases. This result indicates that the style shift is multidimensional, that a single statistic cannot cover the full style risk, and that a combination of triple constraints is necessary. At the same time, it is noted that the introduction of style constraints does not come at the cost of the fidelity index, and the synchronous decline of LPIPS indicates that style consistency and perceived quality are in the same direction under the framework of this paper.

3.4 Joint fine-tuning with routing mechanism

The fusion operator and style constraint are not independent of the routing mechanism of the post-processing module, and its gradient acts on the routing controller and the two flow generator in reverse

in the end-to-end fine-tuning stage, forming a closed-loop adjustment of the output end to the decision layer.

The perceptual loss gradient at the fusion boundary is transmitted back to the routing controller, and its direction is to smooth the routing weight in space. The reason is that when the route weight oscillates violently locally, the fusion bandwidth is repeatedly widened, the residual discontinuity of the boundary increases, and the perception loss increases. After the gradient is transmitted back by the partial derivative of the route weight by the fusion operator, the oscillating route distribution will be punished, and the spatially coherent route distribution will be rewarded [31],[32]. Let the perceptual loss of the fusion output be L_p, and its gradient to the route weight is written as:

$$\frac{\partial L_p}{\partial \omega_p^g} = \frac{\partial L_p}{\partial \hat{x}_p} \cdot \frac{\partial \hat{x}_p}{\partial \alpha_p} \cdot \frac{\partial \alpha_p}{\partial \omega_p^g} \tag{18}$$

The three terms on the right are the gradient of perception loss to fusion output, the gradient of fusion output to fusion coefficient, and the gradient of fusion coefficient to route weight respectively. The third term is introduced by kernel weighted smoothing, whose smoothness ensures the stability of gradient return, so that the controller tends to output spatially coherent weight distribution without losing pixel-by-pixel discrimination ability in fine tuning.

The gradient of the style constraint acts on both stream generators simultaneously. For the generated adversative flow, the gradient of stroke direction and spectrum slope constraint is the strongest in its dominant region, which suppresses texture repetition and high frequency overshoot. For diffusion stream, the gradient of color moment constraint is strongest in its dominant region, inhibiting the excessive creation of deviation from global hue. The two streams further move closer to the style anchor point in their respective advantage regions, and the routing controller maintains the rationality of path allocation in this process. The overall goal of joint fine-tuning is written as follows:

$$L_{\text{total}} = L_g + \lambda_d L_d + \lambda_r L_r + \lambda_p L_p + \lambda_o L_o + \lambda_m L_m + \lambda_s L_s \tag{19}$$

Where L_g is the generative adversarial stream loss, L_d is the diffusion stream denoising loss, L_r is the routing loss, L_p is the perception loss, L_o, L_m and L_s are the three style constraints, and each λ is the weight coefficient. In the fine-tuning stage, the learning rate is reduced to one-tenth of that in the warm-up stage, so that the adjustment of the output gradient to the decision layer is carried out in a gradual manner, avoiding the drastic disturbance of the routing policy in the initial fine-tuning stage. The architecture design of the above joint fine-tuning takes the formalization of the loss function as the core, and the quantitative evidence of its convergence behavior and index returns is given in the experimental section.

4. EXPERIMENTAL VERIFICATION AND ABLATION ANALYSIS

4.1 Data set construction and incomplete hierarchical protocol

The experimental data are collected from the selected and standardized images of murals and paintings in the public cultural relic image collection, and the basic corpus is formed after unified resolution and removal of severe compression artifacts. Based on this, three subsets of defect grades are constructed: L1 minor damage corresponds to spalling points and scratches with a missing area of less than 5%, L2 medium damage corresponds to strip-shaped or block-shaped missing areas ranging from

5% to 25%, and L3 severe damage corresponds to large-area structural fractures exceeding 25%. The generation of incomplete masks does not use random rectangles, but samples according to the real cultural relic disease morphology and takes edge irregularity and connected domain complexity as hierarchical variables, so that the masks within each level have controlled diversity in morphology.

Data division is isolated by cultural relic items, and different incomplete samples of the same cultural relic only appear in one division to avoid homologous leakage. The training set, validation set and test set maintain the same item distribution at the three handicap levels, so that the stratification index can be compared horizontally on the same test benchmark. Table 7 summarizes the size and shape statistics of each subset.

Table 7. Size and shape statistics of the stratified subset of disabilities

Subset	Number of images	Average missing area	Average number of connected domains	Degree of edge irregularity	Number of items
L1 slight loss	2860	2.7%	6.4	1.32	412
L2 medium loss	3140	14.8%	3.1	1.87	447
L3 heavy loss	2470	38.6%	1.6	2.24	356

Table 7 shows that the three subsets show an obvious gradient in the missing area, and the number of connected domains decreases with the impairment level while the edge irregular degree increases, indicating that the heavy loss samples are dominated by a small number of large, connected domains with complex morphology, which is the morphological condition for the routing mechanism to play a role. The ratio of the number of items to the number of images remains stable among the three subsets, ensuring that the cultural heritage diversity of each subset is comparable after separation.

4.2 Comparative experimental design

The baseline covers three categories of methods. Pure generative adversarial methods include the cultural relic restoration model based on gated convolution and the restoration model based on edge prior, which represent the current level of local texture generation. The pure diffusion method is the latent spatial repair model that completes the intra-domain adaptation on the data in this paper, representing the current level of global structure generation. The shallow fusion baseline includes three types: tandem, parallel and distillation: tandem is coarse repaired by generative adversarial stream and then refined by diffusion stream; parallel is fused with fixed weight for the output of the two streams; distillation is guided by diffusion teachers to generate adversarial students. All methods are retrained or fine-tuned under the same training data and the same mask distribution, and the pre-training model adopts the same scale of intra-domain adaptation to ensure the fairness of comparison.

The evaluation indicators simultaneously report the three dimensions of fidelity, perceived quality and style consistency. The peak signal-to-noise ratio and structural similarity are used for fidelity, the similarity of image blocks and Fréchet Inception Distance are used for perceptual quality, and the weighted combination of three statistics of stroke direction KL divergence, color moment distance and spectrum slope deviation are used for style consistency. The hierarchical results are the focus of the analysis because the failure mode of shallow fusion has a clear incomplete rank dependence. Figure 1 shows the hierarchical comparison of the perceived quality of each method at the three levels of

disability.

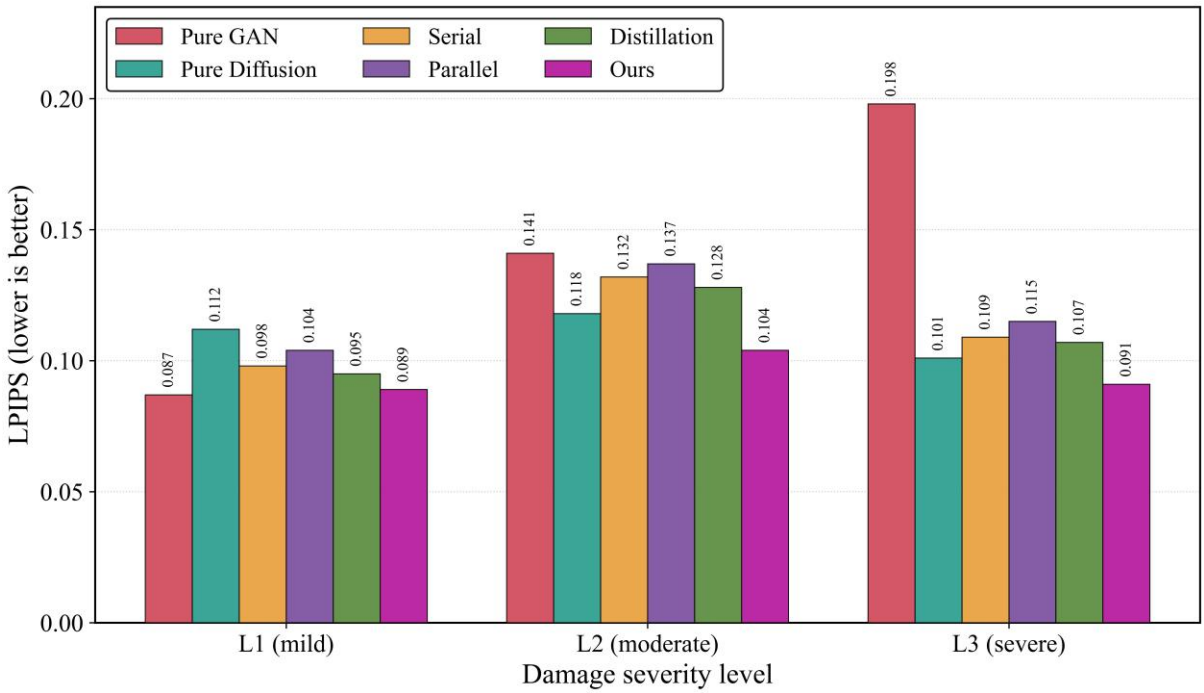


Figure 1. Hierarchical comparison of LPIPS of each method at the three levels of disability

The pure generative advermation method has the lowest LPIPS on L1, which deteriorates rapidly with the increase of the disability level and becomes the worst on L3. The pure diffusion method shows the opposite trend, which is inferior to the generative adversarial method on L1 and close to optimal on L3. The curves of the three types of shallow fusion baselines are in between, with tandem and distilling approaching pure diffusion on L3 but still having a gap, and parallel performing mediocre on all three levels. The curve of the proposed method is close to that of the pure generative adversarial method on L1, and close to that of the pure diffusion method on L3, while the transition interval from L2 to L3 is significantly lower than all the baselines.

Table 8 provides numerical supplementation to the stratified results concerning fidelity and style consistency.

Table 8. Fidelity and style consistency of each method on the three handicap levels

Method	L1 PSNR/SSIM	L2 PSNR/SSIM	L3 PSNR/SSIM	Mean of style distance
Pure GAN	31.24/0.941	26.87/0.883	22.13/0.796	0.187
Pure diffusion	29.86/0.918	27.92/0.901	24.68/0.847	0.142
Serial fusion	30.41/0.926	27.35/0.892	23.94/0.831	0.151
Parallel type	30.12/0.921	27.18/0.889	23.57/0.824	0.158
Distillation type	30.58/0.929	27.46/0.894	24.02/0.835	0.147
Proposed method	31.07/0.938	28.94/0.917	25.31/0.862	0.108

Table 8 shows that the stratification pattern of the fidelity indicators is consistent with the perceived quality. The pure generative adversarial method has the highest PSNR on L1 and lags behind the pure diffusion method by 2.55 dB on L3. The pure diffusion method leads on L3 and lags by 1.38 dB on L1. The PSNR of the three types of shallow fusion baselines on L2 is not more than 27.5dB, while the proposed method reaches 28.94dB, 1.48dB higher than the optimal baseline. The column of style distance mean shows that the proposed method is about 27 percent ahead of the optimal baseline in terms of the three levels of comprehensive style consistency. The serial and distillation methods are close to the pure diffusion method on L3, indicating that shallow fusion can approach its dominant flow on a single level, but it cannot take into account the advantages of the two streams at the same time in the transition interval, which is the distinction between decision-level fusion and output level fusion.

4.3 Ablation experiments

The ablation experiments were stripped one by one according to the design modules, and the order was consistent with the structure of the method section to ensure the independence of each attribution. The first term tests the necessity of a routing controller by replacing it with a fixed route, including two variants of equal weight versus global weight based on mask area. Table 9 presents the hierarchical results for routing ablation.

Table 9. Hierarchical indicators of routing controller ablation

Routing configuration	L1 LPIPS	L2 LPIPS	L3 LPIPS	L2 PSNR	Distance of style
Equal weight	0.121	0.174	0.138	27.41	0.163
Area global weight	0.118	0.169	0.134	27.63	0.157
This paper DARC	0.087	0.104	0.091	28.94	0.108

Table 9 shows that the degradation of fixed routing on L1 and L3 is relatively limited, but the LPIPS on L2 is 67 percent and 62 percent higher than that of the controller in this paper, respectively, and the PSNR is 1.53 dB and 1.31 dB lower. This hierarchical difference accurately corresponds to the value interval of the routing mechanism: under the condition of micro loss and heavy loss, the relationship between the advantages and disadvantages of the two streams is global, and the fixed weight can be approximated. Under the condition of medium loss, texture demand and structure demand in the same missing region are spatially heterogeneous, so pixel by pixel routing becomes necessary. The area global weight is slightly better than the equal weight, indicating that the coarse-grained incomplete statistics contain some information, but are far from sufficient to replace the pixel-by-pixel decision.

The second term examines the contribution of feature-level interactions to output-level fusion. The routing weight is kept unchanged, and the interaction mode is only downgraded to alpha fusion at the output end to isolate the difference between the coupling modes at the decision level and the output level. Table 10 presents the results of this ablation.

Table 10. Comparison of feature-level interaction and output-level fusion

Mode of coupling	L2 PSNR	L2 LPIPS	L3 PSNR	L3 LPIPS	Boundary LPIPS
Output level alpha fusion	28.11	0.129	24.72	0.112	0.146
This paper features level interaction	28.94	0.104	25.31	0.091	0.087

Table 10 shows that under the condition of identical routing weights, feature level interaction brings PSNR gain of 0.83 dB and 0.59 dB on L2 and L3 respectively, and boundary LPIPS decreases by 40%. This result directly proves that the benefits of decision-level fusion do not come from the routing weight itself, but from the mutual constraints of the two streams in the intermediate stage after the weight is involved in the generation process. Even if output level fusion has the same routing information, it can only compromise between already formed errors and cannot correct errors during generation.

The third term tests for the contribution of the marginal consistency fusion operator, which is replaced by fixed kernel poison fusion and simple alpha blending. Table 11 shows the boundary indices for the ablation of the fusion operator.

Table 11. Boundary region index of fusion operator ablation

Fusion operator	Boundary LPIPS	Residual chromaticity shift	Residual frequency break	The residual structure is misaligned
Simple alpha blending	0.163	5.12	0.141	8.7°
Fixed kernel Poisson	0.151	4.38	0.128	7.9°
In this paper, adaptive fusion	0.087	1.96	0.052	3.4°

Table 11 shows that the adaptive fusion in this paper reduces the boundary LPIPS pressure to 53 percent of that of simple alpha blending and 58 percent of that of fixed kernel Poisson, and the three residual discontinuities are all reduced to less than half of that of the comparison method, among which the improvement of structural dislocation is the largest. This result is consistent with the design goal of the fusion operator: the bandwidth varies point by point with the routing confidence to solve the transition band width mismatch, and the kernel shape is oriented along the structure tensor to solve the edge bluntness caused by cross-structure mixing.

The fourth item tests the style anchoring constraint, removes the three statistics separately, and reports the trade-off between the style indicator and the perception indicator. Table 12 presents the results of the style constraint ablation.

Table 12. Style and perception indicators of style anchoring constraint ablation

Constraint configuration	The directional KL divergence	Color moment distance	Spectrum slope deviation	LPIPS
No style constraints	0.412	6.83	0.197	0.142
Remove stroke direction	0.387	2.91	0.059	0.113
Remove color moments	0.152	6.47	0.057	0.116
Remove the spectrum slope	0.149	2.88	0.194	0.121
The complete triple constraint	0.148	2.87	0.058	0.104

Table 12 shows that the three constraints each dominate a statistical dimension, and the removal of any one item only makes its corresponding dimension rise significantly, while the remaining dimensions remain basically unchanged, proving that there is no redundant substitution relationship among the three constraints. The complete triple constraint reaches the lowest level in all style dimensions and the LPIPS is synchronous and optimal, indicating that style consistency and perceived quality are in the same direction rather than conflict under this framework. No single statistic can dominate the full returns, and a triple combination is necessary.

4.4 Interpretability diagnosis of routing behavior

Beyond metrics, independent evidence is needed to determine whether a routing mechanism learns understandable behavior. Figure 2 shows the corresponding relationship between the spatial distribution of routing weight and the type of breakage and the content category of cultural relics in typical breakage scenarios.

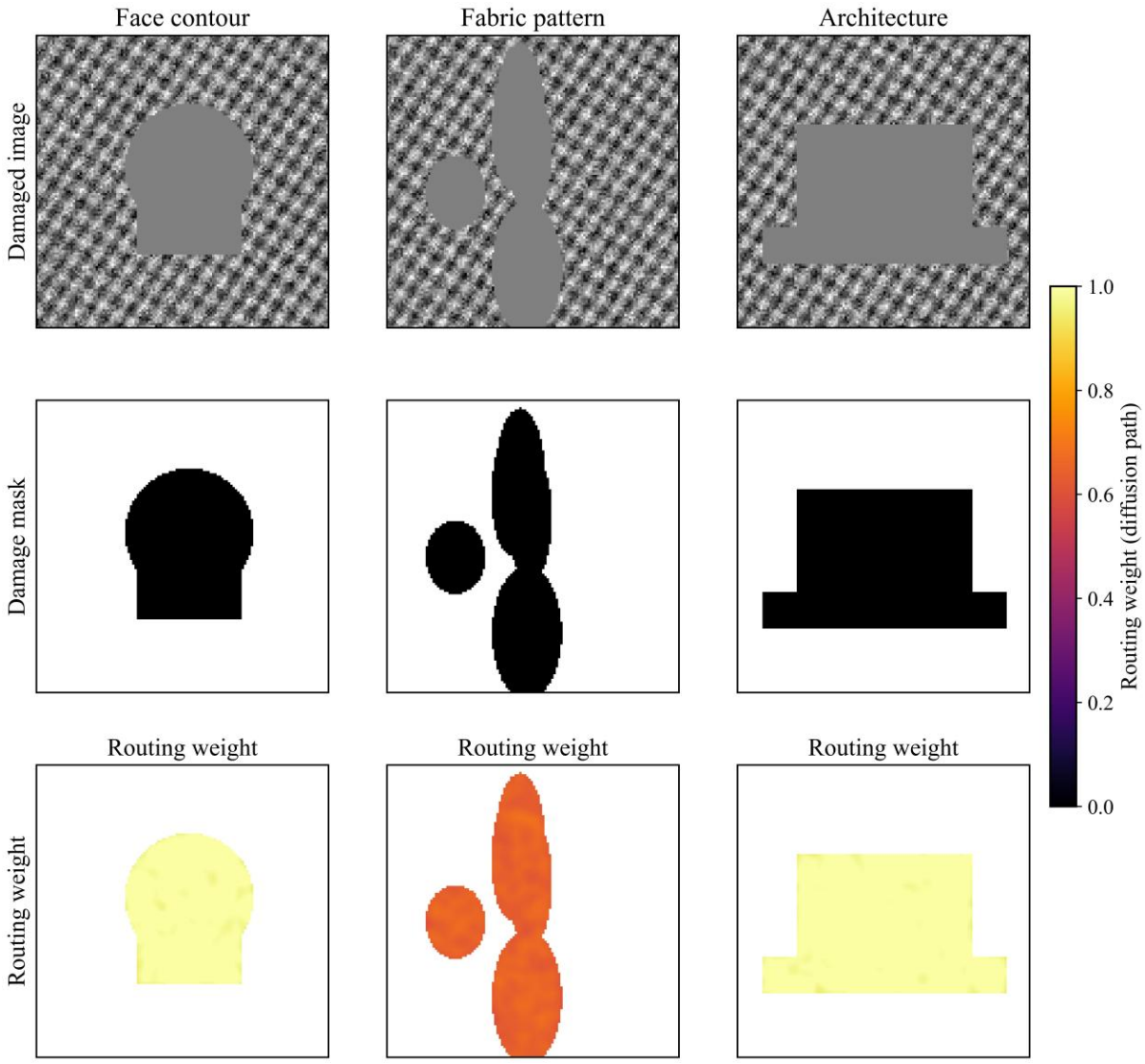


Figure 2. Spatial distribution of routing weights in typical incomplete scenarios

Figure 2 consists of three columns, each corresponding to a type of typical incomplete scene. From top to bottom, the original incomplete image, incomplete mask and routing-weight heatmap are respectively. The first column shows that the facial contour is missing, and the thermal map shows that the high diffusion weight is concentrated at the direction of the facial contour and the five sensory structures, while the texture region is dominated by the generative adversarial stream. In the second column, the repetitive pattern of the fabric is missing. The thermal map shows that the high resistance weight is concentrated in the repetitive pattern area, and the diffusion weight increases at the boundary of the pattern and the direction of the fabric fold. In the third column, the building structure is broken, and the thermal map shows that the diffusion weight is significantly dominant in the central area with a large area of continuous missing, and the opposition weight only appears in the neighborhood of the residual masonry texture. The three columns together indicate that the routing weights highly correspond to the location of the semantic structure and are not random or determined by the mask geometry alone.

To further quantify the uncertainty structure of routing behavior, routing entropy is introduced as a metric, which is defined as the spatial distribution of information entropy of per-pixel routing weights

in the missing region. Statistics show that the region with high value of routing entropy under L2 condition is concentrated near the semantic boundary, that is, the boundary band of different content categories, while the routing entropy of the semantic uniform region is significantly lower. This pattern is consistent with the design expectation: neither of the two streams has an overwhelming advantage at the semantic boundary, so the routing uncertainty should increase. When the semantics is uniform, the path ownership is clear, and the route should tend to be determined. If the routing mechanism only overfits the mask geometry, the entropy distribution should be independent of the semantic boundary, which is rejected by actual observation.

4.5 Robustness under distribution shift and failure case analysis

The stress test examines the degenerate behavior of the method outside the training distribution. The first item is masking shape deviation. Irregular block mask is used in training, and two extreme forms of slender scratch and large area continuous missing are used in test. Figure 3 shows the performance degradation under mask morphological offset.

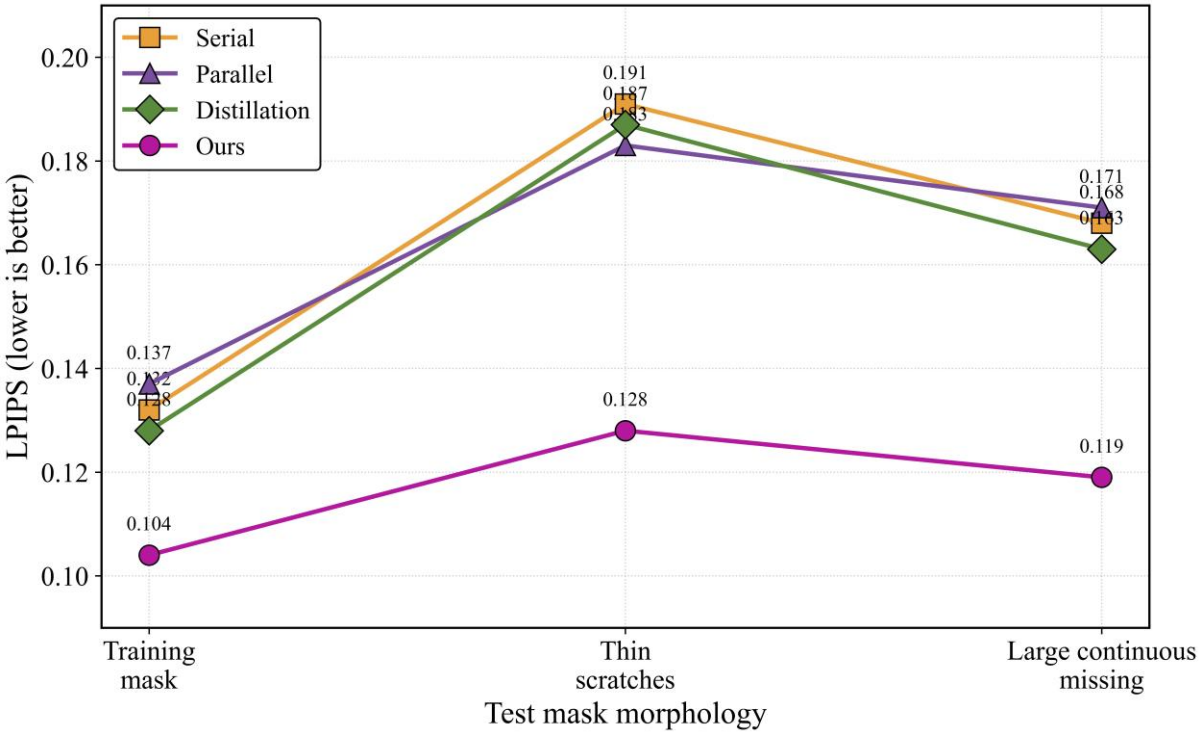


Figure 3. Layered performance attenuation under mask morphological offset

Figure 3 presents LPIPS of the proposed method and the three types of shallow fusion baselines under the three test forms of training mask, slender scratch and large area continuous missing in the form of broken lines. The horizontal axis is the test form, and the vertical axis is LPIPS. The proposed method has the lowest increase on the training mask, the lowest increase on the slender scratch, and the increase on the large area continuous missing is close to that of the pure diffusion method but still lower than all the shallow fusion baseline. The attenuation of the serial and distillation types is the most severe on slender scratches, reflecting that their fixed stage division is difficult to adapt to the mask geometry that is greatly different from the training distribution. The attenuation of the method in this paper is relatively gentle, because the input of the routing controller is local structural cues rather than mask global statistics, and the locality of cues makes it invariant to the overall change of mask morphology.

The second term is cross-material generalization, which is trained on mural data and tested with zero samples on painting cultural relics to observe the changes in routing behavior. The results show that the routing weight shifts to the diffusion stream on the painting cultural relics as a whole, which is consistent with the stronger structural continuity requirement of the painting cultural relics. However, the rise of the style distance is greater than the decline of the fidelity index, indicating that the estimation of the style anchoring constraint under the condition of cross-material is offset, which is a clear weakness of the current method.

Failures need to be reported honestly. When the missing region crosses the boundary of more than two strong texture modes, the routing controller may produce oscillatory routing, that is, the weights of adjacent pixels are repeatedly switched between the two paths, resulting in slight texture seams at the fusion boundary. Figure 4 shows a typical failure.

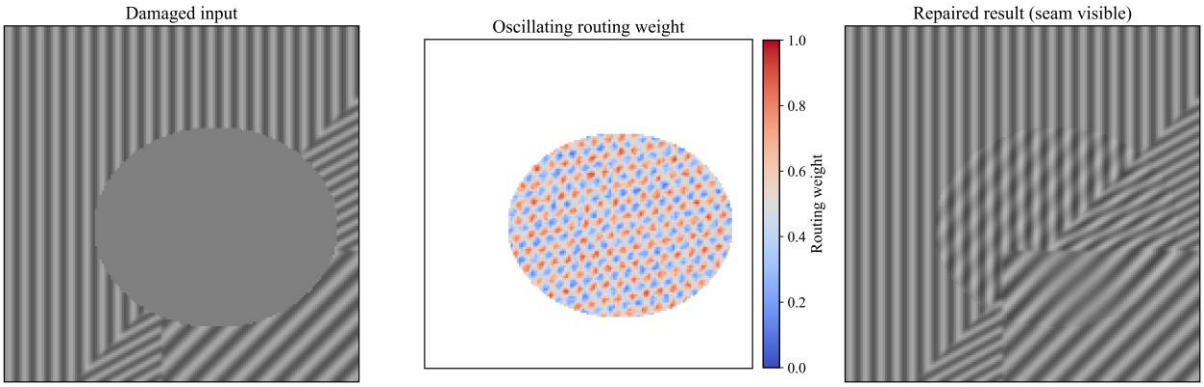


Figure 4. Oscillatory routing failure case across multi-texture pattern boundaries

The left side of Figure 4 shows the original incomplete image, the middle side shows the routing-weight heatmap, and the right side shows the local magnification of the repair result. The thermal map shows the intersection of the three texture patterns, with high-frequency alternating red and blue mottling of weights rather than smooth transitions. The enlarged image shows a slight misorientation of the texture at the seam. The reason for this failure is that the local structural cue itself has a high spatial frequency at the intersection of multiple texture modes. The controller faithfully reflects the oscillation of the cue, but the bandwidth adjustment of the fusion operator cannot fully absorb this high-frequency routing change. The localization of this problem is the direction of future work, and possible improvements include the introduction of local smooth priors in the routing controller, or the application of structural tensor-based anisotropic diffusion to the routing weights in the fusion stage.

5. CONCLUSION

This paper proposes and verifies a decision-level fusion framework for the restoration of cultural relics. The core of this framework is to transform the complementarity between the generative adversative network and the diffusion model from the image compromise at the output to the path allocation in the generation process. The two generation flows adopt deliberately asymmetric structure configuration, the generation adversarial stream adopts gated convolutional code coder to undertake sharp reconstruction of medium and high frequency texture, and the diffusion stream maintains semantic consistency of long-range structure with latent spatial conditional denoising network. The two only share incomplete mask coding and completely isolate feature representation, so as to reach the upper bound of generation ability in their advantage interval before entering the cooperation stage. The

defect-aware routing controller takes local edge density, fragment-connected domain compacting and neighborhood semantic gradient as input, and outputs pixel by-pixel path allocation weight. Its supervision signal is obtained by online mining of the perceived distance between the two-stream generation behavior and the true value, making path selection a learnable decision rather than a preset heuristic rule. The routing weight modulated the cross-attention intensity of the diffusion stream to the texture features of the generative adversarial stream by multiplicative gating and injected the low-level structural prior of the diffusion stream into the generative adversarial stream in an asymmetric manner, so that the two streams formed bidirectional and spatially variable coupling in the intermediate stage of denoising and decoding. The adaptive fusion operator guided by routing gradient generates the fusion bandwidth point by point according to the weight gradient tensor and constructs the anisotropic kernel along the main direction of the local structure tensor, so that the transition band is continuous in the three dimensions of chrominance, texture frequency and structure orientation. The triple style anchoring constraint estimates the stroke direction histogram, pigment aging color moment and texture energy spectrum slope in real time from known intact areas, which are used as dynamic regularities in the generation process to inhibit texture repetition and over creation. Hierarchical experiments show that the proposed framework approaches the optimal performance of pure generative adversarial method and pure diffusion method respectively under the condition of micro-loss and heavy-loss and is significantly better than the three shallow fusion baselines of serial, parallel and distillation in the transition interval from medium loss to heavy-loss. The ablation experiments confirm the independent contribution of routing mechanism and feature-level interaction. The disaggregated results of the boundary and style indicators further support the design validity of the fusion operator and style constraints.

The proposed framework does not introduce a new generation paradigm, and the backbone structure, training objectives and basic generation mechanism of the two generation flows are all derived from the existing achievements of their respective paradigms. The promotion of this paper focuses on the collaborative level, that is, to determine the granularity, stage and way of interaction of the two generation paths with clear division of labor. This definition determines the conditionality of the benefits. When the ability of the single-flow model in its advantage interval continues to improve, the value interval of the routing mechanism moves accordingly, and its relative gain may narrow. When the semantic complexity of the missing region exceeds the respective modeling ability of the two streams, the route can only make a suboptimal choice between the two streams but cannot generate content that is not available to both flows. The effectiveness of style anchoring constraints also depends on the representability of known regions to the overall style of cultural relics. These bounds are made explicit so that the contribution of the framework can be precisely tested and falsified, rather than masking its applicability with blanket statements of superiority.

The current preference mining of routing controllers relies on pixel by-pixel truth supervision, that is, preference labels are constructed by the perceived distance between the results generated separately by two streams and the truth value. This design is effective under the condition of incomplete synthesis or sufficient paired data, while the restoration of real cultural relics faces the fundamental difficulty of having no true value. The historical original appearance of the missing area is unknown, and any single reference image is only one of many reasonable speculations, so the preference label cannot be directly constructed. One possible direction is to turn preference mining from point estimation to ranking learning, in which domain experts give relative preferences rather than absolute truth values for multiple candidate generation results at the same location, and the controller learns path allocation under the ranking constraints, so that expert knowledge is injected into routing decisions in a weakly supervised

form. Another possible direction is to develop reference-free route learning, using the consistency between two flow generation results as a self-supervised signal to improve route uncertainty and trigger a more conservative fusion strategy in areas with low consistency. The feasibility of the two directions depends on the stability of the sorting signal and the discriminant power of the consistency measure respectively, which need to be verified by independent research.

After the expansion of the generative approach to the restoration of cultural relics with model scale and data scale as the core, its further progress will increasingly depend on the fine-grained control of the generative behavior, rather than simply pursuing a larger unified model. In this paper, the differences between different generation paradigms are regarded as resources that can be spatialized, allocated point by point and intervened by process, rather than defects that need to be erased or replaced. The essence of this perspective is to transform the tension between texture and structure, local and global, sharpness and fidelity from the obstacle of fusion to the material of fusion. Its validity has been preliminarily tested in the hierarchical evidence of the transition interval from medium loss to heavy loss, and its boundaries and undone points have also been clearly demarcated in the above discussion.

Abbreviations

GAN, Generative Adversarial Network;

PSNR, Peak Signal-to-Noise Ratio.

Supplementary Material

Not applicable.

Appendix

Not applicable.

Ethics approval and consent to participate.

This study did not involve human participants, animal subjects, or any data requiring ethical approval. Therefore, ethics approval and consent to participate are not applicable.

Acknowledgements

The authors would like to thank the editors of this journal and all the anonymous reviewers who provided valuable comments on this work.

Competing interests

The authors declare that they have no financial or personal relationships that may have inappropriately influenced them in writing this article.

Author contributions

All authors have read and agreed to the published version of the manuscript. The author's contributions are specified as follows: **M.M.:** Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Resources, Data Curation, Writing – original draft, Writing – review & editing, Visualization, Supervision, Project administration.

Funding information

The authors declare that no funds, grants, or other support were received during the preparation of this manuscript.

Data availability

The data that support the findings of this study are available upon request from the corresponding authors, **M.M.**

Disclaimer

The views and opinions expressed in this article are those of the authors and are the product of professional research. It does not necessarily reflect the official policy or position of any affiliated institution, funder, agency, or that of the publisher. The authors are responsible for this article's results, findings, and content.

Declaration of AI and AI-assisted Technologies in the Writing Process

During the writing of this article, the author used ChatGPT for spelling and grammar checking. After using this tool, the author reviewed and edited the content as needed and assumes full responsibility for the final published content.

REFERENCES

[1] Dong, G., Huang, W., Smith, W. A., & Ren, P. (2020). A shadow constrained conditional generative adversarial net for SRTM data restoration. *Remote Sensing of Environment*, *237*, 111602. https://doi.org/10.1016/j.rse.2019.111602

[2] Zhao, T., Chen, G., Pang, C., Seenoi, P., Papukdee, N., Busababodhin, P., & Du, Y. (2025). Hybrid convolutional neural network-graph attention network-gradient boosting decision tree model for seismic impedance inversion prediction. *Journal of Seismic Exploration*, *34*(5), 81. https://doi.org/10.36922/JSE025310051

[3] Saxena, D., & Cao, J. (2021). Generative adversarial networks (GANs) challenges, solutions, and future directions. *ACM Computing Surveys (CSUR)*, *54*(3), 1-42. https://doi.org/10.1145/3446374

[4] Deng, Z., Zhang, H., Liang, X., Yang, L., Xu, S., Zhu, J., & Xing, E. P. (2017). Structured generative adversarial networks. *Advances in neural information processing systems*, *30*. https://doi.org/10.48550/arXiv.1711.00889

[5] Wang, C., Xu, C., Yao, X., & Tao, D. (2019). Evolutionary generative adversarial networks. *IEEE Transactions on Evolutionary Computation*, *23*(6), 921-934. https://doi.org/10.1109/tevc.2019.2895748

[6] Li, R., Chen, L., Zhang, Z., Jampani, V., Patel, V. M., & Zhang, L. (2025, April). Syncnoise: Geometrically consistent noise prediction for instruction-based 3d editing. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 39, No. 5, pp. 4905-4913). https://doi.org/10.1609/aaai.v39i5.32519

[7] Liu, S., Yu, C., Cao, C., Qian, W., & Wang, F. (2024, September). VCD-Texture: Variance alignment based 3D-2D co-denoising for text-guided texturing. In *European Conference on Computer Vision* (pp. 373-389). Cham: Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-72640-8_21

[8] Zhao, T., Chen, G., Pang, C., Seenoi, P., Papukdee, N., & Busababodhin, P. (2025). Time-lapse earthquake difference prediction based on physics-informed long short-term memory coupled with interpretability boosting. *Journal of Seismic Exploration*, *34*(3), 25. https://doi.org/10.36922/JSE025310049

[9] Zheng, H., Yao, Y., Zhou, Y., Luo, J., & Lin, Z. (2026). Pixperfect: Seamless latent diffusion local editing with discriminative pixel-space refinement. *Advances in Neural Information Processing Systems*, *38*, 128266-128292. https://doi.org/10.48550/arXiv.2512.03247

[10] Xu, L., Luo, Q., Chen, L., Ni, T., Lin, Q., Zhao, S., ... & Feng, H. (2026). From Adversarial Competition to Iterative Refinement: A Critical Review of Generative AI for Remote Sensing Scene Classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*. https://doi.org/10.1109/jstars.2026.3670966

[11] Li, X., Ren, Y., Jin, X., Lan, C., Wang, X., Zeng, W., ... & Chen, Z. (2025). Diffusion models for image restoration and enhancement: A comprehensive survey. *International Journal of Computer Vision*, *133*(11), 8078-8108. https://doi.org/10.1007/s11263-025-02570-9

[12] Ding, X., Chen, X., Sui, Y., Wang, Y., & Zhang, J. (2025). Underwater image enhancement using a diffusion model with adversarial learning. *Journal of Imaging*, *11*(7), 212. https://doi.org/10.3390/jimaging11070212

[13] Lu, Z., Li, B., Fu, C., Wu, J., Xu, L., Jia, S., & Zhang, H. (2024, October). Remaining useful life prediction method based on dual-path interaction network with multiscale feature fusion and dynamic weight adaptation. In *Actuators* (Vol. 13, No. 10, p. 413). MDPI. https://doi.org/10.3390/act13100413

[14] Zhao, J., Liu, T., & Sun, L. (2025). Multi-scale fusion maxvit for medical image classification with hyperparameter optimization using super beluga whale optimization. *Electronics*, *14*(5), 912. https://doi.org/10.3390/electronics14050912

[15] Ye, F., & Bors, A. G. (2022). Dynamic self-supervised teacher-student network learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *45*(5), 5731-5748. https://doi.org/10.1109/tpami.2022.3220928

[16] Wang, Y., & Wang, X. (2024). Artificial intelligence in physical education: comprehensive review and future teacher training strategies. *Frontiers in public health*, *12*, 1484848. https://doi.org/10.3389/fpubh.2024.1484848

[17] Liu, M., & He, Q. (2026). Relation-Aware and Constraint-Guided Discrete Diffusion for Design Color Scheme Generation. *Journal of Discovery Core*, *1*(2), 80-109. https://doi.org/10.67541/jdc2610

[18] Liu, L. (2026). Research on Intelligent Generation Algorithm of Interface Icon Based on Diffusion Model. *International Scientific Technical and Economic Research*, *4*(1), 149-167. https://doi.org/10.71451/ISTAER2607

[19] Du, J., & Liu, M. (2026). Intelligent Defect Detection of Design Materials Based on Edge AI. *Journal of Digital Frontier*, *1*(2), 27-57. https://doi.org/10.67541/jdf2609

[20] Ma, X., & Wu, H. (2026). Construction and Real-Time Status Monitoring of a Changxin Palace Lantern Digital Twin Based on Multi-Mode Sensing Data Fusion. *Journal of Digital Frontier*, *1*(1), 185-218. https://doi.org/10.67541/jdf2607

[21] Liu, Y., Feng, S., & Liu, M. (2026). Research on Transformer-Based Action Sequence Modeling of Intangible Cultural Heritage Shadow Play Using Attention Mechanisms. *International Scientific Technical and Economic Research*, *4*(2), 51-77. https://doi.org/10.71451/ISTAER2615

[22] Mouadden, N., Laousy, O., Marini, R., Ong, V., Revel, M. P., Chassagnon, G., ... & Vakalopoulou, M. (2025, September). Conditional Latent Diffusion Models for Irregularly Spaced Longitudinal Radiological Data. In *International Conference on Medical Image Computing and Computer-Assisted Intervention* (pp. 106-115). Cham: Springer Nature Switzerland. https://doi.org/10.1007/978-3-032-05182-0_11

[23] Hussain, M., O'Nils, M., Lundgren, J., & Mousavirad, S. J. (2024). A comprehensive review on deep learning-based data fusion. *IEEE Access*, *12*, 180093-180124. https://doi.org/10.1109/access.2024.3508271

[24] Almutairi, A., & Owais, M. (2025). Reliable vehicle routing problem using traffic sensors augmented information. *Sensors*, *25*(7), 2262. https://doi.org/10.3390/s25072262

[25] Wang, J., Zhang, L., Wang, H., Zhang, C., Su, H., Duan, J., & Hu, J. (2026). Multi-objective collaborative optimization of conical pick group based on factor space theory. *Scientific Reports*. https://doi.org/10.1038/s41598-026-64321-1

[26] Zhang, W., Zhao, Y., Breckon, T. P., & Chen, L. (2017). Noise robust image edge detection based upon the automatic anisotropic Gaussian kernels. *Pattern Recognition*, *63*, 193-205. https://doi.org/10.1016/j.patcog.2016.10.008

[27] Xiao, C., Staring, M., Wang, Y., Shamonin, D. P., & Stoel, B. C. (2012). Multiscale bi-Gaussian filter for adjacent curvilinear structures detection with application to vasculature images. *IEEE Transactions on Image Processing*, *22*(1), 174-188. https://doi.org/10.1109/tip.2012.2216277

[28] Qian, J., Yang, J., & Gao, G. (2013). Discriminative histograms of local dominant orientation (D-HLDO) for biometric image feature extraction. *Pattern Recognition*, *46*(10), 2724-2739. https://doi.org/10.1016/j.patcog.2013.03.005

[29] Shu, C., Ding, X., & Fang, C. (2011). Histogram of the oriented gradient for face recognition. *Tsinghua Science and Technology*, *16*(2), 216-224. https://doi.org/10.1016/s1007-0214(11)70032-3

[30] Robitaille, T. P. (2017). A modular set of synthetic spectral energy distributions for young stellar objects. *Astronomy & Astrophysics*, *600*, A11. https://doi.org/10.1051/0004-6361/201425486

[31] Yu, R., Chen, W., & Zhou, D. (2020). Infrared and visible image fusion based on gradient transfer optimization model. *IEEE Access*, *8*, 50091-50106. https://doi.org/10.1109/access.2020.2979760

[32] Zhang, H., Xu, H., Xiao, Y., Guo, X., & Ma, J. (2020, April). Rethinking the image fusion: A fast unified image fusion network based on proportional maintenance of gradient and intensity. In *Proceedings of the AAAI conference on artificial intelligence* (Vol. 34, No. 07, pp. 12797-12804). https://doi.org/10.1609/aaai.v34i07.6975

APA Citation

Ma, M. (2026). Research on Restoration and Completion Algorithm of Cultural Relics Based on Generative Adversarial Network and Diffusion Model. *Journal of Discovery Core, 1*(3), 1-28. **https://doi.org/10.67541/jdc2613**