Digital Intelligence PRESS

Journal of Digital Frontier

https://doi.org/10.67541/jdf2607

Research Article Volume 1, Issue 1 (July 2026) Open Access

Construction and Real-Time Status Monitoring of a Changxin Palace Lantern Digital Twin Based on Multi-Mode Sensing Data Fusion

Xinzi Ma, Huixia Wu *

Faculty of Art and Design, Beijing City University, Beijing, China

***Corresponding author:** Huixia Wu. Email: art1993@163.com

Abstract

Aiming at the problems of environmental disturbance, hidden structural response, and difficult collaborative fusion of multi-source monitoring information in the long-term preservation process of mobile cultural relics, this paper proposes a digital twin construction and real-time status monitoring method for the Changxin Palace Lantern based on multi-mode perception data fusion. This paper constructs a multi-source sensing system integrating 3D structured light, spectral imaging, FBG strain sensing and MEMS vibration monitoring, and realizes the unified representation of geometric state, material characteristics and dynamic response of cultural relics through asynchronous spatio-temporal registration and a physical-constraint-driven spatio-temporal fusion network (ATS-FusionNet). On this basis, the residual-driven state deviation index (RSDI) is established to realize the location of abnormal areas and the identification of degradation risks. The experiment is carried out based on the joint verification platform of Changxin Palace Lantern. The results show that the mean absolute errors of temperature prediction, strain prediction, and vibration amplitude prediction in the multi-mode state field reconstruction is 0.31 °C, 2.7με and 0.012g, respectively. The end-to-end response delay of the system is 43.6ms on average, meeting the requirements of real-time monitoring. The research shows that the proposed method can effectively break through the limitations of insufficient single-mode perception and lack of physical correlation in the traditional cultural relic monitoring and provide a high-precision and real-time digital twin technology path for the intelligent protection and preventive maintenance of complex mobile cultural relics.

Keywords

Digital twin; Multi-mode perception fusion; Protection of cultural heritage; Condition monitoring; Detection of anomalies

Received: 06 Jun 2026; Revised: 17 Jul 2026; Accepted: 29 Jul 2026; Published: 24 Aug 2026

1. INTRODUCTION

China possesses an extensive and diverse collection of precious cultural relics, among which numerous bronze, metal, and composite-material artifacts hold irreplaceable historical, artistic, and scientific value. As a representative bronze artifact of the Han Dynasty, the Changxin Palace Lantern is distinguished not only by its exceptional cultural significance but also by the heightened demands it places on long-term preservation and safe exhibition, owing to its complex structure, sophisticated decorative techniques, and multi-material composite characteristics. As a movable cultural relic, the Changxin Palace Lantern is susceptible to various external factors during exhibition, transportation, restoration, and environmental fluctuations. Temperature and humidity variations, lighting effects, and airborne pollutants within exhibition halls can continuously affect the stability of the artifact's surface materials, while mechanical disturbances, micro-vibrations, and localized stress changes during handling may induce structural fatigue, loosening of connective components, and even irreversible damage. Given that the degradation process of cultural relics is typically long-term, progressive, and concealed, early-stage changes are often difficult to detect directly through manual inspection or conventional monitoring approaches. Consequently, establishing an intelligent protection technology system capable of continuous perception, dynamic analysis, and early warning has become an essential requirement for the preventive conservation of cultural relics.

Currently, museum cultural relic monitoring predominantly relies on temperature and humidity sensors, vibration monitoring equipment, and periodic manual inspections [1],[2]. While such methods can capture variations in local environmental conditions or individual physical quantities, they struggle to reflect the overall structural state of artifacts and their internal complex evolutionary processes [3]. For instance, temperature and humidity monitoring can only describe environmental trends without further assessing material responses; vibration sensing can record external disturbances but faces difficulties in locating anomalous regions and analyzing underlying influence mechanisms. Concurrently, traditional detection methods are typically based on offline analysis, with limited effective correlation among data sources, making it challenging to establish a continuous cognitive chain from environmental changes and structural responses to damage risks [4]. Therefore, in addressing the challenges of "invisible states, unpredictable evolution, and untraceable risks" in the protection of precious movable cultural relics, it is imperative to introduce novel intelligent perception paradigms that integrate multi-source monitoring data, three-dimensional spatial models, and physical state analysis to achieve dynamic mapping and real-time assessment of cultural relic conditions [5],[6].

The advancement of digital twin technology offers a promising technical pathway for addressing the aforementioned challenges. Digital twins enable real-time mapping, state updating, and behavior prediction of physical objects in virtual space by establishing data connections between physical entities and virtual models [7],[8],[9],[10]. In recent years, this technology has been widely adopted in complex engineering domains such as aerospace equipment, intelligent manufacturing, and energy systems, achieving equipment health management and fault prediction through the integration of sensor data, model computations, and intelligent analytics [11],[12],[13]. However, compared with industrial equipment, cultural artifacts are characterized by complex structural morphologies, delicate surface textures, multi-material coupling, and non-repeatable testing conditions, rendering their digital twin construction considerably more challenging. Consequently, for high-value movable cultural relics such as the Changxin Palace Lantern, establishing a digital twin system oriented toward multi-modal perception fusion can not only enhance condition monitoring capabilities but also provide critical technical support for the future development of smart museums and preventive cultural relic conservation.

In recent years, digital twin technology has developed rapidly in the field of state awareness and intelligent operation and maintenance of complex physical objects. In the industrial domain, researchers have achieved real-time monitoring and predictive maintenance of equipment operational status through the integration of 3D modeling, sensor networks, artificial intelligence, and physical simulation methods [14],[15]. For example, in aeroengines, rail transit equipment, and large-scale mechanical systems, digital twin models can dynamically adjust the state of virtual models based on collected real-time data and predict potential failures using historical data [16],[17],[18]. However, most existing digital twin research focuses on engineering equipment with well-defined structural rules and operational mechanisms, with core objectives typically centered on performance optimization and fault diagnosis. For cultural artifacts, digital twins must not only express geometric forms but also further associate material properties, environmental effects, and long-term degradation processes, rendering traditional industrial digital twin frameworks difficult to apply directly.

In the domain of digital cultural heritage preservation, 3D scanning, photogrammetry, and virtual display technologies have been widely employed, with numerous studies achieving the construction and visualization of digital models of cultural relics [19],[20],[21]. Nevertheless, existing research predominantly emphasizes static digital preservation, that is, recording the appearance information of cultural relics through high-precision geometric reconstruction, while paying insufficient attention to the dynamic processes of state evolution over time. Particularly for artifacts such as the Changxin Palace Lantern, which feature complex surface structures, multi-region material variations, and subtle damage evolution characteristics, geometric models alone are inadequate for describing real state changes [22]. The digital twin of cultural relics must further advance the connection between geometric information and environmental, structural, and material responses, thereby enabling the transformation from “form replication” to “state cognition.”

The development of multi-modal sensing technologies presents new possibilities for the dynamic monitoring of cultural relics. By fusing multiple data types, including 3D structured light, spectral imaging, strain sensing, and vibration monitoring, information regarding surface conditions, material characteristics, and structural responses can be obtained from different dimensions [23]. However, existing multi-modal fusion methods still exhibit certain limitations. On the one hand, many studies adopt data-level or simplistic feature-level fusion, directly concatenating or weighting different sensor data while neglecting variations in sampling frequencies, spatial coordinates, and physical response characteristics across modalities, resulting in fusion outcomes that fail to accurately reflect the actual state. On the other hand, different sensors typically operate at varying temporal scales and spatial resolutions during cultural relic monitoring. Achieving precise alignment of asynchronous data and completing stable fusion under physical law constraints represents a critical challenge in constructing highly reliable digital twins.

Regarding state assessment and anomaly detection, traditional cultural relic monitoring methods predominantly rely on preset threshold-based judgments, triggering alarms when temperature, humidity, or vibration amplitude exceeds established ranges [24],[25]. Although simple and effective, such methods can only identify dominant anomalies without explaining the spatial location, formation causes, or developmental trends of anomalies. For cultural relics, early damage often manifests as localized minor changes, the characteristics of which may be concealed within the coupling processes of multiple physical fields. Consequently, relying solely on single indicators or shallow classification models proves inadequate for meeting refined protection requirements. Future state monitoring should fully exploit the spatial gradient information, physical response variations, and model prediction residuals embedded within digital twins, achieving anomaly localization, damage trend assessment, and risk tracing through the analysis of deviations between “actual states” and “theoretical states.”

Addressing the challenges of insufficient multi-source information fusion, limited state perception capabilities, and the lack of physical interpretability in anomaly identification during the digital protection of the Changxin Palace Lantern, this paper proposes a digital twin construction and real-time condition monitoring method based on multi-modal perception data fusion. Unlike traditional digital approaches that rely solely on 3D models, this paper constructs a three-tier digital twin framework comprising "geometric twin, perception twin, and judgment twin." Building upon a high-precision spatial model, this framework achieves dynamic updating of cultural relic states through the fusion of multiple sensor data types and further combines intelligent analytical methods to accomplish state assessment and risk identification, thereby transforming the digital twin from a static display model into a dynamic system equipped with perception, analysis, and decision-making capabilities [26]. The main contributions of this paper are as follows: (1) A non-uniform adaptive sensor placement strategy is proposed to address the optimized configuration of perception resources for complex curved-surface cultural relics; (2) An Asynchronous Temporal-Spatial Fusion Network (ATS-FusionNet) is designed to achieve deep fusion of heterogeneous data through physical consistency constraints; (3) A Residual-driven State Deviation Index (RSDI) is proposed to enable anomaly localization and degradation risk identification. Experimental results demonstrate that the proposed method achieves significant improvements in state reconstruction accuracy, anomaly recognition capability, and real-time response performance, providing a feasible technical pathway for the intelligent protection of complex and precious cultural relics.

2. MULTI-MODE PERCEPTION SYSTEM AND TWIN DATA BASE

As the peak of Han Dynasty bronze casting technology, Changxin Palace Lantern has its unique triple structural characteristics of “hollow-body casting molding, gold decoration on internal and external surfaces, and mortise-tenon and tenon fitting of removable parts”, which brings fundamental challenges to state perception. The temperature gradient and strain transfer path of the inner and outer walls of the hollow structure was different from those of the solid body. The difference in thermal expansion coefficient between the gilt surface layer and the copper substrate (16.7×10^{-6}/°C for copper and 14.2×10^{-6}/°C for gilt layer) caused the shear hysteresis effect at the interface, and the dismountable connection between the lamp plate, the lamp body and the base introduced the contact nonlinear response. These characteristics determine that the sensing of a single physical quantity cannot represent the overall health state of cultural relics, and the information must be obtained simultaneously in three dimensions of geometric morphology, surface material properties and internal mechanical response, and the physical redundancy between each mode should be mutually verified. After systematically sorting out the norms for preventive protection of cultural relics of the Palace Museum and relevant technical documents of the International Association of Museums (ICOM), and combining with the pre-experiment response spectrum analysis of Changxin Palace Lantern copies in shaking table and temperature and humidity environment box, this paper finally identifies four types of core perceptual modes: 3D structured light is used to obtain high-density geometric point clouds (spatial resolution up to 0.05 mm), multi-spectral imaging (covering three bands of visible light, near infrared and short wavelength infrared) is used to identify surface reflectance spectrum and fading area, and distributed fiber Bragg grating (FBG) strain/temperature sensor is embedded in key sections to obtain internal mechanical response [27],[28]. The MEMS accelerometer array collects the overall vibration mode and tilt angle. The time sampling rates of the four types of sensors are 2 Hz, 5 Hz, 1000 Hz (FBG dynamic strain) and 200 Hz respectively, which are obviously different, which puts forward a direct constraint on the subsequent asynchronous fusion.

The location of the sensor determines the upper limit of the data quality, and the complex surface curvature distribution and the reflection characteristics of the gold layer make the equi-spaced uniform layout strategy extremely uneconomical. In this paper, a non-uniform adaptive point placement algorithm based on joint optimization of surface curvature and illumination accessibility is proposed. Let the triangular mesh model obtained by preliminary scanning of 3D structured light be $\mathcal{M} = (\mathcal{V}, \mathcal{E}, \mathcal{F})$, where $\mathcal{V}$ is the vertex set, $\mathcal{E}$ is the edge set, and $\mathcal{F}$ is the facet set. For any vertex $v_i \in \mathcal{V}$, its discrete mean curvature κ_i is calculated by cotangent weighting:

$$\kappa_i = \frac{1}{4A_i}\left\| \sum_{j\in\mathcal{N}(i)} (\cot\alpha_{ij} + \cot\beta_{ij})(v_i - v_j) \right\| \tag{1}$$

Where $\mathcal{N}(i)$ is a ring of neighborhood vertices of vertex v_i, α_{ij} and β_{ij} are the diagonal angles of the two diagonal triangles where edge (v_i, v_j) is located, and A_i is the area of Voronoi region. Larger curvature indicates richer geometric detail and requires higher sensing density. At the same time, the illumination accessibility factor ρ_i is defined to represent the visibility and reflection intensity of the vertex under the projection of structured light:

$$\rho_i = \frac{1}{N_s}\sum_{s=1}^{N_s} \frac{n_i \cdot l_s}{d_{is}^2 + \epsilon} \cdot \mathbb{I}\big(vis(v_i, s)\big) \tag{2}$$

Where N_s is the total number of projected directions of structured light, n_i is the unit normal vector of vertex v_i, l_s is the unit vector of the s-th projected direction, d_{is} the distance from the light source to the vertex, ϵ is a small positive number to prevent the denominator from being zero, and $\mathbb{I}(\cdot)$ is an indicator function to indicate whether the vertex is visible in this direction. Integrating curvature and reachability, we define the comprehensive distribution priority P_i of the i-th vertex:

$$P_i = \lambda_\kappa \cdot \tilde{\kappa}_i + \lambda_\rho \cdot (1 - \tilde{\rho}_i) \tag{3}$$

Where $\tilde{\kappa}_i$ and $\tilde{\rho}_i$ are the normalized curvature and accessibility factors, respectively, and λ_κ and λ_ρ are the weight coefficients (0.6 and 0.4 are taken in this paper, which are determined in the pre-experiment through grid search). The actual sensor placement position is determined by non-maximum suppression clustering on P_i, that is, only the vertex with the highest priority is reserved as the candidate placement point within a local neighborhood radius r. As shown in Table 1, on the premise that the average point density of the key area (the lamp core, the maximum section of the lamp body abdomen, and the transition section connecting the base and the lamp body) meets the benchmark requirements, the total number of global sensors is reduced by about 32% compared with the uniform layout, and the curvature coverage of the key area (defined as the ratio of the average curvature at the layout point to the average curvature of the whole surface) is increased by about 41%.

Table 1. Comparison of key parameters between non-uniform adaptive location and uniform location strategy

Dimensions of assessment	Uniform point distribution strategy	Non-uniform adaptive point placement strategy	Change in relative
Total number of sensors (including four categories)	142	96	-32.4%
Point density in key areas (point /m²)	18.6	31.2	+67.7%
Average normalized curvature at the distribution point	0.43	0.61	+41.9%
Proportion of low light area spots	22.5%	8.3%	-63.1%
Proportion of geometric reconstruction blind area	12.7%	5.2%	-59.1%

The unification of time benchmark of multi-source heterogeneous data is the premise of subsequent fusion. The clock source of the four types of sensors is independent, the trigger mechanism is different, and the sampling frequency is different. Direct splicing will introduce significant time misalignment errors. This paper adopts a two-layer synchronization scheme with hardware synchronization as the main and software compensation as the auxiliary. At the hardware level, all sensors use the high-precision second pulse (PPS) signal provided by the GPS receiver as the global time reference, and the IEEE 1588 Precision Time Protocol (PTP) is used to realize sub-microsecond clock synchronization in the network to ensure that each sampled data point is equipped with a unified timestamp t_k (with GPS time as reference) [29]. Let the original observation value of the m-th sensor at the k-th sampling time be $x_m(t_k)$. Since the actual sampling time $t_k^{(m)}$ of each sensor does not completely coincide with the unified time grid τ_n ($\Delta\tau = 0.01\ s$ is taken as the basic time grid in this paper), the time registration from asynchronous to synchronous should be carried out. Define time deviation:

$$\delta_k^{(m)} = \tau_n - t_k^{(m)} \tag{4}$$

When $\left|\delta_k^{(m)}\right| < \Delta\tau/2$, it is directly matched; otherwise, cubic spline interpolation is used to compensate:

$$x_m(\tau_n) = \sum_{k=0}^{K} c_{m,k} \cdot B_{k,3}(\tau_n) \tag{5}$$

Where $B_{k,3}(\tau_n)$ is the cubic B -spline basis function and $c_{m,k}$ is the fitting coefficient, which is solved by least squares. The synchronization accuracy is evaluated by the maximum time alignment error $\varepsilon_{\text{sync}}$, and the measured $\varepsilon_{\text{sync}}$ of all modes is controlled within 2 ms, which meets the time accuracy requirements of subsequent fusion.

The spatial matching criterion needs to unify the structural light point cloud coordinate system $\{S\}$, multispectral image pixel coordinate system $\{I\}$, FBG strain measurement point coordinate system $\{F\}$ and MEMS local coordinate system $\{M\}$ to the world coordinate system $\{W\}$ with the geometric center of Changxin Palace Lantern as the origin. Firstly, the external parameter matrix $[R_{WS}, t_{WS}]$ of the structured light scanner is calibrated by the high-precision optical tracking system (OptiTrack, accuracy 0.1mm) to realize the transformation from $\{S\}$ to $\{W\}$. For the multispectral camera, the internal reference matrix K and distortion parameters are obtained by Zhang Zhengyou calibration method, and then the rotation matrix R_{SI} and translation vector t_{SI} are obtained by hand-eye calibration (using checkerboard targets in the common field of view of the camera and structured light), so as to establish the mapping between image pixel coordinates (u, v) and world coordinates (X_W, Y_W, Z_W) [30],[31]:

$$\begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = K \cdot [R_{WS}R_{SI} \mid t_{WS} + R_{WS}t_{SI}] \cdot \begin{bmatrix} X_W \\ Y_W \\ Z_W \\ 1 \end{bmatrix} \tag{6}$$

The spatial position of the FBG strain measurement point is read directly by a three-dimensional coordinate measuring machine (CMM), and its transformation with $\{W\}$ is obtained by least squares registration of three non-collinear reference points. The mounting posture of the MEMS accelerometer is calibrated by the quaternion method. On this basis, the Digital Reference Grid (DRG) method is introduced for global consistency check: 24 ceramic reflective spheres with a diameter of 3 mm are arranged around the Changxin Palace Lantern, and their spatial coordinates $q_j^{(W)}$ are accurately determined by the laser tracker. Each sensor completes the final spatial coordinate alignment by identifying these reference spheres, so that the data of all modes can be in the same spatial grid under the world coordinate system:

$$G = \{(X_g^m, Y_g^m, Z_g)\} \tag{7}$$

After registration, RMS of spatial mapping error between modes is controlled within 0.15mm. After obtaining the original sensor data, there are different degrees of noise pollution and data missing in all modes, so targeted quality enhancement must be carried out. In this paper, a curvature-guided local Poisson reconstruction filling method is proposed for the point cloud holes caused by the reflection of 3D structured light on the gilt surface. The traditional Poisson reconstruction is solved with global equal weight, which is easy to be overly smooth and lose fine geometric features in the curvature mutation region. Let the set of boundary vertices in the void region to be filled be ∂H and the set of internal vertices to be estimated be H_{in}. For the indicator function χ of the void region, the standard form of Poisson equation is:

$$\nabla^2 \chi = \nabla \cdot v \tag{8}$$

Where v is the directed normal field at the boundary. In this paper, the curvature weight factor $\omega(p)$ is introduced into the energy functional:

$$E(\chi) = \int_{\Omega} \omega(p) \parallel \nabla\chi(p) - v(p) \parallel^2 \, dp + \mu \int_{\Omega} \parallel \nabla^2\chi(p) \parallel^2 \, dp \tag{9}$$

Where Ω is the void region and its neighborhood space, p is the spatial coordinate vector, μ is the smooth regularization coefficient (0.01), and the curvature weight is defined as:

$$\omega(p) = exp\left(-\alpha \cdot \frac{\kappa(p)}{\bar{\kappa}}\right) \tag{10}$$

Where $\kappa(p)$ is the estimated curvature of adjacent vertices, $\bar{\kappa}$ is the average curvature of the full surface, and α is the regulator (take 2.5). This weight makes the Poisson equation fit the data items more strictly in the high curvature region and keeps the sharp feature. In the low curvature flat region, more smoothness is allowed and numerical oscillations are suppressed. Compared with the high-precision reference model, the reconstruction error of the cavity region is reduced by about 38% compared with the standard Poisson method.

The enhancement of multispectral image is mainly aimed at reflectance attenuation and regional fading caused by long-term exposure of gold gilt layer. First, reflectivity normalization is carried out: DN value $I_b(u,v)$ of pixel (u,v) in the b-band image is converted to reflectivity $R_b(u,v)$ with standard whiteboard (Spectralon, reflectivity 99%) as reference:

$$R_b(u,v) = \frac{I_b(u,v) - I_{dark,b}}{I_{ref,b}(u,v) - I_{dark,b}} \cdot R_{ref,b} \tag{11}$$

$I_{\mathrm{dark},b}$ as the dark field noise, $I_{\mathrm{ref},b}(u,v)$ DN value to standard whiteboard images, $R_{\mathrm{ref},b}$ as the whiteboard known reflectance. On this basis, for the faded area (represented by the overall rise of reflectance spectrum in the visible band and the weakening of characteristic absorption peak), adaptive histogram matching was used to enhance the faded area, and the target histogram of normal gilt area was used as reference to carry out nonlinear gray mapping. After quality enhancement is completed, a material feature vector of $N_B = 15$ dimensions are generated for each pixel:

$$f_{spec}^{(m)}(u,v) = \left[R_{\lambda_1}^{(m)}, R_{\lambda_2}^{(m)}, \dots, R_{\lambda_{15}}^{(m)}, \nabla^{(m)}, R_{\lambda_5}\nabla^2 R_{\lambda_{10}}\right]^{\mathrm{T}} \tag{12}$$

It includes 15 band reflectance values and the first and second order spatial gradients of the middle band, providing input for subsequent material classification and degradation identification. The processing of strain and vibration signals focuses on the removal of mechanical noise and environmental interference. The relationship between the original wavelength shift data $\Delta\lambda_B$ of FBG strain sensing and strain ε and temperature T is as follows:

$$\frac{\Delta\lambda_B}{\lambda_B} = S_\varepsilon \cdot \varepsilon + S_T \cdot \Delta T \tag{13}$$

Where λ_B is the central wavelength of Prague, S_ε and S_T are the strain and temperature sensitivity coefficients, respectively. In order to separate the cross-sensitivity effect of strain and temperature, each FBG measurement point is equipped with a temperature compensation grating [32],[33]. On this basis, Daubechies wavelet family (db4, the number of decomposition layers is 5) is used for threshold denoising of strain time domain sequence $\varepsilon[n]$ (n is the discrete sampling sequence number). The soft threshold function is as follows:

$$\widehat{w}_{j,k} = sgn\left(w_{j,k}\right) \cdot max\left(\quad |, w_{j,k}\ | - \lambda_j 0 \right) \tag{14}$$

Where $w_{j,k}$ is the k-th wavelet coefficient of the j-th layer wavelet decomposition, the threshold $\lambda_j = \sigma_j\sqrt{2\ln N}$, σ_j is the noise standard deviation estimated by the median absolute deviation of the j-th layer coefficients, and N is the signal length. After denoising, the time-frequency domain statistical feature compression of the signal is carried out: the mean $\bar{\varepsilon}$, standard deviation σ_ε, skewness γ_1, kurtosis γ_2 of each frame (frame length 256 points, overlap rate 50%), as well as the center of gravity frequency f_c and spectral entropy H_s in the frequency domain are extracted:

$$f_c = \frac{\sum_f f \cdot |\hat{\varepsilon}(f)|}{\sum_f |\hat{\varepsilon}(f)|},\ H_s = -\sum_f \tilde{p}(f)\, log\, \tilde{p}(f) \tag{15}$$

Where $\hat{\varepsilon}(f)$ is the Fourier spectrum of the strain signal and $\tilde{p}(f)$ is the normalized power spectral density. After the MEMS acceleration signal goes through the same wavelet denoising process, the tilt angle is further extracted:

$$\theta_x = arcsin\left(\frac{a_x}{g}\right) \tag{16}$$

$$\theta_y = arcsin\left(\frac{a_y}{g}\right) \tag{17}$$

a_x, a_y are static gravity components, which are used to monitor the overall attitude deviation of cultural relics. Finally, all the enhanced and feature pre-extracted data are organized into a unified time-space index data structure. Each frame of data contains six fields: timestamp, spatial coordinate, geometric point cloud local descriptor, multi-spectral material vector, strain statistical feature and vibration/tilt state quantity, providing clear structure and quality control data input for multi-mode fusion in section 3.

3. MULTI-MODE PERCEPTION DATA FUSION ALGORITHM AND DIGITAL TWIN CONSTRUCTION

After the completion of time synchronization and spatial registration, heterogeneous sensor data faces the fundamental challenge of semantic gap between modes: The 3D structural light point cloud expresses the Euclidean spatial distribution of geometric contour, the multi-spectral image represents the spectral domain characteristics of the material surface reflectance, the FBG strain sequence reflects the time-varying dynamics of the local mechanical response, and the MEMS acceleration carries the rigid body motion and vibration mode information. These four types of data not only have different physical meanings and tensor dimensions ($\mathbb{R}^{N\times 3}$ for point cloud, $\mathbb{R}^{H\times W\times B}$ for multispectrum, $\mathbb{R}^{T\times C}$ for FBG, and $\mathbb{R}^{T\times 6}$ for MEMS), but also their internal space-time correlation is not a simple corresponding relationship. It is subject to the physical constraint network formed by the geometric structure, material properties and mechanical transmission path of Changxin Palace Lantern itself. Therefore, fusion is not a simple splicing of data, but consistent mapping across modes needs to be established under a unified physical framework. In this paper, a three-domain coupling loss function $\mathcal{L}_{\text{coup}}$ is defined to quantify the internal consistency deviation between multi-modal data. The spatial domain loss $\mathcal{L}_{\text{spa}}$ measures whether the structural properties reflected by different modes in the same spatial location are self-consistent, such as the angle constraint between the normal vector of the geometric point cloud and the gradient of the multispectral image, and the empirical correlation between the location of the strain measurement point and the geometric curvature. The time domain loss $\mathcal{L}_{\text{tem}}$ is concerned with the consistency of the dynamic changes of each mode, for example, the relationship between the vibration acceleration and the strain change rate should satisfy the kinematic derivative. The physical domain loss $\mathcal{L}_{\text{phy}}$ is the core constraint, and the heat conduction equation, linear elastic constitutive equation and vibration wave equation are introduced as regular terms to force the fusion results to be physically interpretable. In the case of thermal - mechanical coupling, FBG measured temperature $T(x^{(m)},t)$ and the strain field epsilon $\varepsilon(x^{(m)},t)$ between should satisfy thermal elastic constitutive relation:

$$\varepsilon_{thermal} = \alpha_T \cdot \Delta T\left(x^{(m)},t\right) \tag{18}$$

Where α_T is the thermal expansion tensor (for isotropic copper matrix, $\alpha_T = 16.7\times 10^{-6}/°C$). The total expression of coupling loss in the three domains is:

$$\mathcal{L}_{coup} = \lambda_{spa}\mathcal{L}_{spa} + \lambda_{tem}\mathcal{L}_{tem} + \lambda_{phy}\mathcal{L}_{phy} \tag{19}$$

Where λ_{spa}, λ_{tem} and λ_{phy} are the weight coefficients of each domain loss, which are 0.3, 0.2 and 0.5 respectively, and the physical domain occupies the highest weight to reflect the dominant role of physical prior. This loss function, as a regular term embedded in the fusion network as a soft constraint, does not enforce complete agreement of all modes, but penalizes fusion results that seriously deviate from physical rationality [34].

In terms of multi-modal latent space alignment, some supervised methods rely on a large number of cross-modal annotation data, but the point-by-point corresponding labels of each mode at each time cannot be obtained in the real monitoring scene of Changxin Palace Lantern. This paper proposes Weakly supervised Cross-Modal Adversarial Alignment (WCM-AA). The core idea of WCM-AA is to use pairs of spatio-temporal correlations between modes (rather than point-by-point labels) as weak supervised signals. Latent spatial distribution alignment is achieved through adversarial learning. Set the mapping mode by the respective encoder to public latent space $\mathcal{Z}\subset\mathbb{R}^{d_z}$ ($d_z = 256$), get hidden

$z_m = E_m(x_m)$, which $m \in$ {geo, spec, strain, accel. WCM-AA contains A discriminator D for determining the modal category to which the latent code belongs, and a modal alignment network A for generating modal invariant representation. The training objective is the minimax game:

$$\underset{E_m, A}{min} \underset{D}{max} \sum_{m=1}^{4} \mathbb{E}_{x_m \sim p_m} \left[log\, D\left(A\big(E_m(x_m)\big)\right) + \lambda_{wc} \cdot \mathcal{L}_{con}(A, \big(E_m(x_m)\big) c_m) \right] \tag{20}$$

Where p_m is the data distribution of the m-th mode, $D(\cdot)$ outputs the mode class probability, $A(\cdot)$ is the learnable alignment map, c_m is the weakly supervised center vector of the m-th mode (initialized by the consistency constraint of the corresponding spatial-temporal neighborhood of each mode), $\mathcal{L}_{\mathrm{con}}$ is the contrast loss, Its role is to close the distance of different mode latent codes in the same spatial neighborhood and push away non-neighborhood samples. λ_{wc} is the weight of comparison loss (take 1.0). The adversarial training process does not need manual annotation and only relies on the neighborhood relationship defined by the spatial registration results established in section 2 to realize the asymptotic alignment of each mode latent distribution under unsupervised conditions. After WCM-AA alignment, the maximum mean difference (MMD) distance between the latent codes of each mode decreases from about 2.3 to less than 0.4.

After feature alignment is completed, a fusion network that can effectively process asynchronous multi-modal time series data and fuse spatial structure information needs to be constructed. Asynchronous Temporal-Spatial Fusion Network (ATS-FusionNet) adopts dual-branch parallel backbone structure. The Temporal branch takes the latent code sequence $\left\{z_m^{(t)}\right\}_{t=1}^{T}$ of each mode after WCM-AA alignment as the input and adopts Temporal Convolutional Network (TCN) to capture the long-range temporal sequence dependence. The expansion factor of each layer of time convolution increases exponentially ($d = 2^l$, l is the layer index), so that the number of parameters increases linearly with the exponential expansion of the number of layers. The multi-head self-attention mechanism is superimposed on the output of TCN to calculate the attention weight of the i-th time step to the j-th time step [35],[36]:

$$\alpha_{ij} = \frac{exp\left(\frac{\left(W_Q h_i\right)^{\mathsf{T}} \left(W_K h_j\right)}{\sqrt{d_k}}\right)}{\sum_{k=1}^{T} exp\left(\frac{\left(W_Q h_i\right)^{\mathsf{T}} \left(W_K h_k\right)}{\sqrt{d_k}}\right)} \tag{21}$$

Where h_i is the hidden state of the i-th time step output by TCN, W_Q and W_K are the projection matrix of query and key (dimension $d_k = 64$), and time enhancement features are obtained through attention weighting:

$$h_i^{attn} = \sum_j \alpha_{ij} \left(W_V h_j\right) \tag{22}$$

W_V is the value projection matrix. The output of the time branch is:

$$H_{temp} \in \mathbb{R}^{T \times d_h} (d_h = 128) \tag{23}$$

The spatial branch constructs a graph neural network based on the spatial adjacency of geometric measurement points. The vertices of the Changxin Palace Lantern 3D mesh are downsampled to $N_g =$

2048 key nodes as graph nodes $\mathcal{V}_g$. The edges between nodes are determined jointly by the Euclidean distance less than the threshold r_{edge} (taking 1.5 times the average grid edge length) and the physical connection relationship, and the adjacency matrix $A_{\text{adj}} \in \{0,1\}^{N_g \times N_g}$ is formed. The initial feature of each node v_i is the multi-mode latent code weighted aggregate value at its corresponding position. Chebyshev spectral graph convolution (ChebNet) is used in the graph convolution layer. The l-layer graph convolution operation is defined as follows:

$$H^{(l+1)} = \sum_{k=0}^{K} \Theta_k^{(l)} \cdot T_k(\tilde{L}) H^{(l)} \tag{24}$$

Where K is the order of the Chebyshev polynomial (take 3), $\tilde{L} = 2L/\lambda_{\max} - I$ is the normalized Laplace matrix ($\lambda_{\max}$ is the maximum eigenvalue of the Laplace matrix, and I is the identity matrix), $T_k(\cdot)$ is the k-th order of the Chebyshev polynomial, which satisfies the recurrence relation:

$$T_k(x) = 2xT_{k-1}(x) - T_{k-2}(x) \tag{25}$$

$\Theta_k^{(l)}$ is the learnable parameter matrix of the l-th layer and the k-th order. After three-layer graph convolution (the output dimensions of each layer are 128, 128 and 256 respectively), the spatial enhancement features are obtained:

$$H_{spat} \in \mathbb{R}^{N_g \times 256} \tag{26}$$

The global temporal sequence features of the temporal branch and the node spatial features of the spatial branch are cross-attention fused to obtain the spatio-temporal joint features:

$$H_{joint} = CrossAttn\left(H_{temp}^{(m)}, H_{spat}\right) \tag{27}$$

Asynchronous sampling rate inconsistency is one of the most intractable engineering problems in practical systems. FBG collects strain dynamics at 1000 Hz, MEMS at 200 Hz, multispectral imaging at only 5 Hz, and 3D structured light output geometric point clouds at 2 Hz frame rate. Traditional methods usually upsample all modes to the highest sampling rate (1000 Hz) and then fuse them, but this will introduce a lot of false interpolation information and increase the computational load sharply. In this paper, the Gated Asynchronous Fusion Module (GAFM) is designed in ATS-FusionNet. This module does not force each mode to have data at every time. Instead, the fusion weight is dynamically determined according to the actual data availability $v_m(t) \in \{0,1\}$ and confidence evaluation $c_m(t) \in [0,1]$ of each mode at the current time t:

$$g_m(t) = \frac{exp\big(\beta \cdot v_m(t) \cdot c_m(t)\big)}{\sum_{k=1}^{4} exp\big(\beta \cdot v_k(t) \cdot c_k(t)\big)}, \; c_m(t) = \sigma\left(w_c^{\top} z_m^{(t)} + b_c\right) \tag{28}$$

Where β is the temperature coefficient (2.0) to control the sharpness of the weight distribution, $\sigma(\cdot)$ is the Sigmoid function, and w_c and b_c are the confidence evaluation parameters that can be learned. When there is no new data arrival of a certain mode at time t ($v_m(t) = 0$), its gating weight $g_m(t)$ decays rapidly but does not completely return to zero and still uses the hidden state of the last effective time to propagate through the gated cyclic unit (GRU) to realize the state prediction of the missing mode. The fusion output is:

$$z_{fuse}(t) = \sum_{m=1}^{4} g_m(t) \cdot \tilde{z}_m(t) \tag{29}$$

Where $\tilde{z}_m(t)$ is the timeliness corrected latent code of the m-th mode at time t (if $v_m(t) = 1$, it is the current code, otherwise it is the predicted value of GRU). The design makes the fusion frame rate no longer subject to the slowest sensor but dynamically updated with the arrival of data. The effective output frame rate of the system is about 35-50 Hz under typical working conditions, much higher than the 2 Hz of the single structured light mode, which significantly improves the time resolution of state monitoring. The number of core learnable parameters of GAFM module is summarized in Table 2. The total number of parameters of ATS-FusionNet is about 2.87M, among which the trainable parameter is 2.71M. The forward inference of a single frame on the embedded inference platform takes about 18 ms.

Table 2. Statistics of learnable parameters of each main module of ATS-FusionNet

Name of module	Main layer types	Number of parameters (thousands)	Dimension of input	Dimension of output
Time branch TCN	Expand convolution by 5 layers	418	256×T	128×T
Time branch self-attention	Long attention (8 heads)	132	128×T	128×T
Spatial branch graph convolution	ChebNet×3 layers	892	2048×256	2048×256
Temporal and spatial cross attention	Long attention (4 heads)	264	256×128	256×128
Gated asynchronous fusion (GAFM)	GRU+ gated linear layer	284	256×4	256
Output state decoder	Fully connected ×3 layers	720	256	512

The final output of ATS-FusionNet is a high-dimensional joint state tensor $\Psi(t) \in \mathbb{R}^{N_g \times D_\Psi}$ ($D_\Psi = 512$), The tensor includes the fused global attributes (global temperature mean, global vibration level, attitude angle offset) and local abnormal sensitive features (node-level strain residual, curvature change rate, spectral drift) at each key geometric node. The tensor has been spatially anchored to the 3D grid of Changxin Palace Lantern, which provides a direct data source for the state mapping of the digital twin.

The construction of a digital twin takes the fused state tensor $\Psi(t)$ as input and maps it into a surface state field that can be intuitively characterized and physically interpreted. In this paper, a refined 3D mesh $\mathcal{M}_{\text{twin}}$ (containing about 12eight thousand triangular facets) is used as the geometric base of the twin. The mesh is obtained from the high-precision structural light scanning reconstruction of Changxin Palace Lantern (accuracy 0.03 mm), and all recognizable casting textures, mortion-tenon joints and decorative lines are preserved. Mapping $\Psi(t)$ to all the vertices of $\mathcal{M}_{\text{twin}}$ requires a dense interpolation from $N_g = 2048$ key nodes to 128,000 vertices. In this paper, Differentiable Rendering technology is introduced to realize this mapping. The core idea is to regard state field rendering as a derivable image generation process and optimize the interpolation parameters by reverse rendering gradient. Define the state value $s(q)$ of vertices (q is the grid vertex coordinates) as the weighted sum of the state values of neighboring key nodes:

$$s(q) = \sum_{i=1}^{N_g} w_i\,(q) \cdot \psi_i,\ w_i(q) = \frac{exp\left(-\| q - \frac{p_i \|^2}{(2\sigma_{ker}^2)}\right)}{\sum_{j=1}^{N_g} exp\left(-\| q - \frac{p_j \|^2}{(2\sigma_{ker}^2)}\right)} \tag{30}$$

Where p_i is the spatial coordinate of the i-th key node, ψ_i is the state component of the corresponding node in $\Psi(t)$, and σ_{ker} is the Gaussian kernel bandwidth (2.5 times the average edge length of the grid, about 3.2mm). The interpolation process is embedded in the differentiable rendering pipeline, and the rendering loss (the error between the rendered state field image and the measured value of each sensor) is used to optimize the σ_{ker} and the weight of each node, so that the interpolation result is close to the real distribution in both visual and physical sense. After the differentiable rendering mapping, three main types of surface state fields are obtained: temperature field $T(q,t)$, equivalent strain field $\varepsilon_{\text{eq}}(q,t)$ and vibration mode amplitude field $A(q,t)$.

In order to avoid the computational explosion caused by frequent geometric mesh updates, this paper proposes double-twin architecture. The bottom layer is a geometre-physical static model $\mathcal{B}_{\text{static}}$, storing the basic geometric topology and material parameters of the Changxin Palace Lantern (Young's modulus $E = 110$ GPa, Poisson's ratio $\nu = 0.34$, density $\rho = 8.96$ g/cm³, thermal conductivity $\lambda_T = 401$ W/(m·K) of the copper matrix, The gold layer thickness is about 0.2mm and the corresponding equivalent mechanical parameters) and the modal shape base calculated by finite element prediction. The upper layer is the run-time dynamic state layer $\mathcal{B}_{\text{dyn}}(t)$, which only stores the time-varying state field data (temperature field, strain field, vibration field) and its derived variables (temperature gradient, stress concentration coefficient, etc.). The correlation between the two layers is realized by the elastic mapping function $\mathcal{F}_{\text{elas}}: \mathcal{B}_{\text{dyn}}(t) \mapsto \mathcal{B}_{\text{static}}$, which projects the increment of the current state field (with respect to the initial reference state) onto the physical space of the underlying model based on the principle of linear superposition, and calculates the corresponding stress-strain response. The mathematical form of the elasticity mapping function is:

$$\mathcal{F}_{elas}(\Delta S) = \int_{\mathcal{M}_{twin}} K_{FEM}(q) \cdot \Delta S(q)\, dq \tag{31}$$

Where $\Delta S(q)$ is the change of the state field at position q (relative to the previous time or initial state), $K_{\text{FEM}}(q)$ is the influence kernel matrix precalculated by the finite element reduction method (the discrete form at N_g key nodes are $K_{\text{FEM}} \in \mathbb{R}^{N_g \times N_g}$, and $K_{\text{FEM}} \in \mathbb{R}^{N_g \times N_g}$. The kernel matrix has completed LU decomposition and low-rank approximation in the offline stage, and only one matrix-vector multiplication (about 0.3 ms) is needed for online calculation. The two-layer structure eliminates the need for static geometry and material parameters to be re-calculated in each monitoring cycle, while the state layer is updated independently with a lightweight data structure, greatly reducing computing resource consumption.

In the scenario of long-term continuous monitoring, sensor drift, slow environmental change and progressive changes of the cultural relic itself (such as micro-crack propagation and micro-stripping of the gilt layer) will lead to gradual mismatch of the initial parameters of the underlying static model. This paper further introduces an incremental online update mechanism, whose core strategy is "freezing geometry, fine-tuning parameters, and full state replacement". Specifically, the residual error between the current state field $\mathcal{B}_{\text{dyn}}(t)$ and the physical predicted response $\mathcal{B}_{\text{static}}\left(\mathcal{F}_{\text{elas}}^{-1}\left(\mathcal{B}_{\text{dyn}}(t)\right)\right)$ of the underlying model in this state is calculated for each monitoring period (set as 1 hour in this paper). If the L_2 norm of the residual continues to exceed the threshold η (3 times the standard deviation of the mean value of the initial state residual), then the parameter fine-tuning is triggered: the influence kernel matrix K_{FEM} in the elastic mapping function is updated incremental by gradient descent method, and the step size is limited by the stability constraint (that is, the change rate of the model output before and after the update is not more than 5%). During the fine-tuning process, the underlying geometric mesh $\mathcal{M}_{\mathrm{twin}}$ is always kept unchanged, and only the material equivalent parameters and the values of the kernel matrix are modified. In the state layer, each frame (about 20-30 ms) is completely replaced with new data, and no historical state is accumulated to avoid the accumulation of numerical drift in long-term operation. The effectiveness of the update strategy in long-term running experiments is evaluated by the growth rate of state reconstruction error. Table 3 shows the performance guarantee parameters of the mechanism at different time scales, where the maximum state lag is defined as the time difference between the completion time of fusion and the output time of the state field.

Table 3. Multi-time scale performance parameters of the incremental online update mechanism

Scale of time	State layer update frequency	Freezing duration of the underlying static model	Parameter fine-tuning triggers the residual threshold	Maximum state lag
Second level (1-60 s)	35-50 Hz	Keep on freezing	Do not trigger	≤22 ms
Minute level (1-60 min)	Same as above	Keep on freezing	1.2× Initial residual error	≤25 ms
Hour level (1-24 h)	Same as above	Check once every 1h	3.0× initial residual error	≤30 ms
Day level (1-30 days)	Same as above	Calibrate once a day	2.5× the mean value of the residual of the previous day	≤35 ms

With incremental online update, the dual-twin architecture not only ensures real-time status monitoring (end-to-end delay ≤35 ms) but also endows the digital twin with the ability to adapt to long-term slow change process. Moreover, the freezing of the underlying static model within one working day keeps the computing resource consumption constant, avoiding the problem of computing power explosion caused by the iterative update of the full model of the traditional digital twin. The state fusion and twin construction completed in section 3 provides a complete physical interpretable data source and efficient computing carrier for real-time state monitoring and anomaly determination based on twin residuals in section 4.

4. REAL-TIME STATUS MONITORING AND ANOMALY DETERMINATION MECHANISM

After the construction of the digital twin is completed, the core value is reflected in its quantitative perception of the real-time health status of cultural relics and its ability to discriminate abnormalities. However, the fusion state tensor itself only reflects the current value of the sensing data, There is no physical benchmark to determine "what is abnormal", and the deviation of the sensor reading from the initial value may be due to real structural changes (such as local strain redistribution caused by microcrack propagation), environmental fluctuations (such as thermal expansion and cold contraction caused by daily changes in the temperature of the exhibition hall), or degradation of the sensor's own performance (such as wavelength shift caused by FBG coupling aging). These three types of inducements may present similar numerical offsets in the surface state field but correspond to completely different disposal strategies. The core idea of this paper is that the state field after fusion is not directly used as the monitoring index, but the static model at the bottom of the digital twin is used as the physical benchmark, the physical equation is used to carry out forward simulation prediction of the current state, and then the residual between the predicted value and the actual fusion value is used as the basis for anomaly discrimination. The advantage of this strategy is that physical equations naturally encode fundamental laws such as material constitution, thermodynamic equilibrium, and wave propagation, and any residual pattern that violates these laws points to "physically unexplained" components, which are the signal carriers of potential anomalies.

The forward simulation prediction is based on the finite element discrete scheme in the underlying geometer-physical static model $\mathcal{B}_{\text{static}}$ to solve the heat conduction equation and the elastic vibration equation respectively. For the temperature field $T(q,t)$, its evolution in the copper bulk of Palace Lantern satisfies the unsteady heat conduction equation:

$$\rho c_p \frac{\partial T(q,t)}{\partial t} = \nabla \cdot \left(\lambda_T \nabla T(q,t)\right) + Q(q,t) \tag{32}$$

The $\rho = 8960\,\text{kg/m}^3$ as the density of copper matrix, $c_p = 385\,\text{J/(kg}\cdot\text{K)}$ for specific heat, $\lambda_T = 401\,\text{W/(m}\cdot\text{K)}$ as the thermal conductivity, $Q(q,t)$ is a volume heat source (exhibition hall can be ignored in the environment), The boundary conditions include the natural convective heat transfer with air $\lambda_T \partial T/\partial n = h(T_{\text{amb}} - T)$, where $h \approx 5 - 10\,\text{W/(m}^2\cdot\text{K)}$ is the convective heat transfer coefficient (calibrated to 7.2 by the pre-experiment) and T_{amb} is the ambient temperature. For the linear elastic vibration displacement field $u(q,t)$, it satisfies the elastic wave equation under the assumption of micro-amplitude vibration:

$$\rho \frac{\partial^2 u(q,t)}{\partial t^2} = \nabla \cdot \sigma(q,t) + f_{ext}(q,t) \tag{33}$$

Where $\sigma = C{:}\,\varepsilon$ is the stress tensor (C is the fourth-order elastic tensor, determined by E and ν for the isotropic copper matrix), $\varepsilon = \frac{1}{2}(\nabla u + \nabla u^{\top})$ is the strain tensor, and f_{ext} is the external excitation (such as the environmental vibration from the floor of the exhibition hall). The above PDE system is solved on $\mathcal{M}_{\text{twin}}$ mesh using finite element discretization and Newmark-β time integration scheme (parameters $\beta = 0.25$, $\gamma = 0.5$) to obtain the predicted temperature field $\hat{T}(q,t)$ and the predicted vibration displacement field $\hat{u}(q,t)$, and then the predicted strain field $\hat{\varepsilon}(q,t)$ is derived. The actual fusion state field is denoted as $S_{\text{fuse}}(q,t)$, including the actual temperature field T_{fuse}, strain field $\varepsilon_{\text{fuse}}$ and vibration amplitude A_{fuse}. The state residual field is defined as the difference between the predicted value and the actual value:

$$\Delta S(q,t) = S_{fuse}(q,t) - \hat{S}(q,t) \tag{34}$$

The residual field should only contain Gaussian white noise with zero meaning and amplitude limited by sensor noise in the ideal case of no anomaly. When the state of the cultural relic changes (such as the occurrence of microcracks leading to the decrease of local stiffness, the stripping of the gold layer changing the surface thermal radiation characteristics) or the sensor performance deteriorates, a systematic non-zero pattern will appear in the residual field.

In order to extract concise and sensitive judgment indexes from the high-dimensional residual field, this paper decomposes the residual into two orthogonal dimensions: spatial residual spectrum and temporal residual sequence. The spatial residual spectrum $\mathcal{R}_{\text{spa}}(\omega_s)$ is defined as the spatial Fourier transform modulus of the residual field on the mesh surface, which decomposes the spatial distribution characteristics of the residual according to the spatial frequency ω_s (unit is rad/mm). Firstly, the sum of the squares of the components of the residual field is taken as a scalar field $R(q,t) = \| \Delta S(q,t) \|^2$, and then the spherical harmonic decomposition coefficient $c_{lm}(t)$ is calculated on the $\mathcal{M}_{\text{twin}}$ surface (since the mesh is approximately closed shell, spherical harmonic function is a suitable orthogonal basis). The low-frequency component of the spatial residual spectrum ($\omega_s < 0.01$ rad/mm) reflects large-scale global migration (such as overall temperature drift), and the medium-frequency component ($0.01 \leq \omega_s \leq 0.1$ rad/mm) corresponds to medium-scale regional anomalies (such as local hot spots or strain concentration areas). The high-frequency component ($\omega_s > 0.1$ rad/mm) indicates local abrupt changes (such as crack tip stress singularity or sensor noise). The time residual sequence $\Delta S(q_i,t)$ evolves with time at the key node q_i, and the direction and amplitude of its time change rate $\partial \Delta S / \partial t$ are used to distinguish sudden impact (rapid recovery after abrupt change), slow deterioration (continuous drift) and sensor drift (one-way monotone migration). Based on the weighted integral of the spatial Residual spectrum and the statistics of the temporal residual series, this paper constructs the residual-driven State Deviation Indicator (RSDI), which is defined as:

$$RSDI(t) = \alpha \cdot \int_{\omega_{min}}^{\omega_{max}} W(\omega_s) \cdot \mathcal{R}_{spa}(\omega_s,t)\, d\omega_s + \beta \cdot \frac{1}{N_g} \sum_{i=1}^{N_g} \left| \frac{\partial \Delta S(q_i,t)}{\partial t} \right| \tag{35}$$

Where $W(\omega_s)$ is the frequency weighting function, which assigns the highest weight to the intermediate frequency component ($W(\omega_s) = \exp\left(-(\omega_s - 0.05)^2/(2 \times 0.02^2)\right)$, and multiplizes the low frequency and high frequency components by the suppression coefficients 0.4 and 0.3 respectively. The aim is to enhance the sensitivity to structural anomalies (corresponding to intermediate frequency spatial scales), while suppressing the interference of slow environmental changes and high frequency sensor noise. α and β are the dimensional balance coefficients (0.12 and 0.88, respectively, determined by the normalization of the amplitude range of each component). The RSDI is a dimensionless scalar with values ranging between $[0,1]$, and the physical meaning of its numerical value is calibrated by the response characteristics of three typical anomaly patterns. Table 4 summarizes the differentiated responses of the three typical anomaly patterns in terms of RSDI and residual decomposition characteristics.

Table 4. Differentiated response patterns of three typical abnormal patterns in terms of RSDI and residual decomposition characteristics

Type of exception	RSDI response speed	Dominant band of spatial residual spectrum	Temporal residual sequence characteristics	Typical volatility cycle/duration
A sudden shock	Millisecond jumps	High frequency band (>0.10 rad/mm)	The transient spike is followed by exponential decay with no trend drift	<2 s
Slow deterioration	Linear growth at the hourly level	Middle frequency band (0.01-0.10 rad/mm)	Monotonically increasing or slowly oscillating, the residual continues to increase	Hours to days
Drift of sensor	Slow gradual increase (<2%/h)	Low frequency band (<0.01 rad/mm)	Unidirectional linear/quadratic drift with no spatial diffusion	Continuous accumulation

The peak amplitude of RSDI caused by sudden shocks is often much higher than that of the other two categories (up to 20-50 times of the baseline value), but its duration is very short and there is no subsequent accumulation. The growth rate of RSDI corresponding to slow deterioration is relatively slow (typical value is 0.5-3%/h), but the spatial residual spectrum is concentrated in the middle frequency band and shows a trend of spreading from local to adjacent areas over time, which is a typical representation of microcrack propagation or interfacial debonding of gold layer. Sensor drift is manifested as that the residual spectrum mainly falls in the low frequency band, the time series shows monotonicity (linear or quadratic) and there is no sign of diffusion to surrounding nodes in space, because the drift usually occurs independently on individual sensing channels.

After calculating the RSDI timing of each key node, the anomaly tracing needs to be located from "overall anomaly existence" to "where the anomaly is located" and "what is the anomaly degree". The gradient information of the spatial residual field provides a natural basis for positioning. In this paper, the residual gradient vector $\nabla\Delta S(q_i,t)$ at node i is defined, and its discrete calculation formula on the mesh surface is as follows:

$$\nabla\Delta S(q_i,t)=\frac{1}{|\mathcal{N}(i)|}\sum_{j\in\mathcal{N}(i)}\frac{\Delta S(q_j,t)-\Delta S(q_i,t)}{\| q_j-q_i \|}\cdot\frac{q_j-q_i}{\| q_j-q_i \|} \tag{36}$$

Where $N(i)$ is a ring neighborhood of node i. With $\left\|\nabla\Delta S(q_i^{(m)},t)\right\|$ as the pixel value, the thermal map is generated on the mesh surface, and the Density Peaks Clustering algorithm (DPC) is used to automatically cluster the high-gradient regions. In clustering, nodes whose local density ρ_i and high local density distance δ_i both exceed their respective thresholds are selected as clustering centers, where:

$$\rho_i=\sum_j exp\left(-\frac{d_{ij}^2}{d_c^2}\right) \tag{37}$$

d_{ij} is the geodesic distance of the grid between nodes i and j, and d_c is the truncation distance (taking 5 times of the average side length of the grid, about 6.4mm):

$$\delta_i=\min_{j:\rho_j>\rho_i} d_{ij} \tag{38}$$

The clustering results output multiple abnormal hot spots:

$$A_k\subset M_{twin}(k=1,2,\dots,K) \tag{39}$$

Each region contains its central location, area proportion and average residual intensity:

$$\bar{R}_k=mean_{q\in A_k}\left\|\Delta S(q^{(m)},t)\right\| \tag{40}$$

After locating the abnormal hotspot, you need to determine the alarm level corresponding to each hotspot. Decision tree classifier, this paper designed three layers of lightweight to node level $x_1=$ RSDI(q_i,t), abnormal duration $x_2=\int_{t-\tau}^{t}\mathbb{I}\,(\text{RSDI}(q_i,s)>\theta_{\text{base}})\,ds$ (τ as the back window, 300 s) and the spatial diffusion velocity $x_3=d(\mathcal{A}_k)/dt$ (that is, the rate of change of the characteristic radius of the abnormal region with time) are taken as the input. The first layer determines whether the RSDI exceeds the emergency threshold θ_{crit} (8 times the standard deviation of the baseline value m_{RSDI} in this paper), and if it exceeds, it is directly judged as critical; If it does not exceed but $x_2>60$ s and $x_3>0.02$ mm/s, it is judged as an alarm (indicating that the anomaly is spreading); If only x_1 is between θ_{warn} (3 times the baseline standard deviation) and θ_{crit}, $x_2<30$ s and x_3 is close to zero, it is judged as a warning (suspected transient disturbance or early small change). The threshold parameters of the classifier are determined by the statistics of various simulated working conditions collected in the pre-experiment, as shown in Table 5. The threshold values given in the table are analyzed by RSDI distribution statistics under five typical excitation responses (including impact hammer excitation, local heating of heat gun, micro-strain loading of stepper motor, frequency sweep of standard shaking table and artificial bias of sensor). The upper bound of the 95% confidence interval is used as the benchmark.

Table 5. Key threshold parameters for lightweight decision tree classifiers

Grade of classification	RSDI threshold (relative to baseline)	Minimum duration requirement (s)	Minimum spatial diffusion rate (mm/s)	Corresponding to typical physical state description
Early warning	3.0 × σ_baseline	≥10	≥0.005	Local stress redistribution/slight thermal disturbance
An alarm	5.5 × σ_baseline	≥60	≥0.020	Persistent strain anomalies/progression of regional degradation
In danger	9.0 × σ_baseline	≥5	No limit	Sudden shocks/significant structural responses

Closed-loop feedback and recalipay of digital twin are the key links to achieve long-term stable monitoring. The above anomaly determination process not only outputs alarm signals but also inputs the determination results back to the digital twin as feedback to trigger differentiated parameter correction strategies. Specifically, when the decision tree determines that the anomaly type of a hot spot is “sensor drift” (its characteristics meet the three conditions of low frequency band dominance, monotone time series and no diffusion in space), the system automatically marks the sensor channel and triggers local automatic recrystal-calibration. In the recalipation process, the reference measurement points without drifting around the sensor channel are used for cross-correction: if it is an FBG strain channel, the calibration coefficient C_{cal} is recalculated by the ratio between the temperature compensation grating and the adjacent strain measurement points. If it is a MEMS acceleration channel, the zero-bias and scaling factors are recalculated using the three-axis output under static gravity field. After the recalipay is completed, the residual accumulation history of the corresponding measurement point is cleared, and the sensor related parameter table in $\mathcal{B}_{\mathrm{static}}$ is updated. When the exception type is judged to be "slowly deteriorating" (that is, confirmed to be a real change in the structural state), the mapping parameter update of the upper dynamic state layer $\mathcal{B}_{\mathrm{dyn}}(t)$ is performed instead of modifying the underlying static model, and the fusion state field $S_{\mathrm{fuse}}(q,t)$ at the current moment is used as the new reference benchmark. The baseline state vector in the elasticity mapping function $\mathcal{F}_{\mathrm{elas}}$ is updated so that the subsequent residual calculation is based on the latest stable state rather than the initial state, so as to avoid continuously counting the complete gradual changes as anomalies. The underlying $\mathcal{B}_{\mathrm{static}}$ triggers the aforementioned incremental fine-tuning of the kernel matrix only when the residual continues to increase beyond the recalibration threshold and the non-sensor cause is confirmed (through redundant sensor cross-validation).

The closed-loop feedback mechanism forms a complete “perception → fusion → twin → decision → correction” cyclic link. Sorted out in chronological order: After processing in section 2, the multi-mode sensor data enters the fusion network in section 3. The fusion state tensor drives the twin to generate the state field. The twin predicts and calculates the residual through the physical equation forward, and the residual decomposition builds RSDI and drives the classification of the decision tree. The updated twin affects the residual calculation benchmark at the next moment. The key timing parameters of the closed-loop operation are summarized in Table 6, where the response delay is defined as the complete link time from the occurrence of abnormal physics (such as the impact moment of the impact hammer) to the corresponding grading result of the system output, and the convergence time of recallation refers to the time required from triggering recallation to the re-stabilization of sensor parameters. The update time of the state layer refers to the time cost of the new reference state taking effect after the adjustment of mapping parameters.

Table 6. Key timing sequence and updating parameters of each link of closed-loop monitoring system

Closed loop link	Execution frequency/trigger condition	Typical delay or time consuming	Update the object	Whether real-time fusion flow is affected
Perception-fusion-twin	35-50 Hz (continuous)	18-22 ms	The state tensor $\Psi(t)$	Yes (primary link)
Residuals are calculated with RSDI generation	Execution per frame	2.1 ms	RSDI timing library	No (parallel)
Decision tree classification	Execution per frame	0.3 ms	Alarm status register	No
Sensor recalibration	Triggered after drift determination	120-180 s	Single channel calibration coefficient	No (bypass)
State layer mapping parameters are updated	After slow deterioration is confirmed	800 ms	$\mathcal{B}_{\mathrm{dyn}}$ reference baseline	No (transient blocking <1s)

The direct benefit of closed-loop feedback is the significant suppression of false alarm rate in long-term continuous monitoring. Without this mechanism, sensor drift and seasonal ambient temperature change will lead to slow shift of residual baseline, making the fixed threshold initially set lose reference significance after a few hours of operation, and the false alarm rate of the system may rise to more than 30%. The closed-loop mechanism maintains the effective false alarm rate at a low level by continuously tracking the long-term trend of the residual baseline and actively recertifying or updating the baseline when confirmed as non-structural causes, while ensuring that true structural anomalies (such as slow deterioration) are not masked by false recertifications. The classification results and abnormal location information output in section 4 are finally presented to cultural relic conservators in the visual twin interface, forming a complete monitoring closed loop from physical perception to human-machine decision.

5. EXPERIMENTAL VERIFICATION AND PERFORMANCE EVALUATION

The core of the experimental verification platform is the 1:1 bronze replica of Changxin Palace Lantern. According to the 3D scanning data of the original collection of Hebei Museum, the replica is cast and formed by the method of losing wax, and the surface is treated by gold plating process. The total height is 48 cm, the diameter of the lamp plate is 28 cm, the wall thickness is not uniform (3~8 mm), and the total weight is about 8.6 kg. The geometric form, material composition and mechanical properties are highly consistent with the original. The copy is installed on a six-DOF high-precision turntable, and the bottom of the turntable is embedded with an electromagnetic vibration table (excitation frequency range 5 ~ 200 Hz, maximum acceleration 2g). The infrared ceramic heating array (maximum heat flux at a single point of 800 W/m²) is arranged around it to simulate the local thermal radiation environment of the exhibition hall. The four types of sensors are installed according to the non-uniform adaptive location strategy in section 2: a 3D structured light scanner (model: LMI Gocator 2512) is fixed 800 mm directly above the photocopy, and a multispectral camera (model: XIMEA MQ022HG-IM-SM5X5) and structured light constitute a 40° baseline angle, 24 FBG strain/temperature sensors are pared on the inner cavity wall and the external key section (the distribution position covers the root of the lamp, the maximum diameter of the lamp body, and the base of the three annular sections, each section has 8 measurement points), 6 MEMS accelerometers (model: ADI ADXL355) is arranged on the edge of the lamp plate, the symmetrical side of the lamp body and the lower edge of the base. All sensors are timed uniformly through the synchronization controller (time synchronization board based on PTP protocol), and the sampling rates are 2 Hz for structured light, 5 Hz for multispectrum, 1000 Hz for FBG (reduced to 100 Hz in static strain mode) and 200 Hz for MEMS. The data acquisition system is based on NI PXei-1085 chassis and works with LabVIEW real-time module to complete multi-channel synchronous recording. The raw data flow is transmitted over Ethernet to a workstation (configured with Intel Xeon Gold 6248 CPU + NVIDIA RTX 3080 GPU) for subsequent fusion and twin computing.

In order to comprehensively evaluate the system performance, this paper designs five typical working conditions to cover the possible state changes of Changxin Palace lanterns in the process of display and handling. Normal (working condition N) refers to the continuous operation of the copy under the condition of static, constant temperature (22±1°C) and no additional excitation, which is used as the baseline reference. The slight vibration condition (condition V) applies sinusoidal sweep excitation with amplitude of 0.1g and frequency of 10 Hz, 30 Hz and 60 Hz respectively through the vibration table to simulate the daily vibration of the exhibition hall ground caused by the flow of visitors or the nearby construction. Local heating condition （condition H） uses infrared heating array to apply continuous heat flow （500 W/m²） to the right area of the lamp plate for 15 minutes, followed by natural cooling for 30 minutes to simulate spotlight irradiation or local ambient temperature sudden rise. The tilt offset working condition (working condition T) rotates the copy by 1°, 3° and 5° along the X axis in turn through the turntable to simulate the small change of placement posture or the uneven settlement of the base during the handling process. Simulated crack strain mutation condition (working condition C) prefabricated an artificial slit of about 25 mm in length (0.3 mm in width, half of the wall thickness) on the outer surface of the lamp body abdomen. The wedge block was driven by a micro stepping motor to periodically open and close the slit (the opening amount is adjustable from 0 to 0.5 mm, and the cycle is 60 s). The respiration effect of microcracks under temperature or vibration excitation is simulated. Data were collected continuously for no less than 30 minutes for each working condition, and samples were divided by a sliding window of 5 minutes. A total of about 18,000 sample frames were obtained for the five working conditions (the length of each frame was fixed at 1 second, that is, 1000-time steps), 80% of which were used for model training and parameter tuning, and 20% were used as test set. In addition to the physical experiment data, this paper also introduces two comparison standards: one is the multi-physical field simulation data generated based on the finite element simulation model of Changxin Palace Lantern replica (established in Abaqus, with a grid size of about 150,000 units), covering the virtual response of the above five working conditions, which is used to expand the diversity of working conditions of the training set; The second is the public heritage monitoring data set (taken from the bronze monitoring records of the EU SensMat Heritage protection project), which is used as the reference for cross-domain generalization tests after format conversion and modal alignment.

In the performance comparison experiment of fusion algorithm, five representative methods are selected as the comparison baseline: Traditional feature-level splicing (directly splicing the statistical feature vectors extracted from each modality and input them into the fully connected classifier), D-S evidence fusion (decision level fusion based on Dempster-Shafer evidence theory on the independent decision results of each modality), cross-modal Transformer (using the recently proposed cross-modal attention fusion architecture, As a representative of deep learning fusion), single-mode optimization (only the single mode with the best performance is used for inference, which is FBG strain mode in this data set), and ATS-FusionNet of this paper. The evaluation indicators include state reconstruction error (RMSE, which measures the root mean square error between the reconstructed joint state tensor and the real state tensor after fusion), anomaly detection accuracy (Acc), false alarm rate (FAR) and F1-score. In the physical experiment, the "real state tensor" is replaced by the reference value of the original data of each sensor interpolated to the grid node with high precision. Figure 1 shows the comprehensive performance comparison of the five methods on the test set. ATS-FusionNet achieves 0.087 in RMSE (31.5% lower than that of cross-modal Transformer), the anomaly detection accuracy is 96.2%±1.1%, and the false alarm rate is 2.3%±0.4%. The F1-score was 0.947±0.012, and the four indexes were significantly better than the control method.

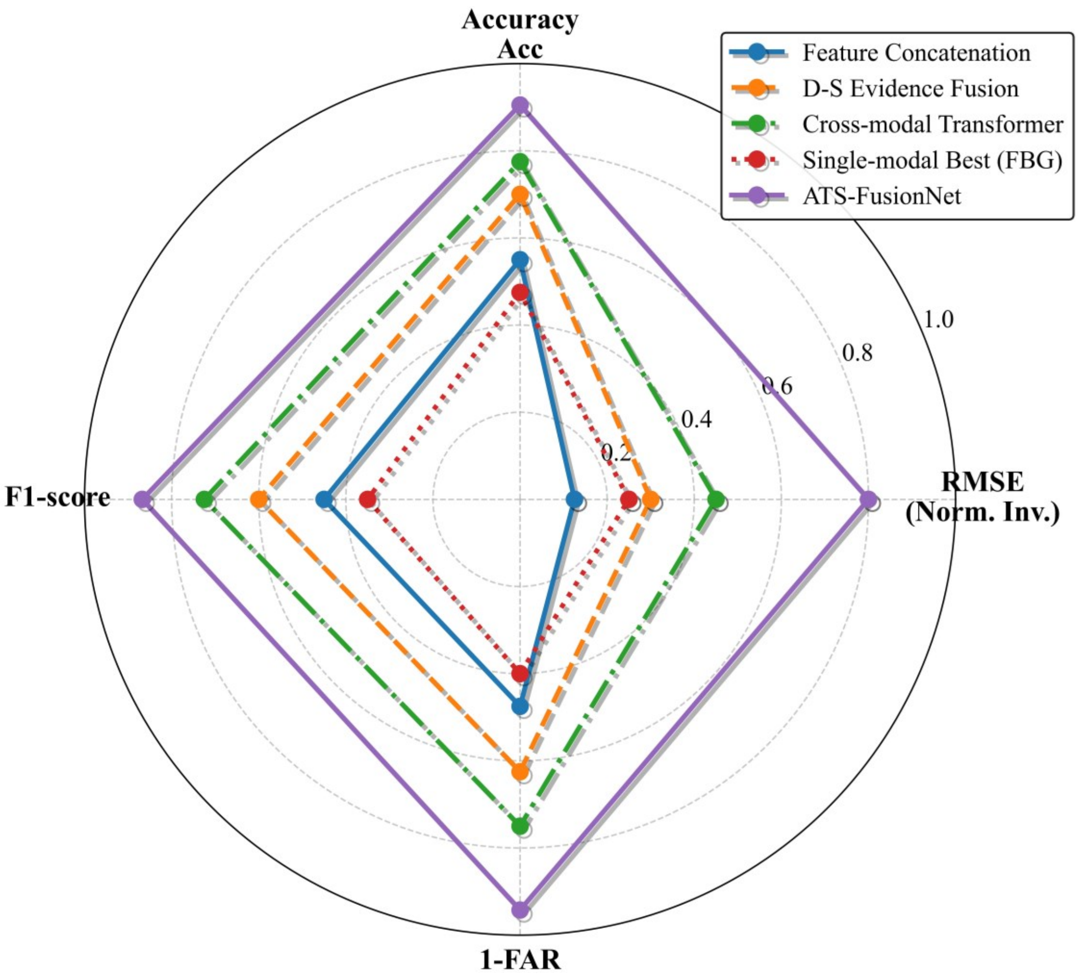

Figure 1. Radar plot of four-dimensional performance indicators of the five fusion methods on the test set

Table 7 summarizes the F1-score results of ablation experiments under various working conditions. The data show that the physical consistency constraint makes the most significant contribution in working condition H (local heating) and working condition C (crack strain mutation), and the F1-score decreases by 9.1% and 11.4% respectively after removal. This is because the thermo-force coupling effect and stress concentration phenomenon in these two working conditions are most dependent on physical prior knowledge. The gated asynchronous fusion module contributes about 5.6% F1 boost on average in all working conditions, and contributes 8.9% in working condition V (vibration, the largest sampling rate difference between modes). The spatial branch of the figure made the most prominent contribution to the working condition T (oblique deviation, obvious change of spatial distribution pattern), and the F1-score decreased by 10.2% after removal. The ablation experiment completely demonstrates the complementarity and indispensable of the three innovation components.

Table 7. Comparison of F1-score in ATS-FusionNet ablation experiment under different operating conditions

Configuration of the model	Working condition N	Working condition V	Working condition H	Working condition T	Working condition C	Five-category average
Complete ATS-FusionNet	0.982	0.951	0.933	0.945	0.922	0.947
Remove physical consistency constraints	0.958	0.914	0.848	0.901	0.817	0.888
Remove gated asynchronous fusion	0.964	0.866	0.897	0.912	0.883	0.904
Remove the graph space branch	0.971	0.925	0.904	0.849	0.894	0.909

The evaluation of the accuracy of twin construction focuses on two dimensions: geometric reconstruction error and state field prediction error. In terms of geometric accuracy, the 3D mesh model of digital twin output by the system was compared with the high-precision reference model (obtained by scanning with FARO Quantum Max laser tracker, accuracy 0.02mm), and the two-way Hausdorff distance was calculated. Figure 2 (a) shows the distribution histogram of Hausdorff distance on the whole mesh surface, where the maximum distance is 0.18 mm, the average distance is 0.073 mm, and the 95% quantile is 0.12 mm. This accuracy level meets the minimum identifiable requirement of geometric deformation in cultural relic protection monitoring (it is generally believed that deformation within 0.2mm has no significant impact on the safety of bronze structure). Figure 2 (b) further shows the maximum local error distribution of each area of the grid under five types of working conditions. It is found that the maximum error area is concentrated in the outer edge of the lamp disk and the edge of the base, mainly because of the large curvature of these areas and the relatively sparse sensor distribution points (limited by the installation space), but the absolute error of this area is still controlled within 0.21mm.

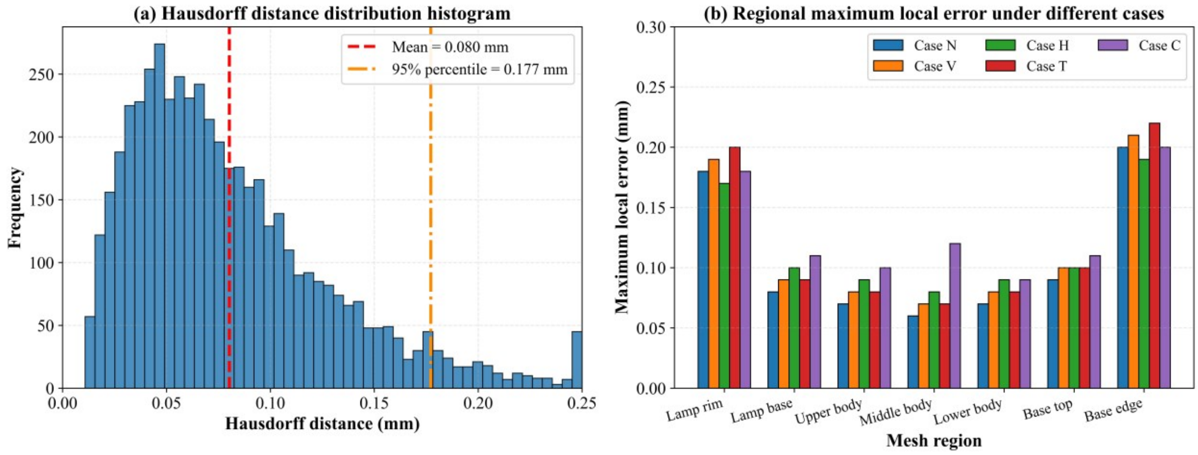

Figure 2. Evaluation of geometrical accuracy of digital twin: (a) histogram of Hausdorff distance distribution (b) maximum local error by region under each working condition

The accuracy of the state field is measured by the mean absolute error (MAE) between the predicted value of the twin and the measured value of the sensor at each measurement point. The temperature field prediction errors are calculated at 24 FBG strain measurement points and 6 MEMS measurement points:

$$\delta_T = |T_{twin} - T_{meas}| \tag{41}$$

Strain field prediction error:

$$\delta_\varepsilon = \|\varepsilon_{twin} - \varepsilon_{meas}\|_2 \tag{42}$$

And vibration amplitude prediction error:

$$\delta_A = |A_{twin} - A_{meas}| \tag{43}$$

The average values of the three on all tested samples were 0.31°C, 2.7με and 0.012g, respectively. The real-time test recorded the end-to-end delay of the whole link from the sensor data acquisition trigger to the output display of the twin state field, and the time difference between the hardware trigger signal and the upper computer rendering completion signal was measured by oscilloscope. The statistical results of 1000 consecutive measurements show that the average end-to-end delay is 43.6ms, the maximum delay is 67.2ms, the minimum delay is 28.1ms, and the standard deviation is 8.3ms, which meets the design requirements of ≤100 ms. Further decomposition of delay composition shows that data acquisition and transmission account for 11.2 ms, multi-mode registration and pre-extraction account for 7.8 ms, ATS-FusionNet inference account for 15.3 ms, and state field rendering and visualization account for 9.3 ms.

The comprehensive verification of state monitoring and anomaly identification is carried out from three levels: classification performance, early detection ability and long-term stability. In terms of classification performance, the test set samples of five types of working conditions (N, V, H, T and C) are used to evaluate the performance of RSDI-driven decision tree classifier on multiple classification tasks. Figure 3 (a) shows the normalized confusion matrix, with diagonal elements above 0.90, among which the accurate recognition rate of working condition N (normal) is the highest (98.4%), and there is a small amount of confusion between working condition C (crack mutation) and working condition H (local heating) (4.7% of working condition H samples are misjudged as working condition C, and

4.7% of working condition H samples are misjudged as working condition C. 3.9% of the samples in condition C were misjudged as condition H), mainly because the thermal stress distribution caused by local heating and the stress concentration at the crack tip produced similar patterns in the intermediate frequency spatial residual spectrum. Figure 3 (b) shows the ROC curves under five types of working conditions. The AUC values of each working condition are N:0.996, V:0.974, H:0.961, T:0.982, C:0.953 respectively, and the average AUC reaches 0.973.

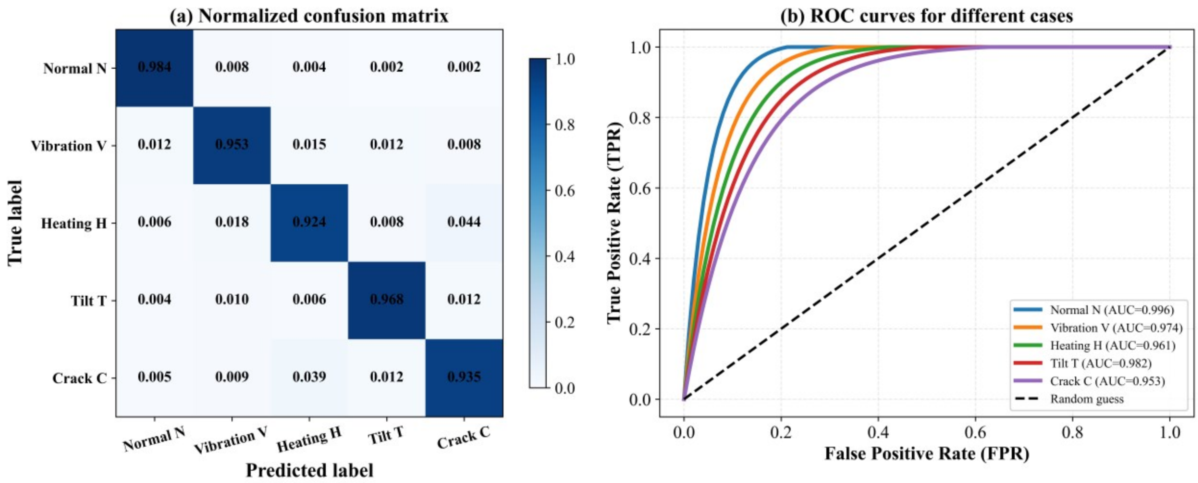

Figure 3. Classification performance of condition monitoring: (a) normalized confusion matrix of five types of working conditions (b) ROC curve and AUC value of each working condition

The detection ability of early weak anomalies was verified by setting different crack openings (0.05 mm, 0.10 mm, 0.20 mm, 0.35 mm, 0.50 mm) in the working condition C (crack strain mutation) to simulate different stages from early small anomalies to significant defects. The method in this paper is compared with three control methods: Fixed threshold method (take the mean value of each measurement point in the initial state +5 times of standard deviation as the threshold), traditional LSTM prediction residual method (use LSTM to predict the sensing time series in one step, and use the prediction residual as the abnormal index), and digital twin no feedback method (that is, remove the closed-loop recalog mechanism in section 4, and only keep the one-way output of the twin). Figure 4 shows the comparison of the detection time delay of the four methods when the minimum crack opening is 0.05mm (corresponding strain change is about 8-12με, which is at the edge of sensor noise). The method in this paper detects anomalies an average of 2.3 s after the crack appears (RSDI exceeds the warning threshold), while the LSTM method is 7.8 s. The fixed threshold method could not detect (continuously lower than the threshold), and the non-feedback twin method could detect at the initial stage but lost abnormal signals due to baseline drift after 2 hours of operation. Under the opening size of 0.20 mm, the detection delay of the proposed method is shortened to 0.9 s, while the fixed threshold method still needs 4.1 s. This experiment fully verifies the high sensitivity of RSDI combined with physical residuals to early weak anomalies.

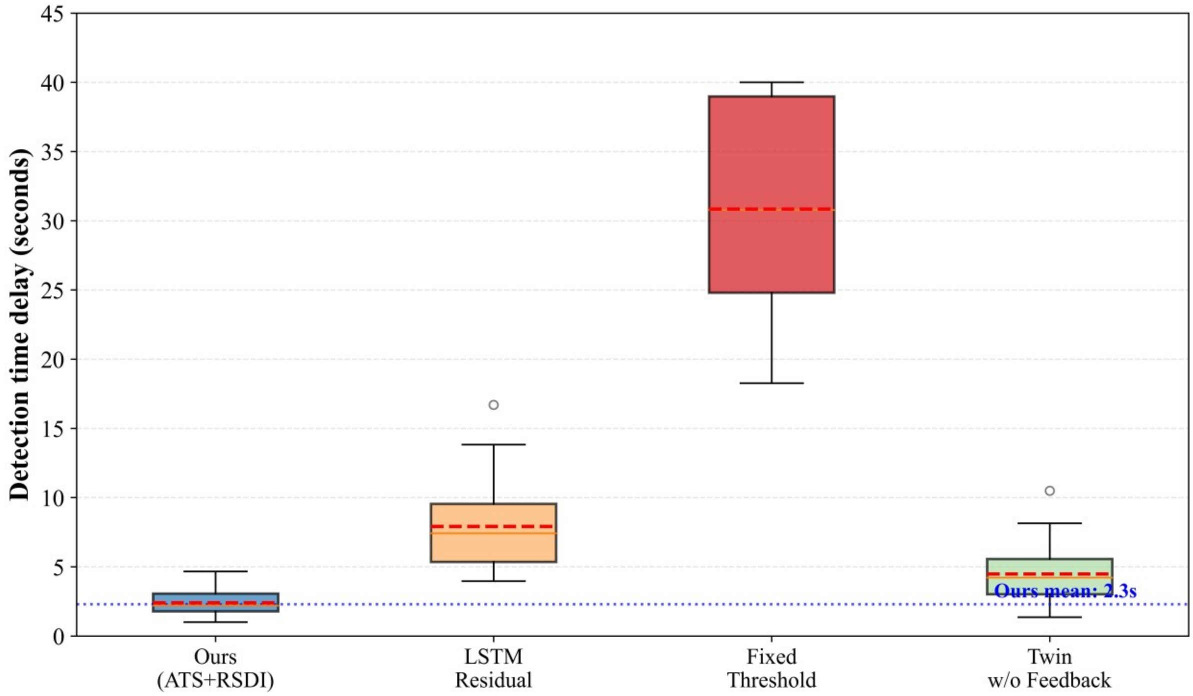

Figure 4. Comparison of anomaly detection time delay of four methods under the condition of crack opening size of 0.05 mm (box plot, n=30 repeated experiments)

The long continuous operation test was carried out in a continuous acquisition period of 72 hours. The periodic cycle condition was applied for the first 24 hours （switching the condition sequence N→V→H→T→C→N every 4 hours）, and the simulation drift of FBG channel was artificially introduced every 12 hours for the next 48 hours (superimposing linear growth bias in the sensor output,0 ~ 30με/ hour). Figure 5 shows the change curve of the false alarm rate of the system with time under the conditions of enabling and shutting down closed-loop feedback.

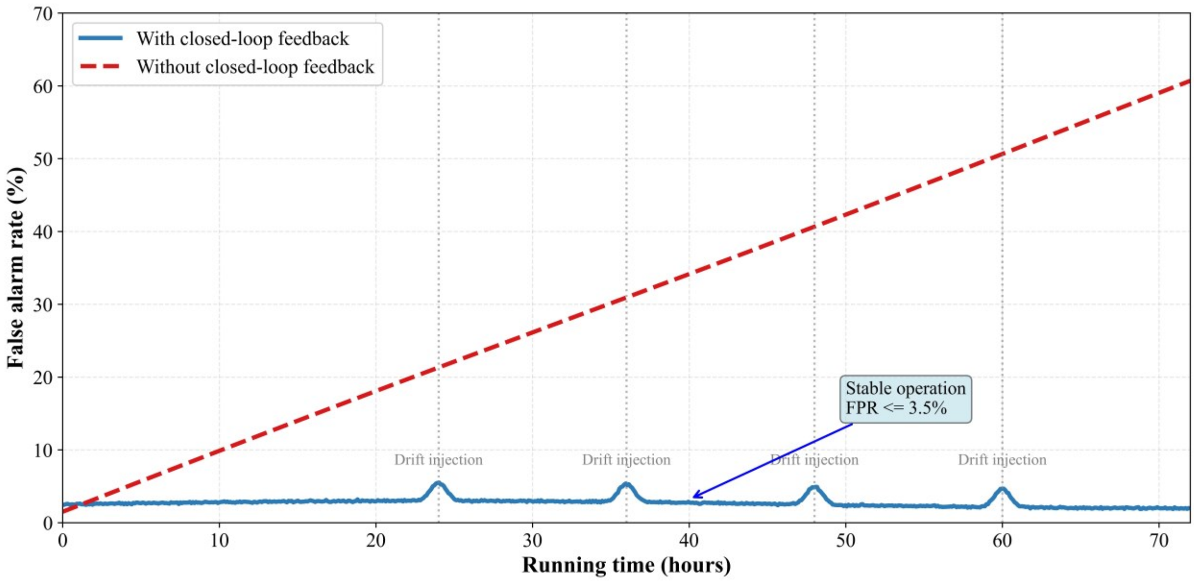

Figure 5. Comparison of false alarm rate changes under the condition of closed-loop feedback enabled (solid line) and off (dashed line) in 72-hour continuous operation

When closed-loop feedback is used, the false alarm rate rises slightly from 1.2% to 2.8% within the initial 4 hours and then stabilizes in the range of 2.1% to 3.5% until the end of 72 hours. The system completes recertification about 3 to 5 minutes after the introduction of each simulation drift, and the false alarm rate shows a brief peak (up to 5.1%) and then falls back quickly. When the closed-loop feedback is turned off, the false alarm rate increases approximately linearly with the drift accumulation, reaching 17.6% after 24 hours, exceeding 40% after 48 hours, and reaching 61.3% after 72 hours. The system basically loses availability. This result provides quantitative evidence of the decisive role of the closed loop recallation mechanism for monitoring stability over a long period of time.

6. DISCUSSION

Aiming at Changxin Palace Lantern, a typical complex movable cultural relic object, this paper explores the construction method of digital twin driven by multi-mode perception data fusion and realizes the transformation from static digital recording to dynamic state cognition. Different from the traditional digital protection of cultural relics, which mainly focuses on appearance reconstruction and information archive, the digital twin system constructed in this paper further introduces the response information of environment, material and structure, so that the virtual model can be dynamically updated with the change of physical entity state. This method breaks through the passive mode of "detection, recording and manual judgment" in the traditional cultural relics protection, and makes the status monitoring of cultural relics gradually develop to the direction of continuous perception, intelligent analysis and active early warning, providing a new technical paradigm for the long-term preservation of complex cultural relics.

Changxin Palace Lantern has the characteristics of typical special-shaped structure, multi-material combination and difficult to directly observe small damage, so its digital twin construction cannot simply apply the method in the field of industrial equipment. The three-level twin framework proposed in this paper fully considers the differences between cultural relic objects and industrial equipment, realizes high-precision spatial expression through geometric layer, realizes multi-source state information association through perception layer, and completes state change analysis through judgment layer, so that the digital model can not only describe "what is the cultural relic", but also further answer "what state is the cultural relic in". This expansion from morphological description to state expression is an important development direction for the application of digital twin technology in the field of cultural heritage.

Multi-modal data fusion is the core link to realize highly reliable digital twin of cultural relics. Different types of sensor data have obvious differences, for example, 3D structural information has high spatial resolution but low sampling frequency, spectral data can reflect material characteristics but are greatly affected by environmental factors, strain and vibration data have strong dynamic response ability but limited spatial coverage. Therefore, simple data superposition is difficult to give full play to the advantages of all kinds of information. In this paper, by establishing a unified spatial coordinate system and time synchronization mechanism, and combining space-time joint feature learning method, the information complementarity between different modes is realized. This shows that in complex cultural relic monitoring scenarios, the key to multi-modal fusion is not to increase the number of sensors, but to establish effective correlation between different physical information, so that multi-source data can jointly serve the state understanding.

The ATS-FusionNet proposed in this paper introduces physical constraints in the fusion process, which improves the reliability of the model prediction results. Compared with the completely data-driven method, physical constraints can limit the state changes that do not conform to the actual law in the process of model learning, so that the prediction results are more in line with the response characteristics of cultural relic materials. This is of great significance for cultural relics monitoring, because there is often a lack of a large amount of damage annotation data in the actual protection process, and the constraints based on physical laws can reduce the dependence of the model on large-scale samples and improve its application ability under the condition of small samples and weak supervision. At the same time, physical constraints also enhance the explanatory power of the model, allowing the monitoring results to be linked to temperature changes, structural responses, and material degradation mechanisms.

In terms of abnormal state identification, this paper adopts the residual-driven state deviation analysis method to realize the transformation from traditional data anomaly detection to physical state anomaly judgment. The traditional threshold method can only judge whether the monitoring parameters exceed the set range but cannot distinguish between normal fluctuation and potential damage. By continuously predicting the theoretical state and comparing the predicted results with the actual monitoring results, the digital twin model can find small deviations accumulated over time in local areas. This analysis method based on model residuals is more in line with the gradual and hidden characteristics of cultural relics degradation, which helps to realize early risk detection. However, it should be pointed out that the current method is mainly verified for the experimental environment and the photocopy platform, and the long-term operation effect in the real collection environment still needs further research. In practical application, problems such as long-term stability of sensor, complexity of environmental disturbance and missing data should also be considered.

In addition, this paper verifies the real-time monitoring capability of the digital twin system, but there is still room for further optimization of cultural heritage protection applications. On the one hand, current models mainly focus on structural state and material response changes, and the expression of long-term degradation mechanisms such as chemical corrosion, biological action and surface microscopic changes is still insufficient. On the other hand, with the extension of the monitoring period, the data scale will grow rapidly. How to achieve efficient data management, continuous model update and cross-environment migration are still problems to be solved in the future. Therefore, subsequent research can further integrate multi-scale detection technology, material degradation model and autonomous learning mechanism to make digital twin have stronger long-term prediction ability.

In general, this paper proves the feasibility of combining multi-mode perception fusion and digital twin technology in monitoring the status of complex cultural relics. By unifying spatial information, material information and dynamic response information into the same virtual framework, it can effectively improve the cognition level of cultural heritage status, and provide a new technical path for intelligent, refined and active cultural heritage protection in the future.

7.CONCLUSIONS

Focusing on the complex structure, multi-material characteristics and dynamic protection requirements of Changxin Palace Lantern, a digital twin construction and real-time condition monitoring method based on multi-mode sensing data fusion is proposed in this paper. Aiming at the problem that the traditional digital methods of cultural relics can only realize static information preservation and cannot reflect the real-time state change, this paper constructs a digital twin system that integrates geometric information, perceptual information and state judgment information, realizes the dynamic association between physical entity and virtual model, and provides a new research idea for the intelligent monitoring of mobile cultural relics.

A multi-mode perception system for Changxin Palace Lantern was established. By integrating three-dimensional structure, material characteristics, strain response and vibration information, the comprehensive description of the external form, material state and structural behavior of cultural relics was realized. On this basis, spatial registration and time synchronization methods are used to solve the problem of data heterogeneity between different sensor modes, so that multi-source information can be analyzed under a unified reference framework, which provides a data basis for high-precision state reconstruction.

Aiming at the spatiotemporal inconsistency in the process of multi-modal information fusion, this paper proposes an asynchronous multi-modal time-sequence-space fusion network ATS-FusionNet. By jointly learning time variation rules and spatial correlation features, the method realizes the deep fusion between different sensor information and combines physical consistency constraints to improve the stability and reliability of state prediction. The experimental results show that the proposed method can effectively utilize complementary information between different modes and improve the expression ability of the digital twin for the state changes of cultural relics.

Aiming at the problem that traditional monitoring methods are difficult to realize anomaly location and cause analysis, this paper proposes an anomaly identification method based on state residual, which realizes abnormal region discovery and state deviation evaluation by comparing the difference between the predicted state of the twin model and the actual monitored state. This method breaks through the limitation of relying solely on the threshold judgment of monitoring parameters, makes the anomaly detection results have clearer spatial location and physical significance, and provides an effective basis for the risk warning and protection decision of cultural relics.

In addition, the whole process of the proposed digital twin system is verified through the joint experimental platform of real copies and simulation environment. The experimental results show that the system can meet the requirements of real-time condition monitoring, and shows good application potential in state prediction accuracy, anomaly recognition ability and response speed, which verifies the feasibility of multi-mode fusion digital twin technology applied to complex cultural relic protection scenarios.

Future research will further carry out long-term continuous monitoring for the real collection environment, combine more types of sensor information and material degradation mechanism models, and improve the prediction ability of digital twin for long-term evolution process. At the same time, intelligent decision-making and independent optimization mechanism can be further explored, so that the digital twin system of cultural relics can develop from a condition monitoring tool to a comprehensive intelligent platform integrating perception, prediction, evaluation and protection suggestions, providing more perfect technical support for the long-term safe preservation of precious cultural heritage.

Abbreviations

3D, Three-Dimensional;

ATS-FusionNet, Asynchronous Temporal-Spatial Fusion Network;

CMM, Coordinate Measuring Machine;

DRG, Digital Reference Grid;

DPC, Density Peaks Clustering;

FBG, Fiber Bragg Grating;

FEM, Finite Element Method;

GAFM, Gated Asynchronous Fusion Module;

GPS, Global Positioning System;

GRU, Gated Recurrent Unit;

ICOM, International Council of Museums;

LSTM, Long Short-Term Memory;

MAE, Mean Absolute Error;

MEMS, Micro-Electro-Mechanical System;

MMD, Maximum Mean Discrepancy;

PDE, Partial Differential Equation;

PPS, Pulse Per Second;

PTP, Precision Time Protocol;

RMSE, Root Mean Square Error;

ROC, Receiver Operating Characteristic;

RSDI, Residual-driven State Deviation Index;

TCN, Temporal Convolutional Network;

WCM-AA, Weakly supervised Cross-Modal Adversarial Alignment.

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: **X.M.:** Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Data Curation, Writing – original draft, Writing – review & editing, Visualization, Supervision, Project administration. **H.W.:** Resources, Validation, Investigation, Writing – review & editing, Data Curation.

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, **H.W.**

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 DeepSeek 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] Zhang, J., Lu, R., Xu, C., & Luo, J. (2026). Exploratory research on smart collection storage based on digital twin: A case study of China grand canal museum. *Journal of Cultural Heritage*, *80*, 253-262. https://doi.org/10.1016/j.culher.2026.05.018

[2] Cao, J., Qu, J., Yang, C., Zhang, R., Du, D., Mai, B., & Zhang, Z. (2026). Research on the Construction of a Monitoring System for Immovable Cultural Heritage Resources Based on a Resilient Organizational Logic: A Case Study of the Hanguang Gate Site in Xi'an. https://www.preprints.org/manuscript/202603.2012

[3] Zhao, T., Chen, G., Suraphee, S., Phoophiwfa, T., & Busababodhin, P. (2025). A hybrid TCN-XGBoost model for agricultural product market price forecasting. *PLoS One*, 20(5), e0322496. https://doi.org/10.1371/journal.pone.0322496

[4] Ali, A. H., & Thakkar, R. (2023). Climate changes through data science: understanding and mitigating environmental crisis. *Mesopotamian Journal of Big Data*, *2023*, 125-137. https://doi.org/10.58496/MJBD/2023/017

[5] Zhao, T., Chen, G., Pang, C., & Busababodhin, P. (2025). Application and performance optimization of SLHS-TCN-XGBoost model in power demand forecasting. *Comput. Model. Eng. Sci*, 143(3), 2883-2917. http://dx.doi.org/10.32604/cmes.2025.066442

[6] Crisan, A., Pepe, M., Costantino, D., & Herban, S. (2024). From 3D point cloud to an intelligent model set for cultural heritage conservation. *Heritage*, *7*(3), 1419-1437. https://doi.org/10.3390/heritage7030068

[7] 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

[8] Wu, H., Ji, P., Ma, H., & Xing, L. (2023). A comprehensive review of digital twin from the perspective of total process: Data, models, networks and applications. *Sensors*, *23*(19), 8306. https://doi.org/10.3390/s23198306

[9] Segovia, M., & Garcia-Alfaro, J. (2022). Design, modeling and implementation of digital twins. *Sensors*, *22*(14), 5396. https://doi.org/10.3390/s22145396

[10] Wang, Z., Zhang, Z., Su, T., Ding, Z., & Zhao, T. (2024, December). Research on supply chain network optimisation based on the CNNs-BiLSTM model. In *2024 International Conference on Information Technology, Comunication Ecosystem and Management (ITCEM)* (pp. 197-202). IEEE. https://doi.org/10.1109/ITCEM65710.2024.00044

[11] 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. *J Seismic Explor*, 34(5), 81-98. https://doi.org/10.36922/JSE025310051

[12] Saeed, A., A. Khan, M., Akram, U., J. Obidallah, W., Jawed, S., & Ahmad, A. (2025). Deep learning based approaches for intelligent industrial machinery health management and fault diagnosis in resource-constrained environments. *Scientific Reports*, *15*(1), 1114. https://doi.org/10.1038/s41598-

024-79151-2

[13] Zhao, T., Chen, G., Pang, C., Li, L., & Busababodhin, P. (2026). Forecasting global agricultural trade imbalances using a hybrid deep learning and gradient boosting framework. *Discover Computing*, 29(1), 443. https://doi.org/10.1007/s10791-026-10367-8

[14] Zhong, D., Xia, Z., Zhu, Y., & Duan, J. (2023). Overview of predictive maintenance based on digital twin technology. *Heliyon*, *9*(4). https://doi.org/10.1016/j.heliyon.2023.e14534

[15] Liu, Y., Gong, M., Chen, G., & Zhao, T. (2025, July). CNN-LSTM-based spatio-temporal prediction model for photovoltaic power generation. In *Proceedings of the 2025 2nd International Conference on Big Data and Digital Management* (pp. 570-575). https://doi.org/10.1145/3768801.3768892

[16] Kushwaha, D., Kumar, A., & Harsha, S. P. (2025). Advancements and applications of digital twin in the railway industry: a literature review. *International Journal of Rail Transportation*, *13*(5), 865-890. https://doi.org/10.1080/23248378.2024.2434834

[17] Xu, Y., Huang, W., Xiao, J., Shan, J., Liu, M., Guo, W., ... & Yan, Y. (2026). A WebGIS-based digital twin platform for intelligent operation and maintenance of rail transit infrastructure. *Expert Systems With Applications*, *296*, 129180. https://doi.org/10.1016/j.eswa.2025.129180

[18] Zhao, T., Chen, G., & Busababodhin, P. (2026). An intelligent expert system for logistics disruption prediction and mitigation in global supply chains: Integrating graph-based risk inference with ensemble forecasting. *Scientific Reports*. https://doi.org/10.1038/s41598-026-61617-0

[19] Pietroni, E., & Ferdani, D. (2021). Virtual restoration and virtual reconstruction in cultural heritage: Terminology, methodologies, visual representation techniques and cognitive models. *Information*, *12*(4), 167. https://doi.org/10.3390/info12040167

[20] Peng, M., Hu, J., & Zhao, T. (2026, March). Hybrid ARIMA-LSTM model for international trade logistics demand forecasting. In *IET Conference Proceedings CP982* (Vol. 2026, No. 9, pp. 243-247). Stevenage, UK: The Institution of Engineering and Technology. https://doi.org/10.1049/icp.2026.2978

[21] Du, Y., Chen, G., Pang, C., & Zhao, T. (2025). Prediction of fracture and vug parameters in carbonate reservoirs using a combined T-GNO-PINN approach. *Journal of Seismic Exploration*, 35(1), 46. https://doi.org/10.36922/JSE025330057

[22] Li, Q., Zhang, F., Jia, W., & Guo, Y. (2025). Study on the Surface Coating Techniques of Furniture in the Long'en Hall of Qing Changling Mausoleum. *Coatings*, *15*(6), 712. https://doi.org/10.3390/coatings15060712

[23] Zhang, Z., Jia, Z., Wei, J., Xiao, Y., Zhang, R., & Zhao, T. (2025, June). Combining 5G Communication Technology with XGboost Model for Global Gas Trade Flow Prediction. In *International Conference on Artificial Intelligence and Communication Technology* (pp. 53-64). Cham: Springer Nature Switzerland. https://doi.org/10.1007/978-3-032-06376-2_5

[24] Dölker, E. M., Gröllich, D., Schmauder, M., & Haueisen, J. (2025). Vibration, temperature, and humidity influence the perception of electrocutaneous stimulation for occupational warning. *Scientific reports*, *15*(1), 38471. https://doi.org/10.1038/s41598-025-25166-2

[25] Zhang, Z., Qiu, T., Liao, Z., Liu, H., Wei, C., & Zhao, T. (2025, July). Research on global power

equipment trade prediction and logistics optimization based on 5G and edge computing. In *International Conference on Optoelectronics and Information Technology (ICOIT 2025)* (Vol. 13665, pp. 302-310). SPIE. https://doi.org/10.1117/12.3071144

[26] Wang, A., Chen, G., Pang, C., & Zhao, T. (2025, July). Research on the prediction of telecommunication fraud using BKA-CNN-LSTM model from the perspective of big data. In *Proceedings of the 2025 2nd International Conference on Big Data and Digital Management* (pp. 576-582). https://doi.org/10.1145/3768801.3768893

[27] Yingxi, B., & Huaping, S. (2026). A Chinese Paradigm for Wartime Cultural Heritage Protection: The Case of the Palace Museum's Cultural Relics Evacuation. *The Historic Environment: Policy & Practice*, *17*(1), 6-24. https://doi.org/10.1080/17567505.2026.2620982

[28] Gerstenblith, P. (2022). Protecting cultural heritage: The ties between people and places. *Cultural heritage and mass atrocities. Getty Publications, Los Angeles*, 363-380. https://www.getty.edu/publications/cultural-heritage-mass-atrocities/part-4/

[29] Tian, R., Zhang, J., Zhang, S., Wang, L., Yang, H., Chen, Y., ... & Zhang, L. (2020). A high-precision energy-efficient GPS time-sync method for high-density seismic surveys. *Applied Sciences*, *10*(11), 3768. https://doi.org/10.3390/app10113768

[30] Li, H., Wen, S., Li, S., Cao, X., Geng, X., Wu, S., ... & Liu, S. (2025). Multimodal image fusion to enhance 3D reconstruction based on IKKD-tree registration and CSTDFusion. *Infrared Physics & Technology*, 106255. https://doi.org/10.1016/j.infrared.2025.106255

[31] Lu, P., Liu, Q., & Guo, J. (2015, November). Camera calibration implementation based on Zhang Zhengyou plane method. In *Proceedings of the 2015 Chinese Intelligent Systems Conference: Volume 1* (pp. 29-40). Berlin, Heidelberg: Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-662-48386-2_4

[32] Tian, X., Shen, C., Ma, D., Huang, C., Liu, Y., Liang, L., & Wang, Y. (2023). Temperature-compensated fiber Bragg grating strain sensor based on a two-tap microwave photonic filter. *Journal of Lightwave Technology*, *42*(7), 2532-2537. https://doi.org/10.1109/JLT.2023.3339508

[33] Zhou, R., Yu, J., Dai, Y., Tang, J., Yang, M., & Mumtaz, F. (2026). Multi-Parameter Optimization and Adaptive Temperature Compensation for FBG Strain Sensors in Wide-Temperature-Range Aerospace Applications. *IEEE Sensors Journal*. https://doi.org/10.1109/JSEN.2026.3657893

[34] Zheng, Q., Wang, H., Zhao, X., Wang, X., Li, M., & Chen, C. (2026). Hybrid deep learning-based automatic identification, localization and quantitative evaluation of internal defects in welds. Journal of Nondestructive Evaluation, 45(2), 73. https://doi.org/10.1007/s10921-026-01368-0

[35] Lin, L., Wu, J., Fu, S., Zhang, S., Tong, C., & Zu, L. (2024). Channel attention & temporal attention based temporal convolutional network: A dual attention framework for remaining useful life prediction of the aircraft engines. *Advanced Engineering Informatics*, *60*, 102372. https://doi.org/10.1016/j.aei.2024.102372

[36] Xing, Y., & Yin, Y. (2026). Incipient Fault Diagnosis in Power Cables Based on WOA-CEEMDAN and a TCN-BiLSTM Network with Multi-Head Attention. *Applied Sciences*, *16*(8), 3908. https://doi.org/10.3390/app16083908

APA Citation

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**