

Immersive Reconstruction of Jinhua Intangible Cultural Heritage Space Based on NeRF and Panoramic Image Stitching

Xueyuan Fang¹

¹College of Information, Zhejiang Guangsha Vocational and Technical University of Construction, Dongyang 322100, Zhejiang, China

Corresponding author: Xueyuan Fang (e-mail: 15057988708@163.com)

ABSTRACT Accurate three-dimensional reconstruction from limited image data and inconsistent geometry remains a central challenge in cultural-heritage digitization and engineering scene modeling. This problem is also relevant to electromagnetic measurement environments, where sparse-view reconstruction can support virtual testing of antenna sites, propagation scenarios, and complex indoor scenes. To solve this issue, this paper proposes an immersive reconstruction method integrating Neural Radiance Fields and deep panoramic image stitching. The method uses the continuous modeling capability of NeRF for geometric restoration and panoramic stitching to enhance texture integrity. Its implementation includes deep-guided multi-frame stitching, continuous geometric modeling and color fitting through NeRF, and texture-stitching optimization using the PatchMatch algorithm. The process shows strong adaptability under limited image input. Experimental results confirm that the NeRF fusion model maintains stable reconstruction even with only four input frames, reaching an SSIM of 0.65 and a chamfer distance of 0.45. It also achieves good texture continuity in challenging areas, with LPIPS scores not exceeding 0.151. The system provides a strong immersive experience, with a distortion score above 0.9, while maintaining practical efficiency at high resolution. These results provide a useful benchmark for high-fidelity geometric and textural reconstruction in heritage spaces and related electromagnetic scene-digitalization tasks.

INDEX TERMS neural radiance fields, panoramic stitching, three-dimensional reconstruction, electromagnetic scene modeling, texture fusion

I. INTRODUCTION

JINHUA intangible cultural heritage space carries rich cultural value, and its three-dimensional digital reconstruction [1,2] provides an innovative way for cultural protection and dissemination. This digital approach to preserving complex, high-value assets is increasingly vital across many sectors. Similarly, applying advanced 3D digital reconstruction techniques can provide an innovative way for cultural preservation and knowledge transfer in the industry, particularly for documenting the rich cultural value embodied in historical machinery, traditional looms, and intricate weaving patterns. The complex structure and diverse texture characteristics of the intangible cultural heritage space require high-precision geometric restoration and delicate texture reconstruction to ensure the immersion and authenticity of the virtual display. The demand for accurate capture of complex structures is critical for digital applications across sectors. This methodology offers a solution for the industry to create high-fidelity 3D models of machinery or fabrics, ensuring that virtual displays achieve the necessary level of immersion for effective training, design, or historical

preservation [3,4]. In the process of three-dimensional reconstruction of Jinhua intangible cultural heritage space, the core challenges faced are mainly concentrated on the contradiction between the integrity of data acquisition and the precision of the reconstruction model. On the one hand, the conditions for onsite acquisition are limited and the image viewing angles are sparse, resulting in missing or discontinuous spatial geometric information in the reconstruction [5,6]. On the other hand, the texture mapping process is difficult to balance detail restoration and overall consistency, especially in complex lighting and multi-material surface areas. Texture distortion and stitching gaps are obvious, affecting visual realism [7,8]. In addition, when dealing with complex structures, single reconstruction technology often faces the problem of high computing resource consumption and insufficient efficiency, which limits the promotion of the model in practical applications [9,10].

II. RELATED WORKS

Scholars have conducted extensive research on the three-dimensional reconstruction of intangible cultural heritage

spaces. Tan Y X proposed a three-dimensional reconstruction method based on MVSNet (multiview stereo network) to achieve high-precision temperature field reconstruction of equipment in thermal infrared images. This method has high efficiency and accuracy advantages in texture enhancement, key frame selection and depth estimation, and improves the reconstruction performance in complex scenes [11]. MVSNet uses deep learning to improve the accuracy of multi-view stereo matching and enhance the recovery of geometric details, but it has shortcomings when dealing with texture discontinuities and complex structures [12,13]. Lee B K used COLMAP to reconstruct three-dimensional images of traditional Korean stone pagodas and verified its effectiveness in high-precision modeling. Studies have shown that this method has good application potential in cultural heritage restoration [14]. COLMAP, as a classic method based on structured beam adjustment, has strong dense reconstruction capabilities, but it is highly dependent on the acquisition perspective and lighting conditions, and texture stitching is prone to gaps [15,16]. PMVS (Patch-based Multi-View Stereo) achieves three-dimensional reconstruction by gradually expanding sparse point clouds. It is suitable for a variety of scenes, but it is relatively weak in capturing high-frequency texture information. These studies have made significant progress in improving spatial structure restoration. However, in irregular cultural spaces, the complete continuity of image textures and the immersive experience of perspective switching are still lacking [17,18]. In order to solve the problems of texture loss and visual incoherence, many studies have tried to use image stitching and depth-guided techniques to optimize local area performance, but the robustness and generalization ability of the overall solution need to be improved.

III. METHODS

Figure 1 shows the NeRF and panoramic image stitching fusion architecture, the core of which is to integrate geometric accuracy and texture integrity. The system achieves spatiotemporal alignment through the sequential image acquisition module synchronized by the IMU (Inertial Measurement Unit). Then, the depthguided panoramic stitching technology is used to construct a highly consistent texture base, the depth estimation network analyzes the multi-level structure of the intangible cultural heritage scene, and the multiview weighted fusion algorithm processes the material mutation area. The NeRF modeling layer realizes geometric reconstruction of continuous perspectives through volume rendering, and its implicit representation effectively overcomes the topological errors of traditional methods in complex carved structures. The final texture optimization module combines back projection and PatchMatch algorithm to ensure the accurate mapping of cultural symbols such as patterns.

A. DYNAMIC ACQUISITION AND PREPROCESSING OF VIEW SEQUENCE IMAGES

Based on lightweight trajectory equipment, sequence images are collected in the indoor and outdoor spaces of the intangible cultural heritage, and frame synchronization is performed by combining IMU trajectory and timestamp data; ORB (Oriented FAST and Rotated BRIEF) features are used for coarse registration, blur and redundant frame images are removed, and input image sets are generated.

B. VIEW SEQUENCE IMAGE ACQUISITION AND TRAJECTORY SYNCHRONIZATION TECHNOLOGY

The scene structure in the intangible cultural heritage space is changeable and the lighting is complex, which leads to the problem of sparse viewing angle and unstable frame in the image acquisition process [19,20]. In this stage, an integrated lightweight trajectory capture module is used to complete the image data acquisition task. The module is equipped with an inertial measurement unit (IMU), which outputs high-frequency accelerometer and gyroscope data during the actual movement process, and performs time-series correlation with the frame information collected by the image sensor to generate a synchronous data stream with image sequence and trajectory alignment. The trajectory alignment process uses linear interpolation combined with a sliding time window strategy to map the attitude information recorded by the IMU in the time domain to ensure a stable matching relationship between the image frame and the position state. In actual operation, the acquisition time of each frame is marked as t_i , and the six-axis data sequence provided by the IMU is represented as $\{a(t), \omega(t)\}$, where $a(t)$ is the three-axis acceleration and $\omega(t)$ is the angular velocity. After using the median filter to suppress the oscillation interference, according to the attitude integration model:

$$R(t_i) = R(t_{i-1}) \cdot \exp(\Delta t \cdot \omega^\wedge(t_{i-1})) \quad (1)$$

The rotation matrix estimation between consecutive frames can be completed. Among them, $R(t_i)$ represents the current frame posture rotation matrix, Δt is the time difference between adjacent frames, and $\omega^\wedge$ is the angular velocity vector expression in the form of an antisymmetric matrix. This step can effectively organize the image frames into a unified time sequence trajectory and reduce the reconstruction distortion problem caused by the posture drift between images.

C. IMAGE FRAME SCREENING AND INITIAL REGISTRATION CONSTRUCTION

The original image sequence contains some blurred frames, repeated frames, and frames with severe motion blur. If they are directly entered into the modeling stage without processing, it is easy to cause feature matching failure and texture misalignment problems. To this end, a quality evaluation mechanism based on image gradient energy distribution is constructed to eliminate images with insufficient

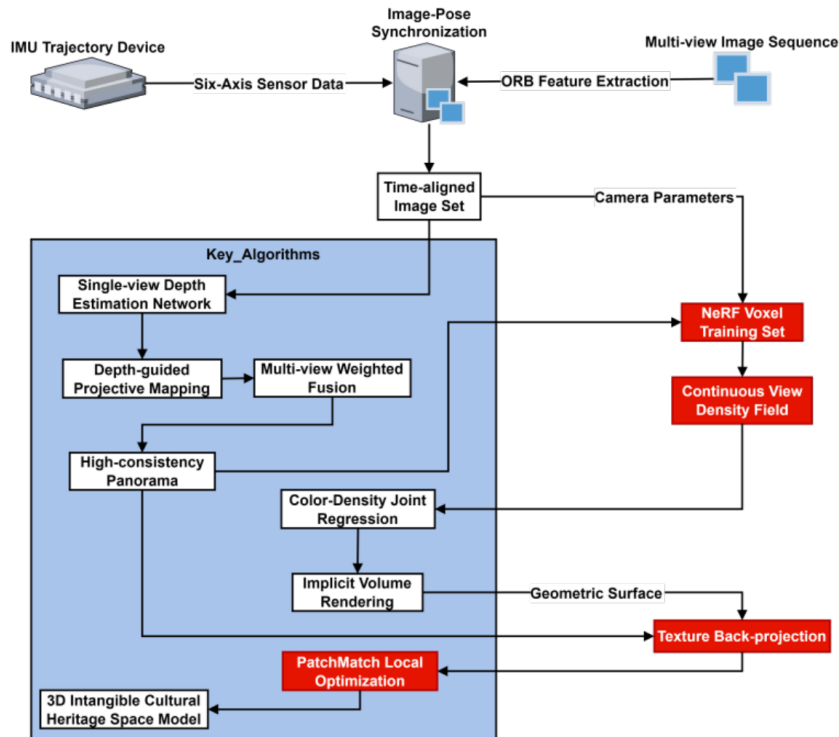


FIGURE 1. NeRF and panoramic image stitching fusion architecture

information. After graying each frame of the image, its gradient modulus map is calculated, and Gaussian filtering is applied to smooth the interference in the local area. The image texture intensity is identified by setting a response threshold, and frames with global gradient amplitudes below a given threshold are removed. The filtering ratio is controlled within 10% to ensure the stability of the collected data.

The initial registration accuracy directly determines the quality of panoramic stitching. In order to improve the registration accuracy, a global consistency evaluation index is introduced for matching verification. Assume that the ORB matching point set of the current image frame and the adjacent frame is M_{ij} , and the transformation matrix estimation adopts the minimum reprojection error model, the essential matrix between two frames is fitted using the RANSAC (Random Sample Consistency) mechanism, and erroneous matching pairs are eliminated. This process screens out low-quality matching results and provides a more stable image infrastructure for the subsequent deep stitching stage.

Figure 2 shows the key data features and their changing trends in the acquisition of intangible cultural heritage space image sequences. Figure 2a reflects the change of IMU attitude estimation error with the number of acquisition frames under different filtering strength conditions. The overall error shows a downward trend, indicating that filtering plays a positive role in reducing sensor noise and integral drift, and improves the stability of trajectory estimation.

The higher the filtering intensity, the more obvious the error attenuation. The estimation error of high filtering stabilizes below 1 degree in the later stage, showing a better attitude estimation effect, reflecting the importance of filtering parameter adjustment in dynamic acquisition. Figure 2b reveals the effect of the time interval between image frames on the number of ORB feature matches.

D. DEPTH-GUIDED PANORAMIC IMAGE STITCHING CONSTRUCTION

This paper uses a pre-trained depth estimation network to predict a single-frame depth map, and performs forward projection registration at the pixel level based on the depth map; the multi-view alignment results are fused to construct a local high-consistency depth panorama for subsequent texture fusion.

E. SINGLE-FRAME DEPTH MAP PREDICTION AND FORWARD MAPPING RECONSTRUCTION

This paper uses DPT (Dense Prediction Transformer) as the depth estimation network with an input resolution of 384×384 . It has been pre-trained in general indoor and architectural scenes and has strong generalization ability for complex geometric structures and weak texture areas. It is suitable for deep inference of high-detail structures such as carvings, beams and columns in intangible cultural heritage spaces.

After the image is input, the network outputs a depth matrix $D_i(u,v)$ with the same size as the original image, where (u,v) is the image pixel position coordinates, and

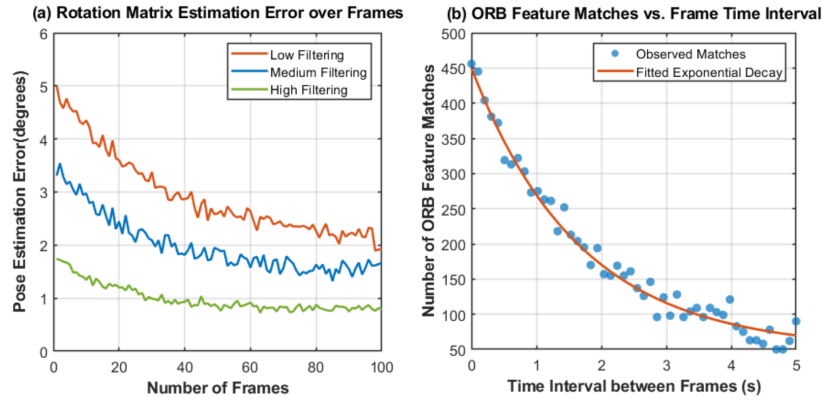


FIGURE 2. Key data features and their changing trends in image sequence acquisition

D_i represents the depth value set corresponding to the i -th frame image. The depth map is used to estimate the position of each pixel in three-dimensional space, and then combined with the trajectory information, it is converted to the reference coordinate system to complete the projection mapping. Set the camera intrinsic matrix to K , and the position of the current pixel in the camera coordinate system is:

$$P_i(u, v) = D_i(u, v) \cdot K^{-1}[u, v, 1]^T \quad (2)$$

$P_i(u, v)$ represents the spatial coordinate point of the pixel in the current frame, and K^{-1} is the inverse of the intrinsic matrix, which is used to restore the true direction vector of the pixel. This projection construction maps the image point by point into three-dimensional space, avoiding the problem of pixel misalignment due to different perspectives in traditional methods, and is particularly suitable for spatial areas with hierarchical structures or occlusion changes. After the spatial coordinate points are generated, they are converted to a common reference perspective by the posture matrix to form a depth point cloud with consistent perspective.

F. MULTI-VIEW ALIGNMENT FUSION AND DEPTH CONSISTENCY SCREENING

After the initial point clouds of each frame are aligned in the common view, it is necessary to solve the overlapping conflicts, boundary misalignment and uneven brightness problems caused by the image fusion between multiple views. In order to ensure the local consistency and global continuity of the splicing, the fusion weight function is introduced in this stage to control the image superposition process, and the unstable area is eliminated in combination with the depth compatibility detection.

Each pixel may have multiple candidate depth values in multiple views, and weighted average fusion is performed on them. Assuming that the image pixel mapped in the common viewing angle of the i -th frame is $I'_i(u, v)$, the fused image output is:

$$I_f(u, v) = \frac{\sum_{i=1}^n w_i(u, v) \cdot I'_i(u, v)}{\sum_{i=1}^n w_i(u, v)} \quad (3)$$

$w_i(u, v)$ is the fusion weight of the pixel position in the i -th frame. The weight calculation introduces depth difference and image gradient as reference factors, and applies lower weights to high-frequency edge areas and areas with drastic depth changes to avoid texture discontinuity problems. Image gradient is used to constrain edge splicing smoothness, and depth difference controls fusion offset within the parallax range to improve image stability and clarity after multi-frame superposition.

During the fusion process, if the difference in the depth value corresponding to a certain pixel under different viewing angles exceeds the set threshold, it is regarded as a disparity inconsistency point, and the splicing participating value at this position is automatically removed, and the filling method is to use the neighborhood block with the smallest gradient difference among adjacent valid blocks. This strategy can effectively eliminate edge tearing and splicing gap problems without relying on the support of large-scale geometric models, and improve local continuity under the guidance of depth maps.

G. GEOMETRIC RECONSTRUCTION MODELING OF NEURAL RADIATION FIELD

Guided by the stitched panorama, it can be fed into the NeRF network frame by frame for perspective modeling; by training the light color function to fit the density and color information of different perspectives in the scene, the three-dimensional spatial volume structure is restored.

H. PERSPECTIVE FRAME-BY-FRAME INPUT AND NERF MODEL TRAINING PROCESS

For the highly consistent image sequence generated by the stitching of the panorama, a neural radiation field model is constructed to achieve continuous perspective modeling of the scene. This study employs the original NeRF architecture as the base radiance field model. The network comprises two MLP layers (256 hidden units each), with position encoding dimensions of 10 (spatial coordinates) and

4 (view directions). Each ray is sampled at 64 points in the coarse network and 128 points in the fine network. The loss function is the mean squared error of RGB pixels. The Adam optimizer is used with an initial learning rate of $5e-4$, batch size of 1024, and a total training iteration of 200,000 steps. First, each frame of the spliced image is used as input, combined with the corresponding camera internal and external parameter information to form a training sample with an accurate projection relationship. During the training process, the neural network densely scans the space by sampling rays, associates the color and density information of the light in the volume space, and gradually approaches the spatial structure of the real scene.

Specifically, select the pixel in the camera's view and emit the corresponding light $r(t)=o+td$ from the camera center outward, where o represents the camera's optical center position, d is the light direction, and t is the ray parameter. The neural network model defines a continuous function mapping of color and density:

$$(\sigma, c) = F_{\theta}(r(t), d) \quad (4)$$

σ represents volume density, $c = (r, g, b)$ represents the color value at the corresponding position, and the function F_{θ} is implemented by a deep neural network represented by the parameter set θ . By accumulating the color and density of multiple sampling points on the ray, the final color of the ray is inferred, and then the pixel value of the input image is fitted.

The synthesis of ray color adopts the volume rendering formulas:

$$\hat{C}(r) = \int_{t_n}^{t_f} T(t)\sigma(t)c(t) dt \quad (5)$$

$$T(t) = \exp\left(-\int_{t_n}^t \sigma(s) ds\right) \quad (6)$$

$T(t)$ is the transmittance of the ray from the near end t_n to the position, indicating the probability that the light is not blocked. $\hat{C}(r)$ is the color estimate obtained by rendering. This mechanism combines the perspective, spatial density and color in the same model to ensure the synchronous reconstruction of illumination and geometric information under continuous perspective.

I. SPATIAL VOLUME STRUCTURE RESTORATION AND MODEL OPTIMIZATION STRATEGY

After the model training converges, the neural network completes the continuous estimation of the density and color of the three-dimensional space, realizing the volume representation of the scene. The volume model can reproduce the real image effect at the training perspective, and also supports free rendering of any new perspective, which greatly enriches the interactive possibilities of the immersive experience. In addition, the regularization term is used to constrain the density function to suppress the generation

of non-physical noise and prevent overfitting. During the parameter optimization process, a joint evaluation is taken for color consistency and geometric coherence to ensure the naturalness and spatial continuity of the model rendering output. The mosaic image can be integrated as a guiding input to enhance the richness of texture information in the training samples and reduce the reconstruction holes and detail blur caused by texture loss in the model.

J. SPATIAL TEXTURE REFINEMENT AND SCENE VOXEL FUSION

Based on the spatial voxels generated by NeRF, the panoramic texture is remapped to the 3D mesh surface; image consistency detection is used to remove unstable texture areas, and the PatchMatch algorithm is combined to optimize the local texture mosaic transition to achieve overall texture refinement.

K. PANORAMIC TEXTURE REMAPPING AND IMAGE CONSISTENCY DETECTION MECHANISM

The spatial voxel model generated by the neural radiation field provides a continuous and delicate geometric basis for the three-dimensional scene. In order to enhance the visual expression, it is necessary to map the high-quality panoramic texture to the three-dimensional mesh surface to ensure the rich presentation of spatial details. The process first accurately locates the pixels of the two-dimensional panoramic image to the surface coordinates corresponding to the three-dimensional voxels through the perspective projection transformation.

Texture mapping uses a reverse projection strategy to calculate the projection position of the 3D mesh vertices in the panoramic image. The mapping formula is defined as:

$$(u, v) = \pi(K, R, p_y, X) \quad (7)$$

X represents the three-dimensional vertex coordinates, π represents the standard perspective projection function, R and p_y represent the camera rotation matrix and translation vector respectively, and (u, v) is the two-dimensional pixel coordinates after projection. This mapping ensures the spatial correspondence between texture information and geometric models, avoiding the common texture misalignment phenomenon.

After the mapping is completed, an image consistency detection mechanism is constructed for the overlapping areas of textures at different perspectives. This mechanism determines the texture stability through the color similarity and gradient information of the local block, and calculates the local consistency index C_{cons} :

$$C_{\text{cons}} = \frac{1}{N} \sum_{i=1}^N \left(\alpha \|I_i - \bar{I}\|_2 + \beta \|\nabla I_i - \nabla \bar{I}\|_2 \right) \quad (8)$$

I_i represents texture blocks at different viewing angles, $\bar{I}$ represents their mean images, ∇ represents image gradient operators, α and β adjust the color and gradient contribution weights respectively, and N is the number of viewing angles compared. In this configuration, $\alpha = 0.6$ and $\beta = 0.4$, the optimization of the weighted sum of LPIPS and edge consistency index on the validation set was achieved through grid search in the pre-experiment, which balanced color fidelity and structural continuity. This indicator is used to remove texture blocks with significant deviations in color or edge information, effectively filter out texture instability areas caused by lighting changes or occlusion, and avoid visual artifacts in stitching.

L. APPLICATION OF PATCHMATCH ALGORITHM AND LOCAL TEXTURE OPTIMIZATION

The stable texture area selected by consistency test uses the PatchMatch algorithm to perform local texture splicing optimization to improve the smoothness of boundary transition and the naturalness of detail connection. The standard PatchMatch configuration is used in this paper: the image block size is 7×7 pixels, the iteration times is 5 rounds, the search radius is 20 pixels, and the neighborhood propagation step is 1 pixel. The algorithm uses fast randomized neighborhood search to find the optimal texture block match and reduce splicing gaps and dislocations. The PatchMatch algorithm in this phase operates on the mapped results of panoramic texture maps on 3D mesh surfaces, optimizing local consistency at multi-view texture stitching boundaries, without participating in NeRF radiance field or depth map optimization.

The specific implementation process is defined as finding the transformation matrix T to maximize the similarity between the source texture block P_s and the target texture block P_t . The optimization objective function is:

$$T^* = \arg \min_T \sum_{x \in \Omega} \|P_s(x) - P_t(T(x))\|^2 \quad (9)$$

Ω represents the set of texture block pixel coordinates, x represents a single pixel position, and $T(x)$ represents the corresponding point after the pixel position is transformed, this transformation is a local pixel offset and does not involve affine or global geometric transformations. PatchMatch uses the initial guess in the neighborhood and quickly converges to the best match through iterative random search and propagation process, significantly reducing the computational complexity of traditional matching methods.

This optimization process focuses on solving the color fault and structural incoherence of texture splicing boundaries, especially for irregular surfaces and multi-material boundary areas. By dynamically adjusting the position and shape of texture blocks, smooth transition of local colors is achieved, eliminating visual discomfort caused by multi-view texture differences.

IV. METHOD EFFECT EVALUATION

A. EXPERIMENTAL DATA AND SETTINGS

The experimental data mainly comes from the field collection of Jinhua intangible cultural heritage space, covering a variety of complex indoor and outdoor scenes, including representative cultural spaces such as the inner hall of ancient buildings and handicraft exhibition areas. The collection process uses a lightweight trajectory device with integrated IMU to ensure the synchronization of image and posture data. The input image sequence is preprocessed, including blurry frame removal and ORB feature matching, to construct high-quality initial registration data. To verify the robustness of the model under sparse input, the experiment sets input conditions with different numbers of frames (4, 8, 16, 32, and 64 frames) to simulate the limited viewing angle in actual acquisition. Figure 3 shows some of the iconic buildings in Jinhua, from left to right: Zhuge Bagua Village, Guzi City, and Wanfo Pagoda.

The evaluation compared the performance of the NeRF fusion model with traditional methods (MVSNet, COLMAP, PMVS) and the original NeRF model in terms of geometric reconstruction, texture continuity and interactive experience. Geometric accuracy was measured by SSIM and Chamfer distance, texture continuity was evaluated by LPIPS and edge consistency index, and immersive experience was quantified by frame rate standard deviation and distortion perception score. In addition, the experiment also tested the computational efficiency of the model at different resolutions (640×480 to 1920×1080) to verify its scalability. The experiment covers 6 independent scenes (4 indoor and 2 outdoor), each scene is captured 64 frames, and divided into training set and test set randomly according to 8:2 ratio.

Table 1 lists in detail the core parameter configuration of trajectory synchronization in the dynamic acquisition phase, including the IMU sampling frequency, image acquisition frame rate, time window length, attitude integration time interval, median filter window size, and attitude reset interval. These parameters ensure high-precision time alignment of attitude information and image data, and reduce the registration deviation caused by time error. The IMU collects six-axis data at a frequency of 200 Hz to ensure the capture of detailed attitude estimation. The image frame rate is maintained at 30fps to achieve a reasonable match between visual information and inertial data. The sliding time window is set to 100 milliseconds to effectively filter out measurement noise and improve the stability of attitude interpolation. The attitude integration time difference is about 33 milliseconds, reflecting the temporal resolution of image acquisition. The median filter window size is 5 sampling points, which is used to smooth the high-frequency interference in the angular velocity data. The trajectory attitude reset is performed every 10 meters to reduce the accumulation of IMU integral drift and ensure the overall path positioning accuracy.



FIGURE 3. Landmark buildings in Jinhua, Zhejiang

TABLE 1. Trajectory synchronization parameter settings

Parameter Settings	Value / Type	Description
IMU Sampling Frequency	200 Hz	Ensures high-frequency attitude output
Image Acquisition Frame Rate	30 fps	Basis for timestamp alignment with IMU
Time Window Length (Sliding Window)	100 ms	Used for temporal interpolation and noise filtering
Attitude Integration Time Interval	Approx. 33 ms	Time difference between adjacent image frames
Median Filter Window Size	5 samples	Used to smooth angular velocity data
Attitude Reset Interval	Every 10 meters	Reduces cumulative error during trajectory

B. TEMPORAL DYNAMIC CHARACTERISTICS OF IMU SIGNALS, RANSAC REPROJECTION ERROR AND ADAPTIVE THRESHOLD

The original acceleration and angular velocity data recorded by the IMU sensor during the acquisition process are analyzed in time series. After the signal is filtered to remove high-frequency interference, the linear acceleration and angular velocity changes of the three axes are extracted respectively, and the component curves are drawn. To enhance the contrast, amplitude differences and disturbance terms were added to the data in the simulation to reflect the intensity of the device's movement in different dimensions in space, revealing the typical characteristics of violent displacement disturbances and smooth posture rotation during acquisition. The adaptive threshold is adjusted dynamically according to the entropy of local inter-frame feature distribution and the sliding statistics of the reprojected residual, and the non-polynomial fitting is used.

Figure 4 shows the time trend of the three-axis acceleration and three-axis angular velocity recorded by the IMU sensor in the acquisition of intangible cultural heritage space images. Time (seconds) is the horizontal axis, and the vertical axes are acceleration (unit: g) and angular velocity (unit: radians/second). The X-axis, Yaxis and Z-axis signals are distinguished by different colors to facilitate the identification and comparison of the three-axis dynamics. The X-axis signal has obvious alternations of positive and negative peaks, and the Yaxis shows a periodic oscillation with a high amplitude. The Z-axis is slightly flat but still maintains ir-

regularity, and the values of the three axes fluctuate between -0.6g and 0.6g as a whole. This large-scale, nonlinear fluctuation indicates that during mobile acquisition, the device is affected by various disturbances such as ground feedback and hand-held shaking, and subsequent filtering and median correction processing must be performed. The change trend of the angular velocity signal is obviously more stable. The overall three-axis angular velocity curve is dominated by low-frequency, low-amplitude oscillations, especially the Z-axis curve, which is basically maintained within ± 0.05 rad/s, indicating that the device attitude adjustment speed is relatively slow. There is a certain periodic fluctuation in the Y-axis rotation, but the amplitude is also much lower than the acceleration signal. This signal feature helps to achieve continuous interpolation of posture between image frames and is an ideal basis for building a high-precision trajectory model. The acceleration signal reflects the strong dynamic changes in the acquisition path and is a sensitive indicator for detecting instantaneous impact and vibration interference; while the angular velocity signal is more stable and lownoise, suitable for posture estimation and perspective alignment. The combination of the two can support the design of the synchronization mechanism between the image and the trajectory, and improve the accuracy of frame organization and alignment.

Figure 5 shows the reprojection error distribution and threshold adaptive strategy of the RANSAC algorithm in 200-frame image registration. The horizontal axis represents the image frame sequence number, the vertical axis is the pixel distance of the error in each frame image estimation, and the column corresponds to the maximum error value of each frame in the minimum reprojection model fitting process. The errors of most frames are concentrated below 0.8 pixels, indicating that the registration is generally stable. Some frames have pixel deviations due to uneven feature distribution or initial posture deviation, and the columns are obviously prominent, which may cause local mismatching risks. The red curve represents the error rejection threshold that is dynamically adjusted at different frame positions. This threshold is generated by combining the complexity of the inter-frame structure and the range of local view angle changes, showing slight fluctuations, and the change range is approximately [0.6, 1.0].

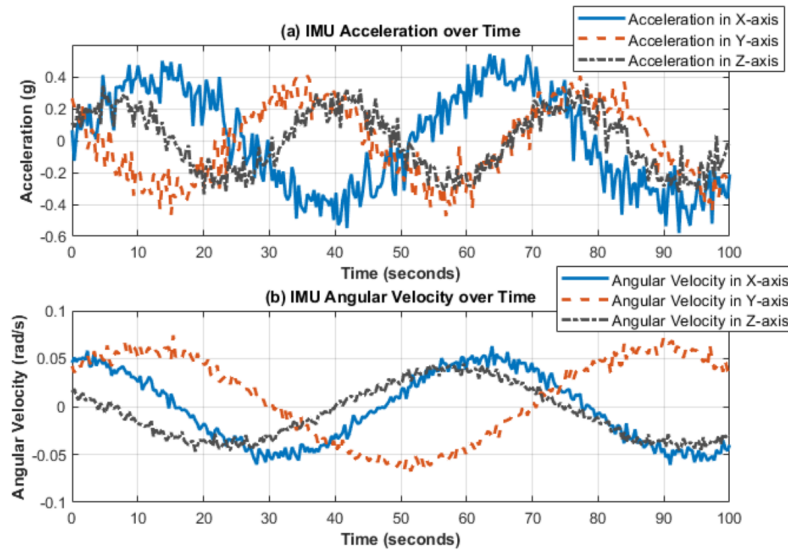


FIGURE 4. Time dynamic characteristics of IMU signals

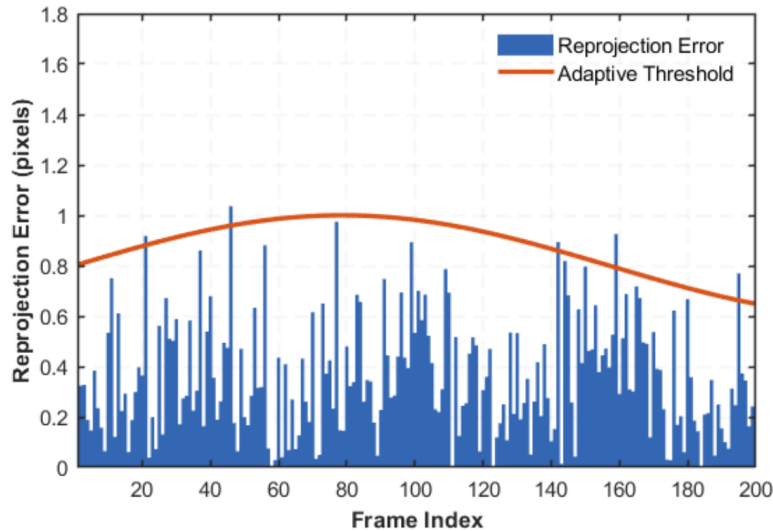


FIGURE 5. RANSAC reprojection error and adaptive threshold

C. SINGLE-FRAME DEPTH MAP 3D POINT CLOUD PROJECTION

This paper performs depth estimation on a single-frame image. The data is generated by deep network estimation from a single frame image of the inner hall of the ancient architecture in Jinhua. First, a pre-trained depth estimation network is used to generate corresponding depth values for each pixel of the input image to form a two-dimensional depth matrix. Then, combined with the camera intrinsic parameter matrix, the two-dimensional coordinates and depth information of each pixel are converted into three-dimensional spatial coordinates to construct a complete set of spatial points. After certain noise processing and filtering, the point set retains the detailed characteristics of the spatial structure. Finally, a three-dimensional scatter plot is used to visualize these spatial coordinate points in a color mapping

manner, where the color corresponds to the size of the depth value, showing the layering and texture changes of the space. This method effectively restores the spatial form of the scene, ensures the geometric correspondence between depth data and image pixels, and lays a solid foundation for multi-view registration and fusion.

Figure 6 shows the distribution of the 3D point cloud generated by the projection of a single-frame depth map, reflecting the actual position and depth relationship of the image pixels in space. The X-axis and Y-axis represent the horizontal plane coordinates of the space, simulating the plane projection position of the image pixels in space; the Z-axis represents the depth value of the corresponding pixel, that is, the distance from the point to the sensor. The undulations and fluctuations of the surface reflect the complex structure and detail differences of the object surface, and can

reveal the layering of texture and geometric information. The local changes in depth reveal the irregularity of the spatial form and reflect the diversity of the surface form and posture of the object during the acquisition process. The point cloud color coding enhances the intuitive perception of depth information and assists in analyzing spatial continuity and edge features. The overall depth value is between 1 meter and 5 meters.

D. COMPARISON OF GEOMETRIC RECONSTRUCTION STRUCTURE ACCURACY

In multiple intangible cultural heritage space samples (such as the inner hall of ancient buildings and the handicraft exhibition area), this paper evaluates the geometric reconstruction structure accuracy of the NeRF fusion model and MVSNet, COLMAP, PMVS and the original NeRF model under different image input densities (4 frames, 8 frames, 16 frames, 32 frames, 64 frames). The selection of 4 frames as the minimum input baseline stems from typical constraints observed during fieldwork at Jinhua's intangible cultural heritage sites: a single round trip typically yields only 3-5 unobstructed key viewpoints. Experience shows that fewer than 4 frames cause traditional NeRF models to develop geometric artifacts and voids due to insufficient cross-line-of-sight constraints, while exceeding 8 frames significantly increases data acquisition workload. Our depthguided stitching method constructs dense, globally consistent texture prior maps, providing NeRF with stable geometric anchors and continuous surface constraints. This effectively mitigates density field ambiguity caused by insufficient light coverage under sparse viewpoints, thereby alleviating typical artifacts. The quantitative evaluation employs a dual-indicator strategy: Chamfer Distance directly measures the Euclidean deviation between the reconstructed point cloud and the real geometry, serving as the core metric for geometric accuracy; while SSIM assesses the luminance consistency between the rendered view and the real image in terms of structural layout, indirectly reflecting the rationality of geometric topology. The actual geometric model was acquired using a high-precision laser scanner (FARO Focus S350), with a point cloud density of ≥ 200 points/m² (unit: mm). The chamfer distance was measured in mm, and the bidirectional nearest neighbor average distance between the reconstructed point cloud and the original geometric model's point cloud was calculated. The SSIM reference image was a rendered view of the laser-scanned reconstruction, matching the resolution of the test image. This dual approach simultaneously validates the model's fidelity in both explicit geometry and visual structure.

Figure 7 shows the performance comparison of the structural similarity (SSIM) and Chamfer distance of the five models under different input image densities. As the number of input frames increases, the geometric accuracy of all models shows an upward trend, reflecting that richer perspective information helps the model capture spatial details and texture features. The NeRF fusion model continues to

lead in the SSIM index, showing that it recovers structural information more accurately and consistently. In the case of image limitation (frame number 4), the SSIM of the NeRF fusion model remains at 0.65. This advantage stems from the fact that the model effectively integrates volume density and color information under continuous viewing angles, enhancing the geometric expression ability of complex scenes. The change in Chamfer distance further proves the precision of geometric reconstruction. The NeRF fusion model shows a lower error, which is only 0.45 under image constraints (frame number 4), reflecting better overlap and spatial consistency between point clouds.

E. TEXTURE RECONSTRUCTION CONTINUITY ANALYSIS

The NeRF fusion model based on deep splicing results is analyzed for texture reconstruction continuity with MVSNet, COLMAP, PMVS and the original NeRF model, mainly in the irregular facades and illumination change areas. The quantitative indicators use LPIPS and Edge Consistency Index to measure the high-frequency texture retention ability and the integrity of image splicing and stitching, respectively.

Table 2 systematically evaluates the performance of five types of models in terms of texture reconstruction continuity based on two key scene areas. "Irregular facades" and "lighting change areas" are selected as evaluation samples to reflect the challenges of structural complexity and lighting interference on texture restoration. The LPIPS index is used to measure the similarity between the model output and the original image at the perceptual level. The lower the value, the closer the texture details are to human perception. The edge consistency index measures the transition smoothness and structural alignment of the image splicing boundary. The higher the value, the more natural the edge transition. From the numerical distribution, the NeRF fusion model shows lower LPIPS values and higher consistency indexes in both irregular facades and illumination change areas, with LPIPS means of 0.151 and 0.137, and edge consistency index means of 0.861 and 0.876, respectively, reflecting its significant advantages in high-frequency texture restoration and splicing continuity control. The achieved high texture continuity ($LPIPS \leq 0.151$) and geometric consistency not only demonstrate numerical advantages but also directly preserve the visual authenticity of key cultural elements in Jinhua's intangible cultural heritage space, including carved window lattices, painted beams, and brocade patterns. These details embody craftsmanship techniques and historical context, and their faithful digital reproduction provides a credible visual archive for academic research, public education, and intergenerational inheritance.

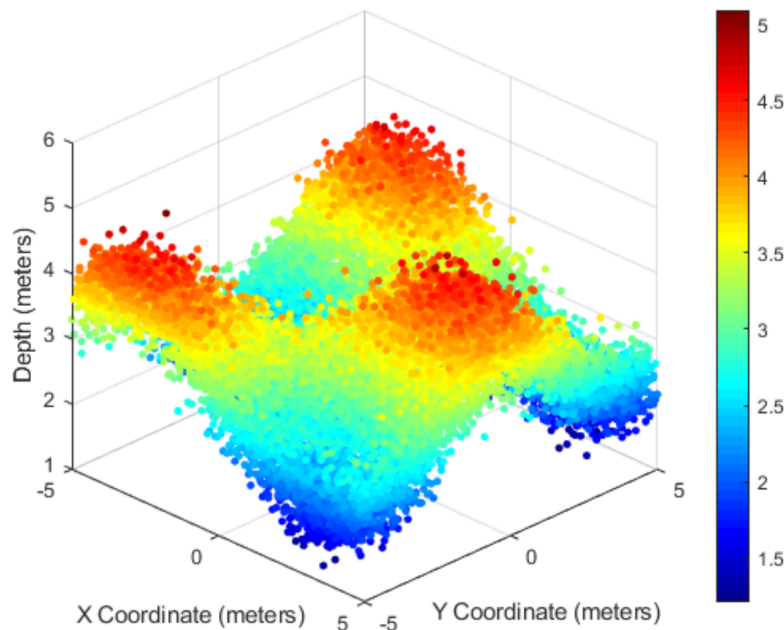


FIGURE 6. 3D point cloud projection of a single-frame depth map

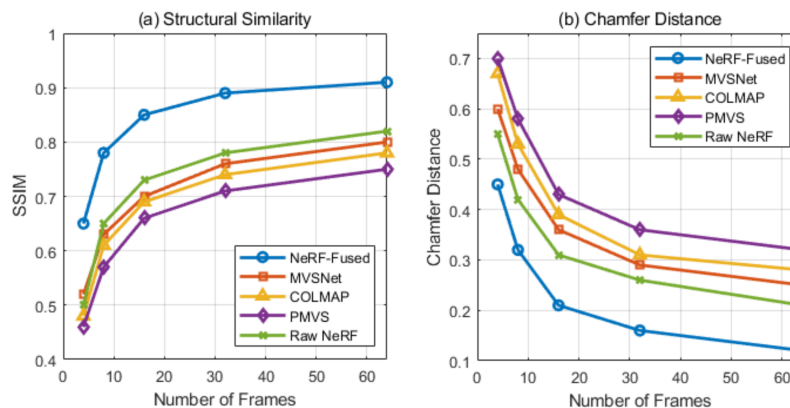


FIGURE 7. Geometric accuracy under different input densities

TABLE 2. Texture reconstruction continuity evaluation

Model	Region Type	LPIPS (Mean $\pm$ Std)	Edge Consistency Index (Mean $\pm$ Std (standard deviation))
NeRF-Fused	Irregular Facade	0.151 $\pm$ 0.012	0.861 $\pm$ 0.010
NeRF-Fused	Lighting	0.137 $\pm$ 0.016	0.876 $\pm$ 0.008
	Variation Zone		
MVSNet	Irregular Facade	0.224 $\pm$ 0.020	0.774 $\pm$ 0.014
MVSNet	Lighting	0.202 $\pm$ 0.019	0.784 $\pm$ 0.012
	Variation Zone		
COLMAP	Irregular Facade	0.245 $\pm$ 0.021	0.743 $\pm$ 0.017
COLMAP	Lighting	0.233 $\pm$ 0.022	0.757 $\pm$ 0.013
	Variation Zone		
PMVS	Irregular Facade	0.273 $\pm$ 0.018	0.721 $\pm$ 0.016
PMVS	Lighting	0.261 $\pm$ 0.022	0.728 $\pm$ 0.015
	Variation Zone		
Raw NeRF	Irregular Facade	0.195 $\pm$ 0.017	0.802 $\pm$ 0.013
Raw NeRF	Lighting	0.183 $\pm$ 0.016	0.815 $\pm$ 0.011
	Variation Zone		

F. USER PERSPECTIVE IMMERSION EXPERIENCE MEASUREMENT

This paper simulates the free movement trajectory of the user's perspective in the reconstructed model to evaluate the smoothness of the interaction between NeRF-Fusion and MVSNet, COLMAP, PMVS and the original NeRF model under continuous perspective switching. The quantitative indicators include the frame rate standard deviation and the perspective distortion perception score. The former is recorded in real time by the system, and the latter comes from the standardized processing of the subjective score of the experimenter (between 0 and 1). The Distortion Perception Score is a normalized metric (range [0, 1]) based on user subjective evaluations, quantifying the perceived artifacts like geometric discontinuities, texture jumps, or projection distortions during continuous viewpoint switch-

ing. A higher DPS indicates reduced visual distortion and stronger spatial consistency during immersion.

Table 3 shows the immersive experience performance of various 3D reconstruction models in typical perspective motion scenes. The evaluation is divided into three trajectory types: horizontal rotation, forward propulsion, and upward observation, which simulate the main observation paths of users in the intangible cultural heritage space. The frame rate standard deviation reflects the real-time response stability during continuous rendering. The smaller the value, the more coherent the picture. The distortion perception score is standardized by the user's subjective evaluation of structural deformation, screen tearing and edge distortion in each trajectory. The higher the score, the better the spatial consistency. The NeRF fusion model shows the consistency characteristics of low fluctuation and high immersion under all three types of trajectories. The standard deviation of the frame rate is controlled within 1.85, and the distortion perception score is maintained above 0.9. In contrast, PMVS and COLMAP show significant fluctuations in the forward and upward movement, indicating that there are problems of geometric fracture and texture jump when the perspective switches rapidly.

TABLE 3. Immersion experience measurement results

Model Type	Motion Type	Frame Rate Std	Distorted Perception Score
NeRF-Fusion	Horizontal Rotation	1.62	0.93
NeRF-Fusion	Forward Propulsion	1.84	0.91
NeRF-Fusion	Upward Observation	1.77	0.92
MVSNet	Horizontal Rotation	3.32	0.79
MVSNet	Forward Propulsion	3.58	0.76
MVSNet	Upward Observation	3.49	0.78
COLMAP	Horizontal Rotation	4.06	0.72
COLMAP	Forward Propulsion	4.38	0.7
COLMAP	Upward Observation	4.25	0.71
PMVS	Horizontal Rotation	4.71	0.69
PMVS	Forward Propulsion	5.02	0.67
PMVS	Upward Observation	4.94	0.68
Raw NeRF	Horizontal Rotation	2.73	0.84
Raw NeRF	Forward Propulsion	2.89	0.82
Raw NeRF	Upward Observation	2.81	0.83

G. EVALUATION OF CONSTRUCTION EFFICIENCY AND MODEL SCALABILITY

The construction efficiency and model scalability of each method are evaluated under different image resolutions (640×480, 1024×768, 1280×720, 1920×1080). Quantita-

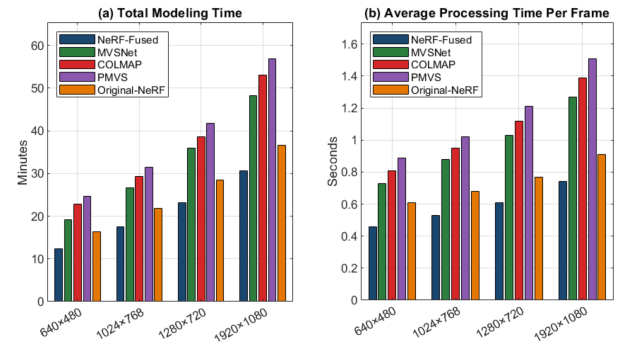


FIGURE 8. Comparison of overall modeling time and average processing time per frame

tive indicators include the complete modeling time (minutes) and the average processing time per frame (seconds), and the effects of NeRF-Fusion in this paper are compared with MVSNet, COLMAP, PMVS and the original NeRF model under the change of scene scale. The experimental platform is NVIDIA RTX 4090 GPU, with the total modeling time including NeRF training and stitching optimization.

Figure 8 shows the overall modeling time and average processing time per frame of the model under different image resolution conditions. When the image resolution increases, the modeling time of all models increases, but the increase rate varies. Among them, the NeRF fusion model still maintains a low overall time (30.6 minutes) at high resolution (1920×1080), reflecting the efficiency advantage of the optimization strategy in parallel structure scheduling and voxel sparse processing. In contrast, the other models showed obvious computational bottlenecks after the image size increased, especially in the texture reprojection and dense point cloud generation stages, where the processing load increased significantly. The average processing time per frame further illustrates the sensitivity of the model to different input scales. The NeRF fusion strategy still maintains a low average frame processing time (0.74 seconds) at high resolution (1920×1080), maintaining a relatively stable processing efficiency, indicating that its network structure is more scalable and adaptable.

V. CONCLUSION

This study focuses on the digital reproduction needs of Jinhua intangible cultural heritage space and adopts an immersive 3D reconstruction method that integrates neural radiation field and deep panoramic stitching. The advanced capabilities of this integrated method for high-fidelity visual capture are directly relevant to industrial applications with similar requirements. Consequently, this approach offers a model for the industry to meet the digital reproduction needs of complex assets, such as creating highly realistic 3D models of historic looms or detailed fabric patterns. Under complex conditions such as limited image acquisition and discontinuous texture, this method enhances texture integrity by constructing multi-frame spliced images and introduces NeRF for voxel modeling under continuous viewing angles,

solving the problems of incoherent structure restoration and insufficient detail restoration. The strength of this approach in overcoming data scarcity and enhancing integrity is highly valuable for industrial digitization. This provides a direct solution model for the industry to solve problems of incoherent structure restoration and enhance the digital representation of fabric textures and complex machinery, even with limited sensor data. Image consistency detection and PatchMatch local optimization mechanism are also superimposed in the fusion process, so that texture mapping shows high stability and visual consistency in boundary transitions and high-frequency areas. In multiple evaluation dimensions, when the model has 4 image frames, the SSIM remains at 0.65, the Chamfer distance is only 0.45, and the LPIPS averages in irregular facades and illumination change areas are 0.151 and 0.137 respectively. The edge consistency index averages are 0.861 and 0.876 respectively, the frame rate standard deviation under three types of trajectories is controlled within 1.85, and the distortion perception score is maintained above 0.9. This method shows superiority over traditional structure reconstruction or single neural modeling in terms of geometric accuracy, texture coherence, and interactive experience. The research findings indicate that the developed framework offers a highly valuable technical path for the digital reconstruction of complex cultural scenes, particularly when applied to the preservation of intangible cultural heritage.

AUTHOR CONTRIBUTIONS

Xueyuan Fang designed, collected and analyzed the data, and drafted the manuscript. Xueyuan Fang conducted the study, critically revised the manuscript for important intellectual content, and gave final approval of the version to be published. Xueyuan Fang participated fully in the work, take public responsibility for appropriate portions of the content, and agreed to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

CONFLICTS OF INTEREST

The author declares no conflict of interest.

FUND SUPPORT

In 2024, Zhejiang Province Philosophy and Social Science Planning "Provincial and municipal cooperation" project "Research on the path of reconstructing the sense of Place of Jinhua non-relic Image in the Digital Media Era", project number 24SSHZ157YB

ACKNOWLEDGEMENTS

Not applicable.

REFERENCES

- [1] D. Shao, K. J. Zoh, and Y. Xie, "The Spatial Differentiation Mechanism of Intangible Cultural Heritage and Its Integration with Tourism Development Based on Explainable Machine Learning and Coupled Coordination Models: a Case Study of the Jiang-Zhe-Hu in China," *Heritage Science*, vol. 12, no. 1, pp. 1-22, 2024, doi: 10.1186/s40494-024-01528-3.
- [2] B. Tu and Y. Chen, "Research on the Realization Path of Intangible Cultural Heritage Inheritance from the Perspective of Public Cultural Services: a Case Study of Yao Long-Drum Dance in S Town," *Humanities and Social Science Research*, vol. 8, no. 2, pp. 30-30, 2025, doi: 10.30560/hssr.v8n2p30.
- [3] J. Zhao, "Digital Protection and Inheritance Path of Intangible Cultural Heritage Based on Image Processing Algorithm," *Scalable Computing: Practice and Experience*, vol. 25, no. 6, pp. 4720-4728, 2024, doi: 10.12694/scpe.v25i6.3287.
- [4] İ. D. A. Ergin, "Digital Approach in Conservation of Heritage: 3D Virtual Reconstruction Applications in Ancient Cities," *Journal of Architectural Sciences and Applications*, vol. 8, no. 2, pp. 969-987, 2023, doi: 10.30785/mbud.1312738.
- [5] A. G. Pirbazari and S. K. Tabrizi, "RecordIM of Iran's Cultural Heritage Using an Online Virtual Museum, Considering the Coronavirus Pandemic," *ACM Journal on Computing and Cultural Heritage (JOCCH)*, vol. 15, no. 2, pp. 1-14, 2022, doi: 10.1145/3500925.
- [6] B. Liang, K. Sriuksa, N. Wongpongkham, and A. Sriuksa, "Digital Reconstruction of the Snake King Festival Intangible Cultural Heritage in Zhanghu Town: a Multidimensional AI-enabled Animation Approach," *Journal of Ecomodernism*, vol. 3, no. 8, pp. 9995-10012, 2024, doi: 10.62754/joe.v3i8.5612.
- [7] F. Colonnese, "Architectural Intangible Heritage and Graphic Reconstruction," *Terminological and Philological Notes. SCIRES-IT-Scientific REsearch and Information Technology*, vol. 13, no. 1, pp. 15-30, 2023, doi: 10.2423/j22394303v13n1p15.
- [8] D. Yan, "From Physical to Virtual: Enhancing the Representation of Intangible Cultural Heritage Using Mixed Reality," *Journal of Posthumanism*, vol. 5, no. 2, pp. 1541-1556, 2025, doi: 10.63332/joph.v5i2.537.
- [9] H. U. Jeong, J. Kim, J. Kong, and R. Y. C. Kim, "Applied Practices on Digital Historical Data Transformation Based on Intangible Cultural Heritage with Metaverse Approach," *International Journal of Advanced Smart Convergence*, vol. 13, no. 3, pp. 279-286, 2024, doi: 10.7236/IJASC.2024.13.3.279.
- [10] S. Zeng, X. Li, J. Tan, and X. Li, "Research on Digital Activation and Derivative Morphogenetic Design Methods for Peking Opera Intangible Cultural Heritage Based on Motion Capture," *The Design Journal*, vol. 28, no. 3, pp. 502-524, 2025, doi: 10.1080/14606925.2025.2467621.
- [11] Y. X. Tan, S. S. Fan, X. D. Zeng, and R. X. Wang, "3-D Reconstruction for Thermal Infrared Images of Substation Equipment Based on Visual SLAM and Cascade MVSNet," *IEEE Sensors Journal*, vol. 24, no. 7, pp. 11693-11704, 2024, doi: 10.1109/jsen.2024.3360439.
- [12] T. Wang and V. J. L. Gan, "Multi-View Stereo for Weakly Textured Indoor 3D Reconstruction," *Computer-Aided Civil and Infrastructure Engineering*, vol. 39, no. 10, pp. 1469-1489, 2024, doi: 10.1111/mice.13149.
- [13] H. Huang, X. Yan, Y. Zheng, J. He, L. Xu, and D. Qin, "Multi-View Stereo Algorithms Based on Deep Learning: a Survey," *Multimedia Tools and Applications*, vol. 84, no. 6, pp. 2877-2908, 2025, doi: 10.1007/s11042-024-20464-9.
- [14] B. K. Lee, B. Kim, W. J. Yoo, M. Ahn, and S. J. Han, "A Study on the Application of ColMap in 3D Reconstruction for Cultural Heritage Restoration," *Journal of the Korea Society of Computer and Information*, vol. 28, no. 8, pp. 95-101, 2023, doi: 10.9708/jksoci.2023.28.08.095.
- [15] F. Rahman and A. Nayak, "Drone-Based 3D Modeling for the Conservation and Restoration of Heritage Buildings," *AIS-Architecture Image Studies*, vol. 6, no. 1, pp. 46-55, 2025, doi: 10.48619/ais.v6i1.1090.
- [16] A. Keyvanfar, A. Shafaghat, and M. S. F. Rosley, "Performance Comparison Analysis of 3D Reconstruction Modeling Software in Construction Site Visualization and Mapping," *International Journal of Architectural Computing*, vol. 20, no. 2, pp. 453-475, 2022, doi: 10.1177/14780771211066876.
- [17] J. Zhang, J. Zhang, S. Mao, M. Ji, G. Wang, and Z. Chen, "GigaMVS: a Benchmark for Ultra-Large-Scale GIGAPIXEL-Level 3D Reconstruction," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 11, pp. 7534-7550, 2021, doi: 10.1109/TPAMI.2021.3115028.
- [18] Q. Zou and F. Liu, "3D Reconstruction of Optical Building Images Based on Improved 3D-R2N2 Algorithm," *Tehnički vjesnik*, vol. 30, no. 5, pp. 1594-1602, 2023, doi: 10.17559/TV-20230323000469.
- [19] G. Mazzacca, A. Karami, S. Rigon, E. M. Farella, P. Trybala, and F. Remondino, "NeRF for Heritage 3D Reconstruction," *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*,

- vol. 48, no. M-2-2023, pp. 1051-1058, 2023, doi: 10.5194/isprs-Archives-XLVIII-M-2-2023-1051-2023.
- [20] D. Harisanty, K. L. B. Obille, N. E. V. Anna, E. Purwanti, and F. Retrialisca, "Cultural Heritage Preservation in the Digital Age, Harnessing Artificial Intelligence for the Future: a Bibliometric Analysis," *Digital Library Perspectives*, vol. 40, no. 4, pp. 609-630, 2024, doi: 10.1108/DLP-01-2024-0018.