npj | heritage science

Article

https://doi.org/10.1038/s40494-026-02385-y

An SfM system for mural digitization with attention-guided feature matching and robust sparse reconstruction

Check for updates

Kaiyi Fang[1], Zhongyu Min[2] & Changyu Diao[1,3,4] ✉

Murals hold profound historical and artistic value, but their inevitable deterioration makes mural cultural heritage digitization increasingly urgent. Traditional Structure-from-Motion (SfM) methods often fail on mural data due to repetitive textures, low-texture regions, and the absence of camera metadata, resulting in poor feature matching and unstable reconstruction. To address these issues, this paper proposes a mural-oriented SfM system that integrates an attention-guided feature matching algorithm with a customized sparse reconstruction pipeline, including focal length estimation and edge-based bundle adjustment. Experiments conducted on mural datasets from the Mogao Grottoes demonstrate that the proposed system significantly improves reconstruction accuracy and robustness, providing a reliable technical foundation for large-scale mural digitization and offering a practical solution for preserving and studying mural heritage.

Mural paintings in grottoes, such as those in the Mogao Grottoes, are vital cultural heritage that reflect historical, artistic, and religious values while documenting intercultural exchange. Prolonged exposure to environmental and human factors has led to pigment fading, image blurring, and structural damage. Mural digitization enables high-precision data acquisition and preservation, providing scientific support for conservation, research, and cultural transmission[1]. A primary goal in this field is generating clear and undistorted frontal views of murals, which is an essential step impacting the effectiveness of all downstream tasks. Due to the large scale of murals, high-resolution cameras are typically used to capture different regions of the surface (Fig. 1), and these images are then merged via computational techniques to generate high-resolution orthographic views.

There are two primary strategies for merging local mural images: image stitching and 3D reconstruction. Traditional 2D methods, based on grayscale registration and blending, are inefficient for high-resolution murals and lack robustness[2]. Although early stitching methods often produced visible seams and geometric distortions (Fig. 2), recent advances have significantly alleviated these problems by preserving geometric consistency and object-level structure. Gao et al.[3] pioneered seam-driven approaches to minimize visible artifacts, Du et al.[4] proposed a geometric structure-preserving warp to reduce distortion, and Cai and Yang[5] further introduced object-level constraints for natural image stitching. Despite these impressive strides in visual consistency, stitching methods largely rely on homographies or mesh warping that assume quasi-planar scenes or fixed camera centers. Nevertheless, for mural data characterized by complex wall curvature, uneven lighting, subtle perspective variations, and scarce or repetitive textures (Fig. 3), stitching methods may still lead to misalignments or distortions[6]. Therefore, this study adopts a 3D reconstruction approach to obtain geometrically accurate and distortion-free orthographic mural views. Liu Liming[7] introduced 3D reconstruction for planar murals, but its accuracy can be reduced in complex scenes due to SfM feature matching errors[8]. Later studies improved SfM with MVS[9] or developed high-fidelity SfM-based stitching techniques[10]. Deep learning methods such as Pixel-Perfect SfM[11] further enhance accuracy and efficiency, but they often fail to maintain robustness under the complex textures and lighting of mural images.

To further improve mural 3D reconstruction, recent research has focused on two key components of the SfM pipeline: feature point matching and sparse reconstruction. Deep learning–based matching methods, such as SuperGlue[12], have introduced attention mechanisms to improve accuracy and robustness, while efficiency-oriented variants[13–15] and end-to-end approaches such as LoFTR and MatchFormer[16–22] further enhance performance but remain constrained by computational cost and model complexity. Sparse reconstruction has evolved from Bundler[23] to more efficient systems such as VisualSFM[24,25] and COLMAP[26,27], which optimize triangulation and bundle adjustment. Although recent learning-based approaches, including Pixel-Perfect SfM[11], DeepSfM, and DeepV2D[28–30], as well as techniques leveraging sparse reconstruction in NeRF and 3D Gaussian Splatting[31–34] have shown promise, their robustness and generality remain

[1]Zhejiang University, School of Art and Archaeology, Hangzhou, Zhejiang, China. [2]Zhejiang University, College of Computer Science and Technology, Hangzhou, Zhejiang, China. [3]Zhejiang University, Center for Balance Architecture, Hangzhou, Zhejiang, China. [4]Zhejiang University, Laboratory of Art and Archaeology Image, Hangzhou, Zhejiang, China. ✉e-mail: dcy@zju.edu.cn

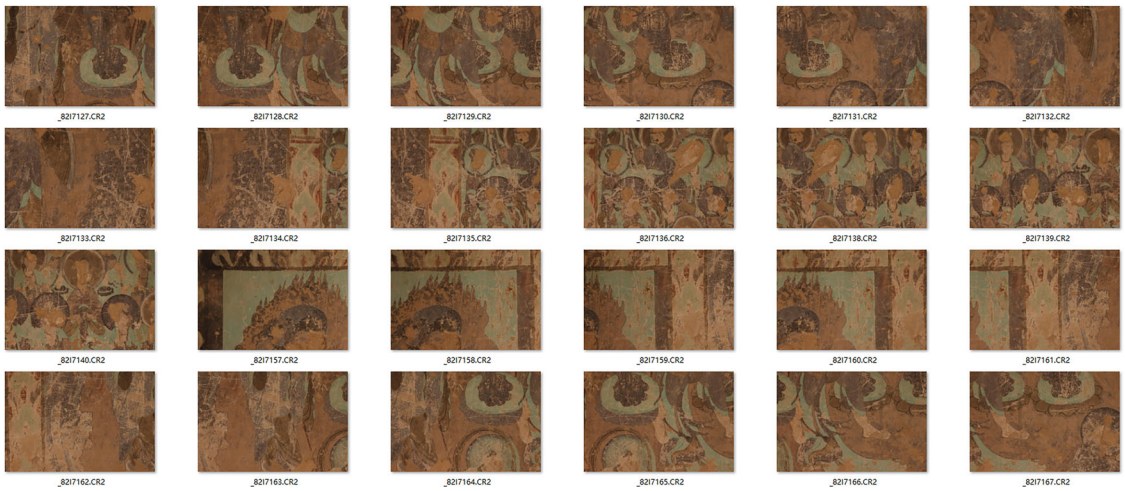


Fig. 1 | Local frontal mural views captured by high-precision camera.

Fig. 2 | Illustration of image stitching applied to grotto murals.

Fig. 3 | Scarce (left) or repetitive textures (right) in mural images.

limited. Consequently, traditional mathematically grounded sparse reconstruction pipelines continue to be the dominant framework in practical applications.

To address the challenges of large-scale mural digitization under complex geometric and texture conditions, we propose a mural-oriented Structure-from-Motion (SfM) system tailored to real-world grotto environments. The proposed system integrates an attention-guided feature matching module to enhance correspondence reliability in texture-ambiguous regions, together with a robust sparse reconstruction pipeline adapted to mural-specific acquisition conditions.

Extensive experiments on real-world mural datasets from the Mogao Grottoes, as well as the public MuralDH dataset, demonstrate that our approach enables stable and geometrically consistent reconstructions. The results show that the proposed system outperforms existing pipelines based on stitching and SfM in challenging mural scenarios, providing a reliable technical foundation for large-scale cultural heritage preservation.

Methods: an SfM system tailored for mural data

The robustness of SfM-based reconstruction depends on image acquisition conditions. In mural digitization, uneven grotto wall geometry often violates the planar or fixed-view assumptions of traditional image stitching. Although SfM relaxes these constraints by jointly estimating camera poses and scene structure, it still requires sufficient image overlap and reasonably uniform spatial coverage for stable reconstruction. Accordingly, we design the acquisition protocol to ensure adequate overlap between adjacent views, while the proposed attention-guided matching module alleviates the resulting computational burden and enables geometrically consistent reconstruction.

The SfM system proposed in this study consists of two core modules: feature point extraction and matching, and sparse reconstruction. The feature matching module integrates an attention mechanism-based algorithm to significantly improve matching accuracy and efficiency. Meanwhile, the sparse reconstruction module incorporates custom-designed focal length estimation and bundle adjustment strategies to enhance the

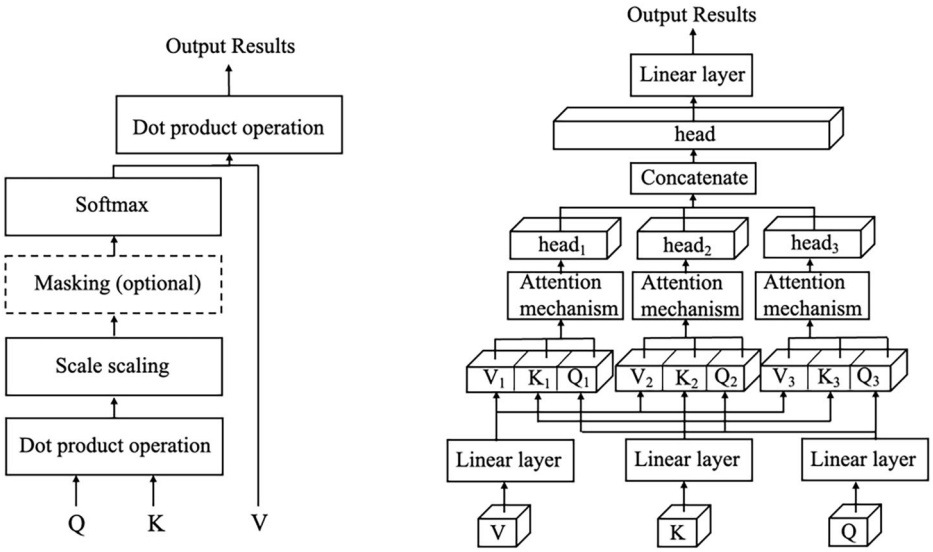


Fig. 4 | General architecture of the attention mechanism in our backbone. (Left) Flowchart of the Attention Mechanism. (Right) Flowchart of the Multi-Head Attention Mechanism.

Fig. 5 | Example of feature point matching in overlapping mural segments.

robustness and precision of the reconstruction process. In this paper, the term attention-guided matching algorithm refers to the feature matching module within the proposed SfM system.

Attention-guided feature matching for mural images

The feature matching module adopts an attention-based matching paradigm to establish reliable correspondences under challenging mural conditions. Inspired by attention-based message-passing formulations, we redesign the matching process to better accommodate mural-specific constraints, specifically repetitive texture ambiguities and high computational costs.

In our framework, each feature point is represented by a high-dimensional embedding encoding both visual descriptors and spatial information. These embeddings are projected into query, key, and value representations to facilitate dynamic contextual aggregation, the general architecture of which is illustrated in Fig. 4. This process enables the model to suppress implausible correspondences by leveraging contextual consistency, which is vital in mural regions lacking strong local discriminative cues. To enhance this interaction, a Multi-loop Attention network is introduced to expand the effective receptive field across iterative stages, allowing the system to resolve local ambiguities that a single-pass mechanism may fail to distinguish.

Building upon this attention-guided formulation, the system incorporates two core architectural departures:

(1) Efficiency Optimization: To handle the thousands of high-resolution images required for mural digitization, the iterative Sinkhorn algorithm is replaced with a lightweight vector cross-product-based similarity computation. This strategy reduces computational overhead by approximately 77% while maintaining comparable accuracy.

(2) Geometric Constraint Integration: We introduce a learnable spatial attention gate into the final matching probability computation. Unlike generic matchers that rely primarily on visual similarity, our method explicitly incorporates learnable 2D position embeddings to modulate matching scores. This mechanism functions as a soft geometric prior and effectively suppresses matches that are visually similar but spatially inconsistent with the learned geometric layout.

Replacement of the Sinkhorn Algorithm. The Sinkhorn algorithm is traditionally utilized to solve optimal transport problems for finding correspondences between feature point sets. However, the computational burden of this iterative process is significant for large scale mural digitization where thousands of high resolution images must be processed. Recent studies have introduced more efficient alternatives to address such bottlenecks, including LoFTR and LightGlue. LightGlue in particular demonstrates high performance when image pairs exhibit limited geometric variation and consistent scales. Since mural acquisition typically meets these conditions due to our rail based capture protocol, we adopt a inner product strategy as a lightweight replacement for Sinkhorn to enhance matching efficiency.

Through the multi-head attention network, we obtain $\mathbf{x}_A$ and $\mathbf{x}_B$, derived from ${}^{(l-1)}\mathbf{x}^{A}_{cross}$ and ${}^{(l-1)}\mathbf{x}^{B}_{cross}$, respectively.

$$S_{i,j} = \mathrm{Linear}\left(\mathbf{x}_i^{A}\right)^{\top} \mathrm{Linear}\left(\mathbf{x}_j^{B}\right), \forall (i,j) \in A \times B \quad (1)$$

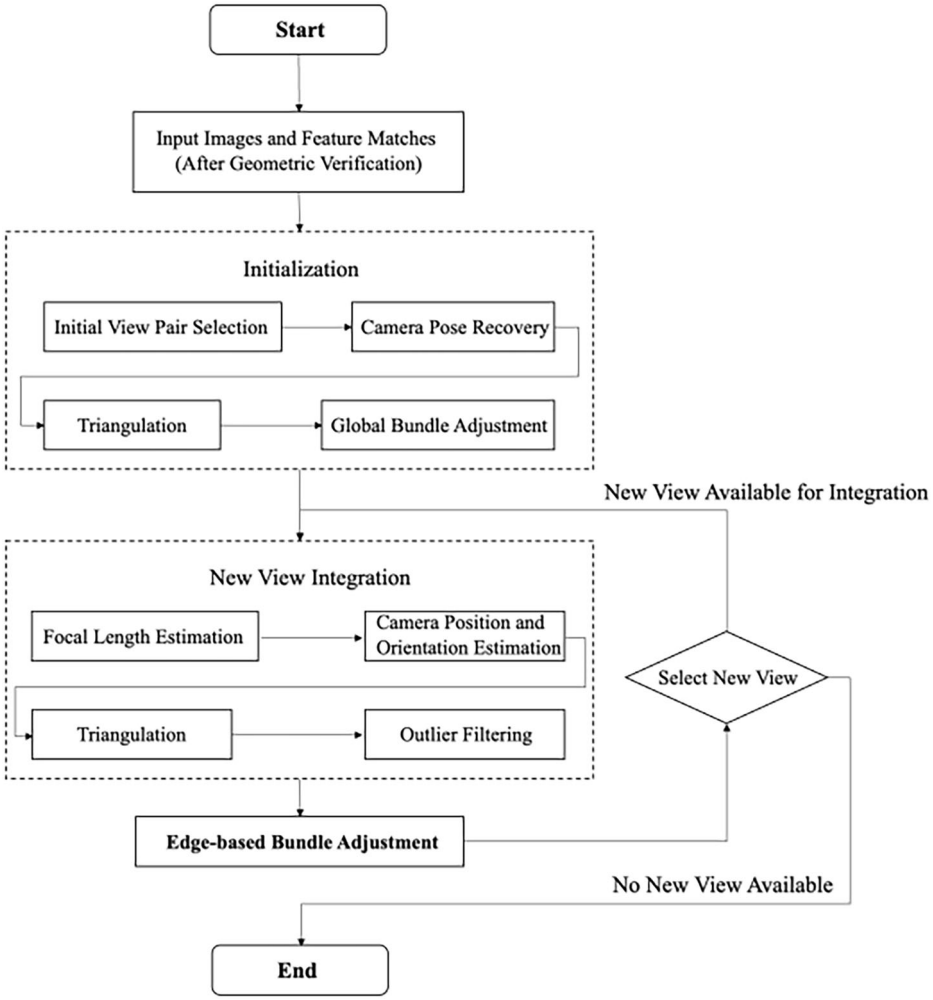


Fig. 6 | Sparse reconstruction workflow tailored for murals.

$$P_{i,j} = \mathrm{Softmax}_{k\in A}(S_{k,j})_i \cdot \mathrm{Softmax}_{k\in B}(S_{i,k})_j \tag{2}$$

In Eq. (1), Linear denotes a learnable linear transformation with bias. $S_{i,j}$ represents the similarity score between the i^{th} feature point in image A and the j^{th} feature point in image B. Equation (2) transforms the similarity matrix $\mathbf{S}$ into a probability matrix $\mathbf{P}$, where $P_{i,j}$ denotes the probability that the feature points i and j are matched. If $P_{i,j}$ is the maximum in both its row and column, the two points are considered a match.

Incorporating Feature Point Location Information. It is observed that correctly matched feature points tend to be concentrated within overlapping regions of the corresponding image areas (Fig. 5). Based on this insight, we incorporate learnable spatial awareness into the final matching stage of the algorithm to explicitly model geometric consistency.

Figure 5 illustrates two partially overlapping mural segments, where the red-framed region indicates the actual area of geometric overlap. Feature matches located outside this region are geometrically invalid because they fall into non-overlapping zones where epipolar constraints cannot be satisfied. To leverage this observation, we introduce a learnable position embedding mechanism in Eq. (3) instead of a rigid distance heuristic. This mechanism projects 2D coordinates into a high-dimensional feature space to capture spatial dependencies. The term $\mathbf{WW}^\top$ computes the interaction between these position embeddings, effectively functioning as a spatial attention gate. This gate modulates the visual matching score by enforcing spatial coherence: it encourages the network to prioritize matches that are not only visually similar but also spatially consistent with the learned geometric layout of the mural. By learning these spatial relationships, the model acts as a soft geometric prior, suppressing geometrically implausible matches (e.g., random texture matches in non-overlapping zones) and improving overall reconstruction robustness.

$$\mathbf{w}_i = \mathrm{Linear}(\mathbf{p}_i)$$

$$\sigma = \mathrm{Sigmoid}\left(\mathbf{WW}^\top \mathrm{Linear}(\mathbf{x})\right) \in [0,1] \tag{3}$$

In Eq. (3), $\mathbf{p}_i$ denotes the pixel coordinate of the i^{th} feature point. It is projected via a learnable linear transformation into a high-dimensional position embedding $\mathbf{w}_i$. We then stack all embeddings $\mathbf{w}_i$ to form the position matrix $\mathbf{W}$. Consequently, the term $\mathbf{WW}^\top$ represents the spatial interaction matrix, capturing the geometric distances between all feature locations. Similarly, Linear($\mathbf{x}$) applies a learnable transformation to the visual features $\mathbf{x}$, resulting in a score vector (rather than a single value) that indicates the initial match likelihood. Finally, we multiply the interaction matrix $\mathbf{WW}^\top$ with this score vector and apply a Sigmoid function to obtain σ, where the i^{th} element σ_i represents the probability that feature point i has a valid correspondence.

By incorporating $\mathbf{x}_A$ and $\mathbf{x}_B$ into Eq. (2), and subsequently adding σ^A and σ^B, the final matching probability is computed using Eq. (4).

$$P_{i,j} = \sigma_i^A \sigma_j^B \mathrm{Softmax}_{k\in A}(S_{k,j})_i \mathrm{Softmax}_{k\in B}(S_{i,k})_j \tag{4}$$

Sparse reconstruction pipeline tailored for mural data

This paper proposes a sparse reconstruction pipeline specifically designed for mural scenarios, introducing two key innovations:

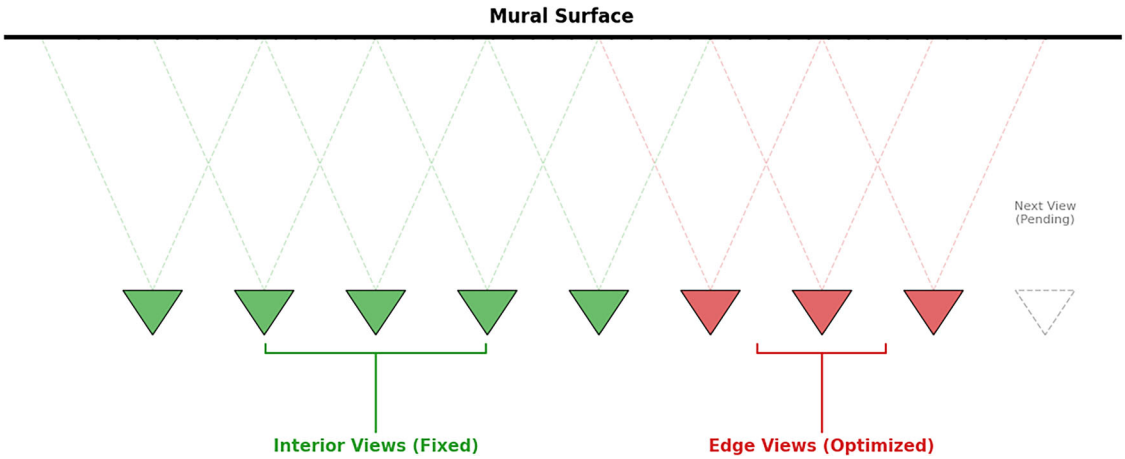


Fig. 7 | Schematic illustration of the edge-based bundle adjustment strategy.

(1) a mural-oriented camera focal length estimation module, and
(2) an edge-based bundle adjustment module.

These improvements effectively address the limitations of other open-source sparse reconstruction systems, which often fail to estimate camera focal lengths accurately or converge to physically meaningless solutions during bundle adjustment optimization. Comparative experiments demonstrate the robustness and accuracy of the proposed method, with both modules performing well on mural datasets.

Overall Sparse Reconstruction Pipeline. As illustrated in Fig. 6, the sparse reconstruction pipeline designed for murals comprises two main stages: initialization and new-view integration. The initialization step estimates the intrinsic and extrinsic parameters of the initial camera pair and generates a partial sparse point cloud of the scene. Subsequently, additional views are incrementally added to the reconstruction, updating both the sparse point cloud and camera parameters.

The reconstruction process begins with an initial camera pair and incrementally registers new views to the existing model. At each iteration, feature correspondences are established and camera poses are estimated using PnP with RANSAC. The reconstruction process terminates when a newly added view produces fewer than a predefined number of inlier matches with the existing model, indicating that no further reliable registration is possible. In our implementation, this threshold is empirically set to 50, following typical practice in SfM systems such as COLMAP and VisualSFM, where values between 30 and 100 are commonly used. The resulting model then serves as the final output of the sparse reconstruction. The proposed pipeline integrates two novel modules, mural-specific focal length estimation and edge-based bundle adjustment, while adopting other standard components from open-source software such as COLMAP and OpenMVG[35].

Focal Length Estimation for Mural Cameras. Camera focal length is crucial for model accuracy as it directly affects 3D reconstruction and motion estimation. Inaccurate focal lengths can lead to increased reprojection errors and incorrect camera pose estimates. Due to manual verification and format conversions, the EXIF metadata in mural images is often altered or missing, making direct focal length retrieval infeasible. Therefore, it is necessary to re-estimate the focal length through self-calibration.

This paper adopts an enumeration strategy over a range of candidate focal lengths, selecting the value that yields optimal camera pose estimation. Given the nature of mural imagery, the variation in focal length across different views is relatively small. Thus, the focal length of new cameras is estimated using the average focal length of existing reconstructed cameras, reducing computation.

If the number of cameras in the current reconstruction is below a threshold, a wide-range enumeration is performed. Once the threshold is reached or exceeded, the estimation is confined to a small range around the average focal length. The detailed pseudo-code for mural camera focal length estimation and pose solving is provided in Supplementary Algorithm 1.

In our implementation, we assume that the mural region of interest has been approximately cropped from the input images prior to focal length estimation. This assumption is substantiated by the specific acquisition protocols of the Mogao Grottoes datasets used in this study. The images were captured using high-precision digital cameras mounted on mechanical rails, ensuring consistent camera-to-wall distances and stable framing that focuses primarily on the mural surface. Consequently, the variation in intrinsic parameters is minimized, validating the reliability of our estimation strategy. When substantial background elements are present in the input, additional preprocessing (e.g., cropping or masking) may be necessary to ensure accurate intrinsic parameter estimation.

Edge-based Bundle Adjustment. This paper proposes an edge-based bundle adjustment strategy, which differs from global bundle adjustment by optimizing only a subset of cameras and their associated 3D points. By strictly constraining the objective function and reducing the dimensionality of the solution space, this method achieves more accurate results with lower computational cost.

The algorithm first determines whether each camera has sufficient feature correspondences in other images. If a camera has already been well-optimized through previous reconstructions, it is excluded from further optimization. Typically, the images requiring optimization lie at the periphery of the reconstructed mural scene. As new camera views are incrementally added from the outer regions of the mural, previously reconstructed views that were initially located at the boundary gradually become enclosed by newer views. A view is considered "interior" if all other views that share matching feature points with it have already been successfully registered into the reconstruction. In this case, the view is assumed to have sufficient geometric constraints and is excluded from further optimization, as its parameters have typically converged through earlier iterations. The optimization also filters 3D points based on their projection relationships to the selected cameras, retaining only those reconstructed from feature correspondences in the optimized views.

Figure 7 demonstrates the camera states during the incremental reconstruction process. (Left, Green) "Interior Views" represent cameras that have been fully registered and enclosed by newer views; their parameters are fixed to reduce computational load. (Right, Red) "Edge Views" represent the newly added cameras at the boundary of the reconstruction; these views, along with their associated 3D points, are subject to active optimization to ensure accurate registration.

Considering the uniqueness of mural imagery, this method introduces constraints on camera focal lengths and radial distortion parameters. Since mural data is usually captured using high-precision digital cameras mounted on rails, the focal lengths are expected to be close to the average,

Table 1 | Experimental results on the mural dataset

Feature Extraction Method	Feature Matching Method	Time (s)	Recall (%)	Accuracy (%)
SuperPoint	Nearest Neighbor	**0.02**	69.7	61.5
	SuperGlue	0.54	86.9	80.4
	LightGlue	0.11	86.3	79.9
	Proposed Method	0.12	**90.1**	**84.2**

Bold marks the best result per group.

Table 2 | Ablation study results

Feature Extraction Method	Replace Sinkhorn	Feature Point Position Enhancement	Time (s)	Recall (%)	Accuracy (%)
SuperPoint	no	no	0.54	86.9	80.4
	yes	no	0.12	87.0	81.2
	no	yes	0.53	89.2	83.9
	yes	yes	0.12	90.1	84.2

and radial distortion is minimal. An upper bound is imposed on the distortion parameter during optimization.

$$\min \sum_{i\in N}\sum_{j\in M}\left(\mathbf{u}_{ij}-\pi(\mathbf{C}_j,\mathbf{X}_i)\right)^2+\alpha\left(f(\mathbf{C}_j)-\bar{F}\right)^2 \quad (5)$$

$$\text{s. t. } k(\mathbf{C}_j)\le K, \forall j\in M$$

The loss function of the edge-based bundle adjustment is defined in Eq. (5). Let M denote the set of reconstructed cameras and N the set of 3D points. $\mathbf{u}_{ij}$ is the projection of 3D point $\mathbf{X}_i$ onto the view of camera $\mathbf{C}_j$. α is the weighting factor, $\bar{F}$ is the average focal length in M, and K is the upper limit on radial distortion. $f(\mathbf{C}_j)$ and $k(\mathbf{C}_j)$ denote the focal length and distortion of camera $\mathbf{C}_j$, respectively. $\pi(\mathbf{C}_j,\mathbf{X}_i)$ is the projection function mapping 3D point $\mathbf{X}_i$ to its 2D coordinates in the image plane of camera $\mathbf{C}_j$.

In Eq. (5), the first term is the reprojection error, consistent with standard bundle adjustment. The second term penalizes deviations in focal length, with α controlling the balance between the two. A constraint is imposed to ensure that the distortion coefficients of all reconstructed cameras do not exceed K. The projection function $\pi(\mathbf{C},\mathbf{X})$ calculates the 2D coordinates $(\tilde{\mathrm{x}},\tilde{\mathrm{y}})$ of the 3D point on the image plane and measures the distance to the observed pixel coordinates $(\mathbf{u}_{ij},\mathbf{v}_{ij})$.

$$\begin{bmatrix}\tilde{\mathrm{x}}\\ \tilde{\mathrm{y}}\end{bmatrix}=\frac{\mathrm{f}}{\mathrm{w}}\begin{bmatrix}\mathrm{u}\\ \mathrm{v}\end{bmatrix}+\begin{bmatrix}\mathrm{c_x}\\ \mathrm{c_y}\end{bmatrix}+\begin{bmatrix}du\\ dv\end{bmatrix} \quad (6)$$

The 2D projection coordinates $(\tilde{\mathrm{x}},\tilde{\mathrm{y}})$ are computed using Eq. (6), where u, v, w represent the 3D point's coordinates in the camera coordinate system. f, $(\mathrm{c_x},\ \mathrm{c_y})$, k denote the focal length, principal point, and distortion coefficient, while du, dv are the distortion-corrected values for u and v (Eq. 7).

$$du=\frac{k}{w^3}(u^2+v^2)u, dv=\frac{k}{w^3}(u^2+v^2)v \quad (7)$$

The optimization process is carried out using the Ceres Solver. The full pseudo-code of the edge-based bundle adjustment is provided in Supplementary Algorithm 2.

Results

The experiments in this study focus on the mural images from the Mogao Grottoes in Dunhuang, Gansu Province. The test dataset comprises 9 representative sets of mural images provided by the Dunhuang Academy, covering Caves 12, 19, 72, 148, 172, 322, 390, and 428 of the Mogao Grottoes, as well as selected murals from grottoes in Subei County. A total of 1800 images were used, covering a broad range of typical mural categories. The implementation details and training procedure are provided in Supplementary B.

Evaluation of the attention-based mural feature matching algorithm

In this experiment, the accuracy and recall rate are used to evaluate the performance of the feature point matching algorithm. Since the murals in the 9 test sets are planar, the homography matrix $\mathbf{H}$ between the local frontal-view images can be precomputed to assess matching correctness. The 5-pixel threshold is adopted as a reasonable and commonly used criterion under known planar homography, given that our rail-based acquisition protocol maintains a consistent camera-to-wall distance and thus a uniform physical error scale across all datasets. All compared methods are evaluated under this identical threshold, ensuring a fair and consistent performance comparison.

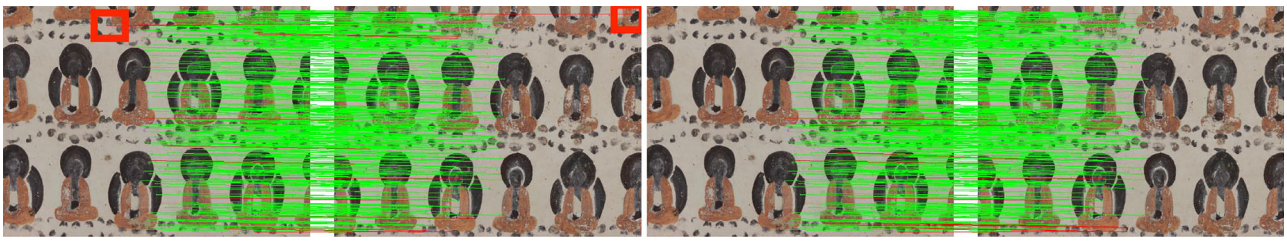
Definition of accuracy: A matched point pair (P_i^A, P_i^B) is considered correct if feature point P_i^A from image A is mapped to image B using homography $\mathbf{H}$, and its distance to the corresponding point P_i^B is less than 5 pixels. Accuracy is the ratio of the number of correct matches to the total number of matches.

Definition of recall: A point P^A from image A is mapped via homography $\mathbf{H}$ to image B as P^{-A}. If there exists a feature point P_i^B in image B within 5 pixels of P_i^{-A}, P_i^B is considered covered. Recall is the ratio of the number of correct matches to the total number of such covered points P_i^B.

Main Evaluation Results. To verify the effectiveness of the proposed algorithm, it is compared with several classical methods. For algorithms that require training, the same training parameters and datasets as this section are used to ensure fair comparison. The compared methods are as follows:

Nearest Neighbor Matching[36]: This method compares the descriptors d^A and d^B extracted by the SuperPoint feature extraction algorithm from images A and B, respectively. It is a classical baseline method for feature point matching.

SuperGlue[12]: Since the proposed method is an improvement of SuperGlue, it is included for comparison.



(a) (b)

Fig. 8 | Example of enhanced feature point matching. a Without Feature Point Position Enhancement **b** With Feature Point Position Enhancement.

Table 3 | Mural coverage rates

Mural Group	Overlap Coverage (%)
Cave 72	58
Cave 148	56
Cave 322	50
Cave 428	52

LightGlue[11]: A modified version of SuperGlue, this algorithm reduces inference time while maintaining comparable performance on simple datasets.

From Table 1, it is evident that nearest neighbor matching yields the worst results. SuperGlue and LightGlue exhibit similar accuracy and recall rates on the mural dataset. However, the proposed method outperforms all others in both recall and accuracy. Furthermore, the inference speed of the proposed method approaches that of LightGlue and is significantly faster than SuperGlue. Overall, the proposed algorithm performs best on the mural dataset.

Ablation Study. As shown in Table 2, both improvements proposed in this study contribute effectively to performance on the mural dataset. Replacing the Sinkhorn algorithm improves inference speed, though it has limited impact on accuracy and recall. In contrast, introducing feature point location awareness significantly improves both metrics. With the two improvements combined, the algorithm achieves a 77.78% reduction in processing time, along with a 3.2% increase in recall and a 3.8% increase in accuracy. The effectiveness of location-aware matching is demonstrated in the following example.

As shown in Fig. 8, incorporating feature point location significantly improves matching results. Figure 8a and b show left and right image pairs. Green lines represent correct matches, and red lines indicate mismatches. In Fig. 8a, red rectangles highlight mismatched points located outside the overlapping region. These mismatches are corrected in Fig. 8b after location-aware enhancement. The number of correct matches in dense regions also increases.

In summary, the ablation study validates that:

(1) replacing the Sinkhorn algorithm with vector cross-product;
(2) adding feature point location information in the matching stage

—both significantly enhance the model's performance on mural feature matching tasks.

Sparse reconstruction experiments for mural data

A subset of 381 images were used for this experiment. Each mural group consists of about 100 images, with coverage rates shown in Table 3. To ensure accuracy, the feature matching results of the test set were manually verified to eliminate incorrect matches, so any errors in reconstruction are solely due to the sparse reconstruction process. All methods, including the proposed one and four open-source baselines, use the same feature point matching inputs to ensure fairness.

Comparison with Existing Sparse Reconstruction Methods. To evaluate the effectiveness of the proposed sparse reconstruction pipeline, we compare it with four classical open-source SfM systems:

OpenMVG[35]: A C++ library for multiple-view geometry. Version 2.1 was used.

MVE[37]: Integrates SfM, MVS, and surface reconstruction.

VisualSFM[38]: Quickly constructs sparse point clouds and integrates MVS.

COLMAP[36]: A widely used C++-based SfM and MVS system in fields such as CV, cultural heritage, VR, and game development. The evaluation metric is the average reprojection error (in pixels).

According to Table 4, OpenMVG and MVE fail to reconstruct the test set. VisualSFM achieves the smallest reprojection error on Cave 322 data, but the scene is misaligned and thus not valid. COLMAP reconstructs Caves 72, 322, and 428 successfully but fails for Cave 148, and its reprojection error is higher than ours. Our method successfully reconstructs all test sets with the lowest reprojection errors.

Figure 9 shows the sparse reconstruction of Cave 148 produced by our system's GUI. The frontal view (Fig. 9a) demonstrates uniform camera distribution and full coverage, while the side view (Fig. 9b) confirms the accurate recovery of the planar structure. These results verify that our pipeline improves both the robustness and precision of mural reconstruction.

We evaluated the focal length estimation module against a baseline using direct EXIF data without optimization. The baseline fails for Caves 72 and 428 and produces significant pose errors for Caves 148 and 322 (Supplementary Table 1). As shown in Fig. 10, our method yields realistic camera poses, whereas the baseline results in chaotic placement due to focal length inaccuracies.

To validate the edge-based BA module, we compare it with traditional global BA on the same four datasets. Table 5 shows that the baseline exhibits higher reprojection errors, longer computation times, and complete failure for Cave 148. As illustrated in Fig. 11, the baseline method suffers from significant camera drift (Std. = 0.83 m). It also shows a noticeable "bowing effect," which is a common degeneracy in near-planar SfM caused by inadequate geometric constraints. In contrast, our proposed system reduces the standard deviation of camera center deviation to 0.21 m. This reduction represents a substantial improvement in planar consistency that aligns with the mechanical precision of our rail-based acquisition system. Consequently, the recovered camera poses remain physically plausible even under limited spatial coverage.

SfM system for murals and its practical applications

The complete pipeline of our mural-oriented SfM system is illustrated in Fig. 12. We conducted comparative evaluations against several open-source tools using a dedicated mural dataset. To comprehensively evaluate its performance, mural image sets from each cave are divided into two test groups of different scales. The first group contains approximately 100 images per mural set, while the second group contains around 200 images per set. The coverage area of murals varies across these test groups, with overlap rates ranging from 50% to 65%.

Table 4 | Sparse reconstruction comparison results

Test Group	Total Images	Reconstruction Method	Successfully Reconstructed	Max Reconstructed Views	Reprojection Error (px)
Cave 72	100	OpenMVG	No	—	—
		MVE	No	—	—
		VisualSFM	No	100	0.702
		COLMAP	Yes	100	0.110
		Proposed Method	Yes	100	**0.094**
Cave 148	90	OpenMVG	No	—	—
		MVE	No	—	—
		VisualSFM	Yes	90	0.302
		COLMAP	Yes	42	0.298
		Proposed Method	Yes	90	**0.275**
Cave 322	92	OpenMVG	No	—	—
		MVE	No	—	—
		VisualSFM	No	91	**0.089**
		COLMAP	No	92	0.136
		Proposed Method	Yes	92	0.105
Cave 428	99	OpenMVG	No	—	—
		MVE	No	—	—
		VisualSFM	Yes	99	0.200
		COLMAP	Yes	99	0.121
		Proposed Method	Yes	99	**0.097**

Bold marks the best result per group.

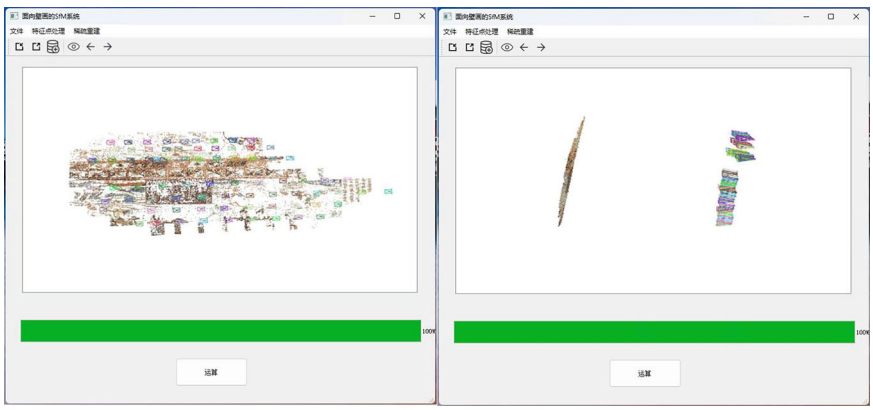


Fig. 9 | Our system's GUI. a Frontal View **b** Side View.

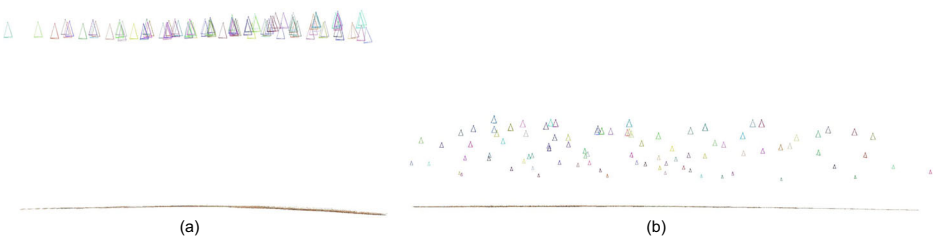


Fig. 10 | Side view comparison for cave 148. **a** Proposed Method Results **b** Baseline Method Results.

This study compares the proposed SfM system with several open-source alternatives: COLMAP, VisualSFM, MVE, and OpenMVG. Among them, MVE and OpenMVG fail to reconstruct most of the test sets, producing errors consistent with earlier comparison experiments. Therefore, their results are excluded from the following evaluation tables.

The sparse reconstruction results of both groups using the proposed system are illustrated in Supplementary Information Fig. 3 and 4. Since the system lacks a graphical interface, the results are visualized in the VisualSFM GUI. The detailed comparative results are provided in Tables 6 and 7. While VisualSFM successfully reconstructs all murals with complete sparse point

Table 5 | Edge-based bundle adjustment results

Test Group	Reconstruction Method	Correctly Completed Reconstruction	Max Reconstructed Views	Reprojection Error (px)	Time (min)
Cave 72	Baseline Method	No	100	0.189	4.2
	Proposed Method	Yes	100	**0.111**	**3.6**
Cave 148	Baseline Method	No	40	0.394	1.7
	Proposed Method	Yes	90	**0.301**	3.4
Cave 322	Baseline Method	Yes	92	0.342	6.5
	Proposed Method	Yes	92	**0.121**	**5.3**
Cave 428	Baseline Method	Yes	99	0.231	5.4
	Proposed Method	Yes	99	**0.104**	**5.0**

Bold marks the best result per group.

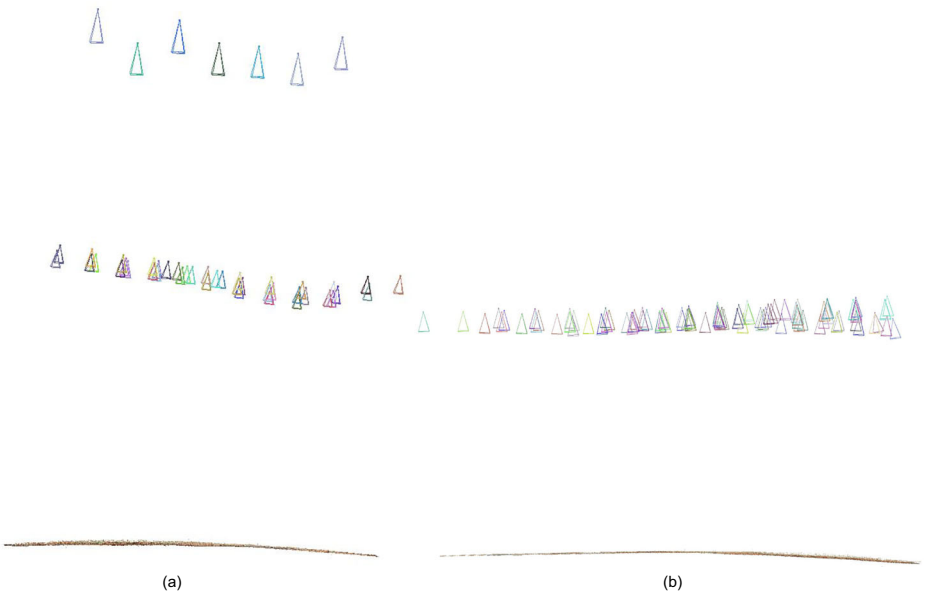


Fig. 11 | Comparison of bundle adjustment results. a Baseline Method Results (Std. of camera center deviation: 0.83 m) **b** Proposed Method Results (Std. of camera center deviation: 0.21 m).

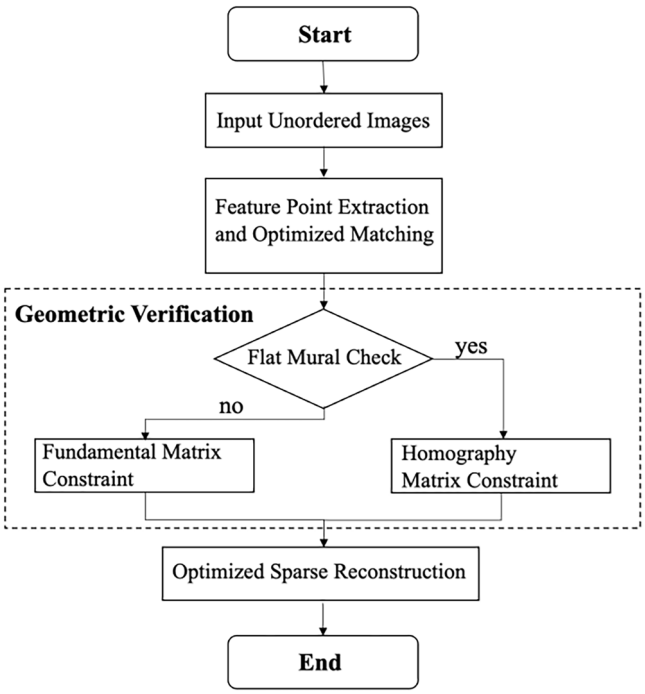


Fig. 12 | 12. SfM system workflow for murals.

clouds, it suffers from distortion and large pose estimation errors. COLMAP performs better on certain caves but struggles on large-scale sets (e.g., 148, 172, 390, and 428), where unrestricted bundle adjustment leads to cumulative errors and missing views.

In contrast, the proposed SfM system completes sparse reconstruction on all test cases, producing accurate sparse point clouds and camera extrinsics. Overall, it demonstrates the highest robustness and precision on mural data.

The ultimate goal of this research is to obtain clear and complete frontal views of murals. This section combines the proposed SfM system with external tools to reconstruct 3D geometry and extract frontal images, demonstrating its practicality. Notably, some mural data could not yield clear frontal views without the proposed system, highlighting its necessity. The reconstruction of Cave 148 serves as an illustrative example.

To obtain a complete and clear frontal view of mural surfaces, the proposed SfM system performs feature extraction, matching, and sparse reconstruction. The resulting sparse point cloud is then imported into RealityCapture (RC) for dense reconstruction and orthographic projection.

Figure 13 shows the reconstruction pipeline for Mogao Grotto No. 148. Subfigures (a–d) illustrate intermediate results, while (e) presents the final orthophoto. Mosaic artifacts near the edges of (e) are caused by insufficient image coverage in those regions; nevertheless, all local views are successfully integrated into a coherent

Table 6 | Comparison results of the first test dataset group

Test Cave	Total Images	System Used	Successfully Completed Reconstruction	Max Reconstructed Views	Reprojection Error (px)
Subei County Grottoes	98	VisualSFM	No	98	0.276
		COLMAP	Yes	100	0.133
		Proposed System	Yes	100	**0.131**
Mogao Grotto No.72	100	VisualSFM	No	100	0.753
		COLMAP	Yes	100	0.114
		Proposed System	Yes	100	**0.102**
Mogao Grotto No.148	99	VisualSFM	No	90	0.327
		COLMAP	No	42	0.322
		Proposed System	Yes	90	**0.157**
Mogao Grotto No.322	92	VisualSFM	Yes	91	0.078
		COLMAP	Yes	92	0.121
		Proposed System	Yes	92	**0.111**
Mogao Grotto No.428	99	VisualSFM	No	99	0.226
		COLMAP	Yes	99	0.104
		Proposed System	Yes	99	**0.110**
Mogao Grotto No.12	99	VisualSFM	No	99	0.291
		COLMAP	Yes	99	0.137
		Proposed System	Yes	99	**0.101**
Mogao Grotto No.172	100	VisualSFM	No	100	0.209
		COLMAP	No	24	0.296
		Proposed System	Yes	100	**0.181**
Mogao Grotto No.19	100	VisualSFM	No	100	0.212
		COLMAP	Yes	100	0.141
		Proposed System	Yes	100	0.142
Mogao Grotto No.390	100	VisualSFM	No	100	0.449
		COLMAP	Yes	100	0.132
		Proposed System	Yes	100	**0.112**

Bold marks the best result per group.

and clear frontal image, demonstrating the practicality of our approach.

For comparison, Fig. 14 shows the results obtained using RC's built-in SfM module, which exhibits significant alignment errors and texture artifacts. In contrast, our system delivers more accurate camera poses and higher-quality orthophotos.

Finally, Fig. 15 presents frontal views of nine mural datasets reconstructed using our pipeline, further validating its robustness and usability for large-scale mural digitization.

Generalization analysis on the public MuralDH dataset

To further validate the generalization capability and robustness of the proposed SfM system beyond the data provided by the Dunhuang Academy, this paper conducts comparative experiments on the public dataset MuralDH[39]. Since most existing public mural datasets consist of single high-resolution images rather than the multi-view sequences required for SfM reconstruction, we constructed a simulated reconstruction dataset based on MuralDH. Specifically, we selected 500 high-quality mural images from the dataset. Adopting a sliding-window cropping strategy, each original image was segmented into 121 sub-images with an overlap ratio controlled between 45% and 65%, simulating the camera path and image overlap typically found in real-world mural digitization. This process generated a large-scale test set containing 60,500 images across 500 groups of sequences. Such a large amount of data provides a sufficient basis for a comprehensive model performance evaluation.

The experimental results are shown in Table 8, demonstrating the comparison of the proposed SfM system with VisualSFM and COLMAP on the processed MuralDH dataset. We evaluated the methods based on Reconstruction Success Rate, Time Efficiency, and Reprojection Error.

Compared with VisualSFM and COLMAP, the proposed system demonstrates significant advantages across all metrics. In terms of reconstruction success rate, our system successfully reconstructed 353 groups (70.6%), whereas VisualSFM and COLMAP struggled with the repetitive textures inherent in the MuralDH style, leading to frequent failures or partial reconstructions. Regarding accuracy, the proposed system achieved the lowest average reprojection error (0.127 px), which is significantly better than the 0.155 px of COLMAP and 0.189 px of VisualSFM, indicating a superior capability in recovering precise camera poses and scene geometry.

In addition, the computational efficiency of our system is notable. The average processing time per group was only 1.552 minutes, which is significantly faster than the other two methods. This efficiency is attributed to the attention-guided feature matching module, which accelerates convergence, and the edge-based bundle adjustment, which reduces the optimization load.

Figure 16 visually compares the sparse point clouds across four representative scenes. While VisualSfM and COLMAP struggle with sparsity and gaps in regions with repetitive or weak textures, the Proposed System consistently yields the densest and most complete reconstructions. Notably, in the highly repetitive scenario (column 4), our method effectively

Table 7 | Comparison results of the second test dataset group

Test Cave	Total Images	System Used	Successfully Completed Reconstruction	Max Reconstructed Views	Reprojection Error (px)
Subei County Grottoes	200	VisualSFM	No	200	0.297
		COLMAP	Yes	200	0.145
		Proposed System	Yes	200	**0.139**
Mogao Grotto No.72	200	VisualSFM	No	200	0.481
		COLMAP	Yes	200	0.123
		Proposed System	Yes	200	**0.098**
Mogao Grotto No.148	199	VisualSFM	No	199	0.503
		COLMAP	No	95	0.131
		Proposed System	Yes	199	**0.124**
Mogao Grotto No.322	208	VisualSFM	No	208	0.443
		COLMAP	Yes	208	0.130
		Proposed System	Yes	208	**0.102**
Mogao Grotto No.428	204	VisualSFM	No	204	0.310
		COLMAP	No	203	0.132
		Proposed System	Yes	204	**0.130**
Mogao Grotto No.12	199	VisualSFM	No	199	0.694
		COLMAP	Yes	199	0.131
		Proposed System	Yes	199	**0.119**
Mogao Grotto No.172	199	VisualSFM	No	199	0.228
		COLMAP	No	45	0.301
		Proposed System	Yes	199	**0.153**
Mogao Grotto No.19	200	VisualSFM	No	200	1.145
		COLMAP	Yes	200	0.138
		Proposed System	Yes	200	**0.129**
Mogao Grotto No.390	200	VisualSFM	No	198	0.201
		COLMAP	No	199	0.133
		Proposed System	Yes	200	**0.109**

Bold marks the best result per group.

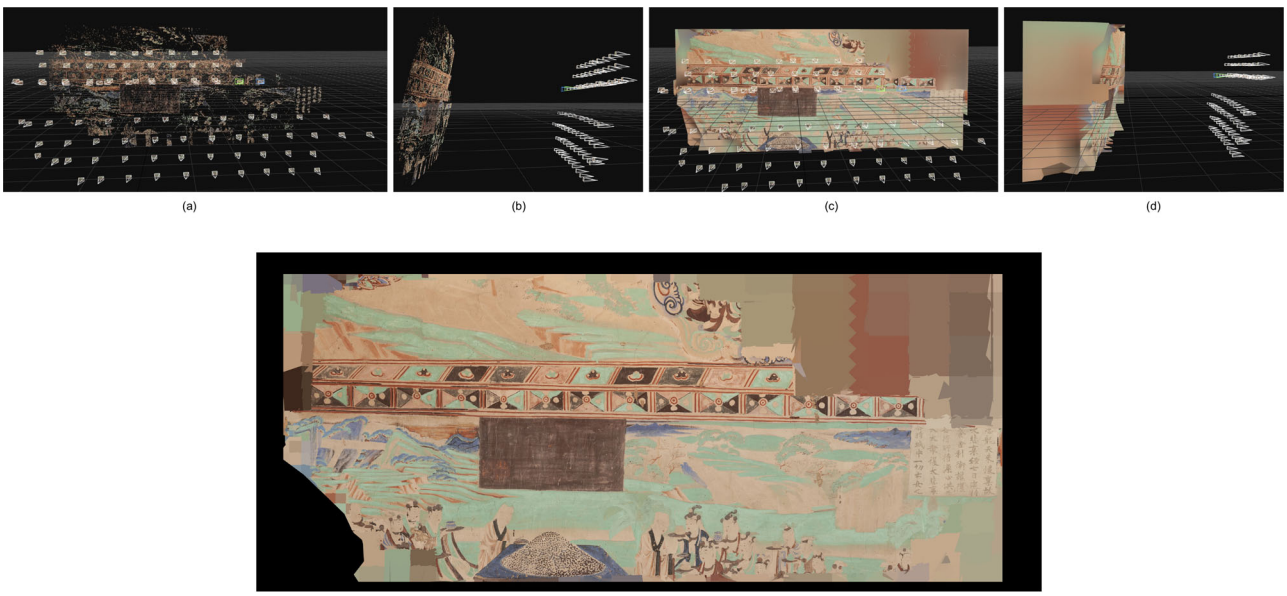


Fig. 13 | a, **b** Frontal and side views generated by RC using the output from the proposed SfM system. **c**, **d** Frontal and side views of the model reconstructed by RC. **e** The final orthographic image.

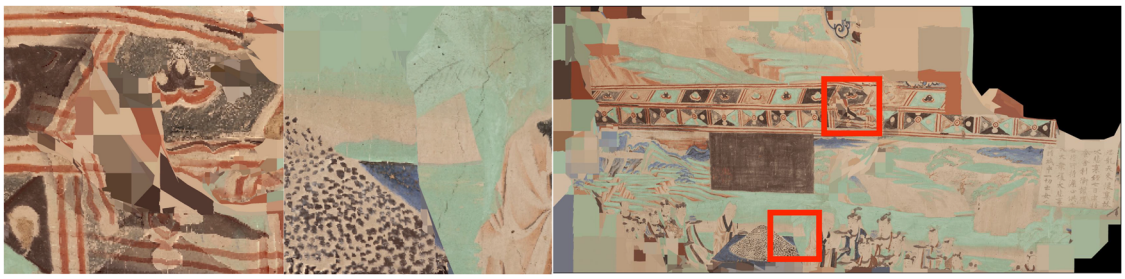
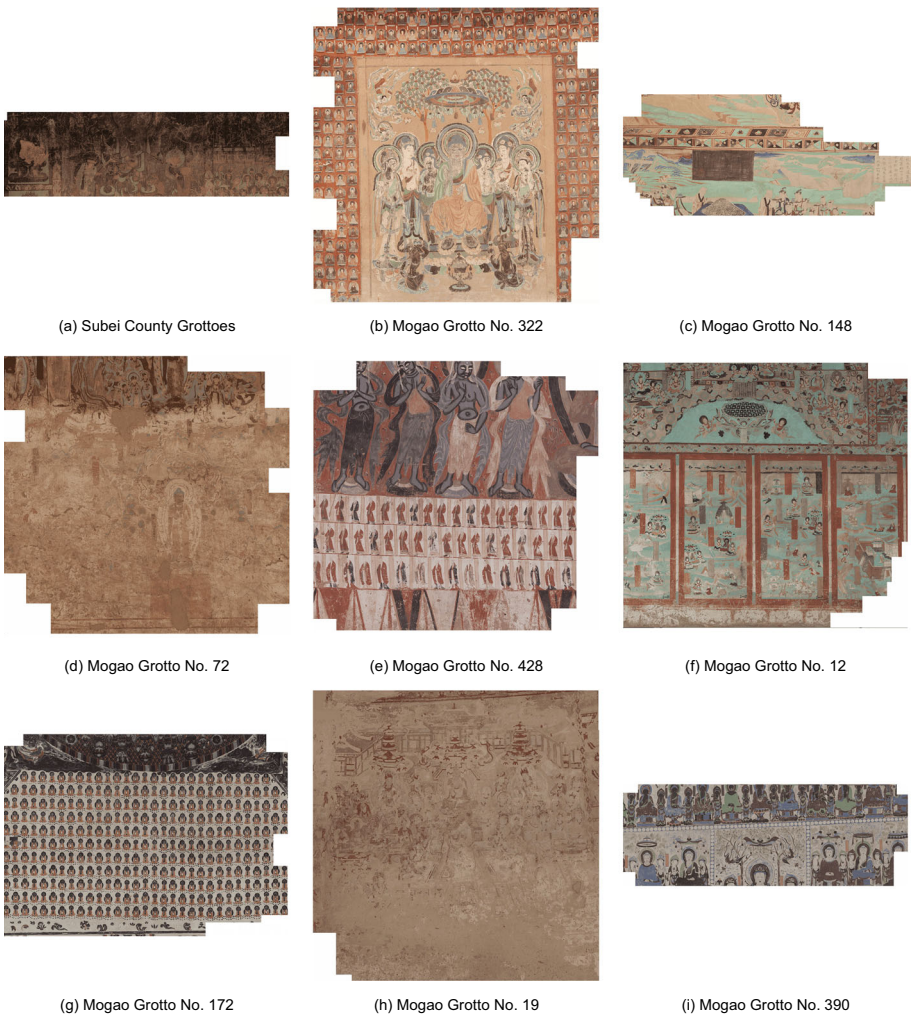


Fig. 14 | Image artifacts produced by the built-in SfM module of RC software.

Fig. 15 | Nine sets of mural frontal views obtained by integrating the proposed SfM system with RC software.



resolves ambiguity to maintain structural coherence, corroborating the quantitative results in Table 8.

In conclusion, experiments on the public MuralDH dataset confirm that the advantages of the proposed system are not limited to specific grotto environments. The system exhibits strong generalization, high accuracy, and real-time performance, demonstrating its broad applicability to other mural cultural heritage digitization tasks involving complex planar textures.

Discussion

Reliable feature correspondence remains a key bottleneck for large-scale mural reconstruction. Unlike urban scenes, mural surfaces are characterized by repetitive patterns, faded pigments, and weak local gradients, where visually similar regions are widely distributed. Under such conditions, even a small number of incorrect matches can propagate through the SfM pipeline and lead to unstable camera pose estimation or reconstruction failure, particularly in near-planar scenarios with constrained camera motion.

The effectiveness of the proposed system lies in embedding mural-specific geometric inductive bias into the correspondence estimation process. By softly constraining matches to spatially plausible overlapping regions, the attention-guided strategy suppresses geometrically implausible correspondences caused by repetitive textures, without imposing rigid geometric assumptions. This design is especially suitable for grotto

Table 8 | Comparative experimental results on the MuralDH dataset

System Used	Reconstruction Success Rate	Avg. Reconstructed Views	Time (min)	Reprojection Error (px)
VisualSfM	28.4%	105	1.685	0.189
COLMAP	50.2%	90	3.042	0.155
Proposed System	**70.6%**	**112**	**1.552**	**0.127**

Bold marks the best result per group.

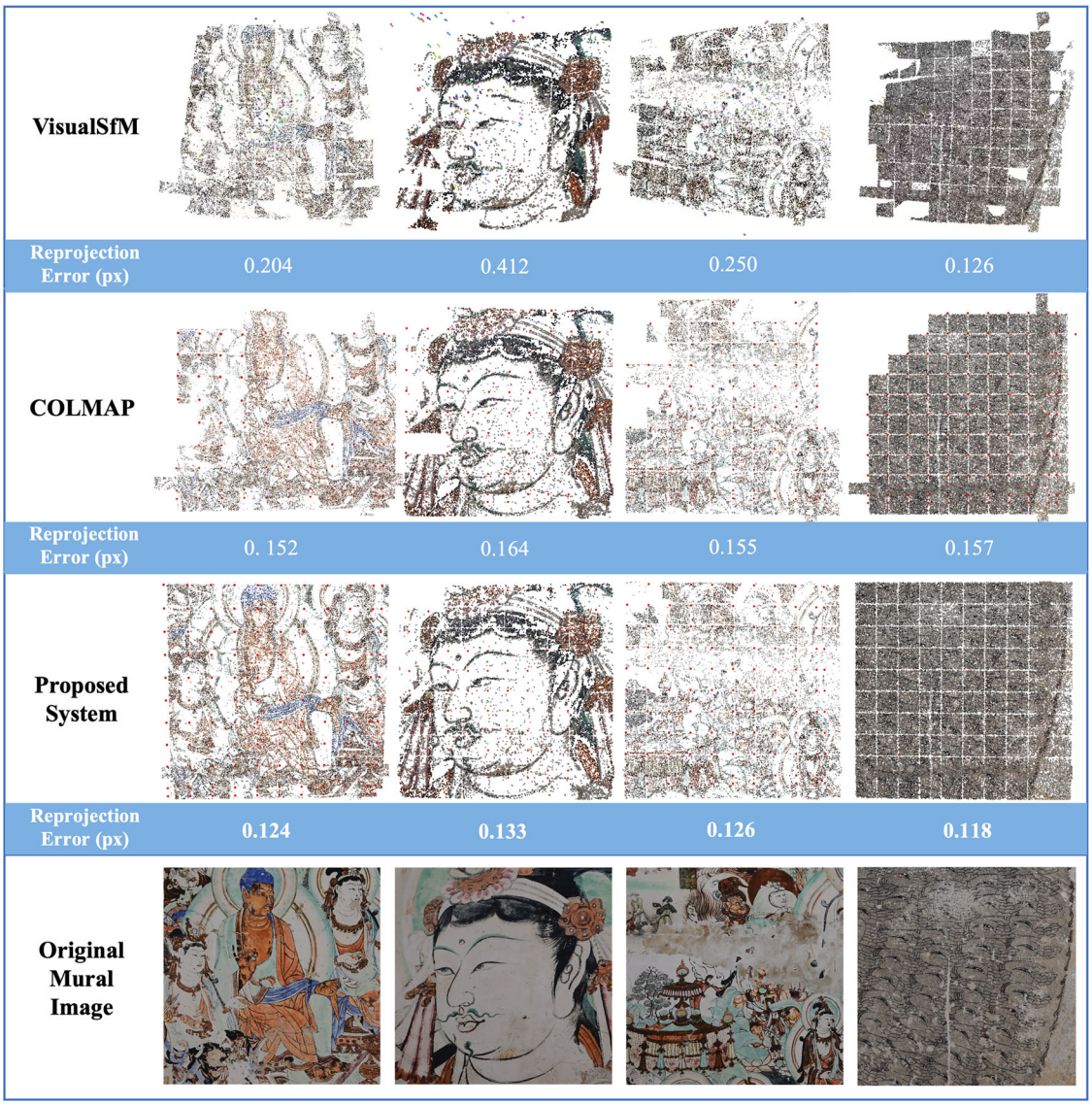


Fig. 16 | Qualitative comparison of sparse reconstruction results on the MuralDH dataset.

environments, where near-planar geometry and controlled acquisition provide weak but consistent spatial regularities.

Improvements in sparse reconstruction further demonstrate the importance of adapting SfM pipelines to mural acquisition conditions. In practice, unreliable camera intrinsics and unrestricted global bundle adjustment often lead to physically implausible solutions. By regularizing focal length estimation and selectively optimizing boundary views, the proposed approach reduces optimization ambiguity and mitigates common degeneracies such as camera drift.

Despite these improvements, the system still relies on sufficient image overlap, which remains an inherent limitation of SfM-based approaches. Future work may explore integrating dense correspondence refinement or optical flow–based strategies to further enhance robustness under extremely weak texture conditions.

Data availability

All data generated or analyzed in this study are included in this article and its supplementary materials. The MuralDH dataset used in this study is publicly accessible at: https://github.com/tearsheaven/MuralDH.

Code availability

The custom source code for the SfM system developed in this study is not publicly available due to confidentiality agreements with the Dunhuang Academy and institutional restrictions associated with an ongoing research project. To support reproducibility, we provide detailed algorithmic descriptions, pseudo-code, and complete experimental settings in the manuscript and Supplementary Materials. Partial implementations or representative modules may be released in the future, subject to institutional approval.

Received: 22 September 2025; Accepted: 11 February 2026;
Published online: 21 March 2026

References

1. Fan, J. Application of digital technologies in the protection and exhibition of the Dunhuang Grottoes. *Dunhuang Res* **6**, 1–3 (2009).
2. He, W. Generation of panoramic images of cylindrical murals based on multi-view images. Thesis, East China Normal University (2021).
3. Gao, J. et al. Seam-driven image stitching. *Eurographics (Short Papers)* 45–48 (2013).
4. Du, P. et al. Geometric structure preserving warp for natural image stitching. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 456–465 (IEEE, 2022) *2022*.
5. Cai, W. & Yang, W. Object-level geometric structure preserving for natural image stitching. In *2025 AAAI Conference on Artificial Intelligence (AAAI)*. **39**, 2 (AAAI Press, 2025).
6. Mei, Y. et al. DunHuangStitch: unsupervised deep image stitching of Dunhuang murals. *IEEE Trans. Vis. Comput. Graph.* **31**, 4226–4240 (2025).
7. Liu, L. Study on ultra-high resolution mural digitization based on 3D reconstruction. Thesis, Zhejiang University (2016).
8. Lai, Z. Research on 3D reconstruction methods and applications for immovable cultural relics based on improved NeRF. Thesis, Fujian Agriculture and Forestry University (2024).
9. Kholil, M., Ismanto, I. & Fu'Ad, M. N. 3D reconstruction using structure from motion (SfM) algorithm and multi view stereo (MVS) based on computer vision. *IOP Conf. Ser. Mater. Sci. Eng.* **1073**, 012066 (2021).
10. Shen, W. High-fidelity stitching technique for large-format mural images based on SfM. Thesis, Zhejiang University (2011).
11. Lindenberger, P. et al. Pixel-perfect structure-from-motion with featuremetric refinement. In *IEEE/CVF International Conference on Computer Vision (ICCV)*. 5987–5997 (IEEE, 2021).*2021*
12. Sarlin, P. E. et al. SuperGlue: learning feature matching with graph neural networks. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 4938–4947 (IEEE, 2020).
13. Chen, H. et al. Learning to match features with seeded graph matching network. In 2021 *IEEE/CVF International Conference on Computer Vision (ICCV)*. 6301–6310 (IEEE, 2021).
14. Shi, Y. et al. ClusterGNN: cluster-based coarse-to-fine graph neural network for efficient feature matching. In 2022 *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 12517–12526 (IEEE, 2022).
15. Lindenberger, P., Sarlin, P. E. & Pollefeys, M. Lightglue: local feature matching at light speed. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*. 17627–17638 (IEEE, 2023).
16. Sun, J. et al. LoFTR: detector-free local feature matching with transformers. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 8922–8931 (IEEE, 2021).
17. Wang, Q. et al. MatchFormer: interleaving attention in transformers for feature matching. In *2022 Asian Conference on Computer Vision (ACCV)*. 2746–2762 (Springer, 2022).
18. Dai, K. et al. OAMatcher: an overlapping areas-based network with label credibility for robust and accurate feature matching. *Pattern Recognit* **147**, 110094 (2024).
19. Chen, H. et al. ASpanFormer: detector-free image matching with adaptive span transformer. In *2022 European Conference on Computer Vision (ECCV)*. 20–36 (Springer, 2022).
20. Yu, J. et al. Adaptive spot-guided transformer for consistent local feature matching. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 21898–21908 (IEEE, 2023).
21. Cao, C. & Fu, Y. Improving transformer-based image matching by cascaded capturing spatially informative keypoints. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*. 12129–12139 (IEEE, 2023).
22. Zhu, S. & Liu, X. PMatch: paired masked image modeling for dense geometric matching. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 21909–21918 (IEEE, 2023).
23. Snavely, N., Seitz, S. M. & Szeliski, R. Photo tourism: exploring photo collections in 3D. *ACM SIGGRAPH 2006 Papers* 835–846 (ACM, 2006).
24. Wu, C. Towards linear-time incremental structure from motion. In *2013 International Conference on 3D Vision (3DV)*. 127–134 (IEEE, 2013).
25. Wu, C. SiftGPU: a GPU implementation of SIFT. http://cs.unc.edu/~ccwu/siftgpu (2007).
26. Schönberger, J. L. et al. Pixelwise view selection for unstructured multi-view stereo. In *2016 European Conference on Computer Vision (ECCV)*. 501–518 (Springer, 2016).
27. Schönberger, J. L. & Frahm, J. M. Structure-from-motion revisited. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 4104–4113 (IEEE, 2016).
28. Wang, J. et al. Deep two-view structure-from-motion revisited. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 8953–8962 (IEEE, 2021).
29. Wei, X. et al. Deepsfm: structure from motion via deep bundle adjustment. In *2020 European Conference on Computer Vision (ECCV)*. 230–247 (Springer, 2020).
30. Teed, Z. & Deng, J. DeepV2D: video to depth with differentiable structure from motion. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 13564–13573 (IEEE, 2020).
31. Mildenhall, B. et al. NeRF: representing scenes as neural radiance fields for view synthesis. *Commun. ACM* **65**, 99–106 (2021).
32. Kerbl, B. et al. 3D Gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.* **42**, 1–14 (2023).
33. Xu, L. et al. Grid-guided neural radiance fields for large urban scenes. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 8296–8306 (IEEE, 2023).
34. Lee, D. et al. DP-NeRF: deblurred neural radiance field with physical scene priors. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 12386–12396 (IEEE, 2023).
35. Moulon, P. et al. OpenMVG: open multiple view geometry. In *2016 International Workshop on Reproducible Research in Pattern Recognition*. 60–74 (Springer, 2016).
36. DeTone, D., Malisiewicz, T. & Rabinovich, A. SuperPoint: self-supervised interest point detection and description. In *2018 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. 224–236 (IEEE, 2018).
37. Fuhrmann, S., Langguth, F. & Goesele, M. MVE — a multi-view reconstruction environment. *GCH* **3**, 2 (2014).
38. Elbayad, M. et al. Depth-adaptive transformer. *arXiv preprint arXiv:1910.10073* (2019).
39. Xu, Z. et al. A comprehensive dataset for digital restoration of Dunhuang murals. *Sci. Data* **11**, 955 (2024).

Acknowledgements
This work is supported by the National Natural Science Foundation of China (Grant No. 62332015) and the Major Project of Philosophy and Social Sciences of Zhejiang Province (Grant No. 25SYS01ZD).

Author contributions
K.F. and Z. M. conceptualized the structure of the manuscript; K.F. and Z.M. completed the methodology and experiments; K.F. and Z.M. developed the conceptual framework and wrote a draft; K.F. and Z.M. and C. D. revised the manuscript; K.F. and Z.M. and C. D. reviewed the manuscript; and C. D. secured funding. All authors read and approved the final manuscript.

Competing interests
The authors declare no competing interests.

Additional information
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s40494-026-02385-y.

Correspondence and requests for materials should be addressed to Changyu Diao.