IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 5 June 2024, accepted 30 June 2024, date of publication 5 July 2024, date of current version 15 July 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3424293

RESEARCH ARTICLE

Refined 3D Modeling of Complex Models Based on Stereo Vision

HAOTIAN MING[1], QI LI[1,2], HAIBO XIA[1], AND PENG LI[1]
[1]School of Civil and Architectural Engineering, Guangxi University of Science and Technology, Liuzhou 545006, China
[2]Key Laboratory of Disaster Prevention & Mitigation and Prestress Technology of Guangxi Colleges and Universities, Liuzhou 545006, China

Corresponding author: Qi Li (liqi@gxust.edu.cn)

This work was supported in part by the Central Guidance Local Science and Technology Development Fund Projects under Grant 2022SRZ0101, in part by Guangxi Young and Middle-Aged Teachers' Scientific Research Basic Ability Improvement under Project 2023KY0346 and Project 2022KY0344, in part by Guangxi Key Research and Development Plan under Grant 2023AB38108, and in part by the Open Project Fund for Key Laboratory of Disaster Prevention & Mitigation and Prestress Technology of Guangxi Colleges and Universities under Grant GXKDTJ018.

ABSTRACT Despite the decreasing costs of stereo vision technologies, traditional 3D reconstruction methods still face challenges related to operational complexity and high costs, which somewhat limit their widespread adoption in various technical applications. Addressing these challenges, this paper presents an improved 3D reconstruction method combining COLMAP and OpenMVS, particularly suited for use with standard consumer-grade imaging devices such as smartphones and drones. By employing Structure from Motion (SFM) and Multi-View Stereo (MVS) techniques, this study significantly enhances image processing speed and achieves substantial improvements in model accuracy and detail reproduction. Systematic experimental validation has demonstrated that the combination of COLMAP and OpenMVS outperforms other open-source tools and combinations in terms of reconstruction speed and precision. This finding highlights the extensive potential and applicability of this combined approach in applications such as virtual reality, robotic navigation, and the digitization of cultural heritage.

INDEX TERMS Structure from motion, multi view stereo, 3D reconstruction, model refinement.

I. INTRODUCTION

With the popularity of portable camera devices such as smartphones, digital cameras, and motion cameras and the development of image-based multi-view stereo vision technology, it is easier and more accessible for people to take photos. The demand for 3D application technology increases sharply, such as virtual reality [1], augmented reality [2], robot navigation [3], cultural heritage protection and digitization [4], etc. 3D reconstruction technology has received extensive attention. It has shown bright application prospects in many fields and provides a solid foundation for better environmental perception and scene understanding. The rapid progress of computer hardware and algorithms provides more computational support for 3D reconstruction technology. Sensor technology is also constantly improving and reducing costs (such as lidar, depth cameras, etc.). Its development provides more data sources for 3D reconstruction technology, which significantly improves the ability to obtain 3D data [5]. According to different requirements, the acquisition of 3D data can be divided into two ways: active (such as lidar) and passive (camera). Hu et al. used the multi-stage Terrestrial Laser Scanner (TLS) integration method to conduct fine measurements and 3D modelling of ancient buildings [6]. Although active methods such as lidar scanning can obtain high-precision 3D data, the cost of the scanning equipment is relatively expensive for ordinary users and small-scale applications.

Moreover, the performance of lidar scanning could be improved in better lighting conditions, and it is easy to be interfered with by strong light, which leads to increased measurement errors. More importantly, lidar scanning does not contain any semantic information [7], and the passive acquisition method of data acquisition through cameras because of its relatively low cost, high flexibility, and the advantages of obtaining geometric and semantic information at the same time, can improve the accuracy of lidar scanning. It has been widely used in recent years. For example, using 3D

The associate editor coordinating the review of this manuscript and approving it for publication was Zhan-Li Sun.

reconstruction technology based on UAV photography, Yu et al. designed a low-cost, simple, efficient, and high-precision technical process for obtaining and 3D reconstruction of fine geometric texture on the surface of significant immovable cultural relics, which can be presented to the public in an all-round and close range. They can be used as a new data collection method and data source for 3D GIS platforms [8].

Structure from Motion (SFM) is a 3D reconstruction method based on image sequence, which infers the geometric structure of a 3D scene by analyzing camera motion and the structure of objects in ordered or unordered images [9]. The strength of SFM lies in its ability to process image data from a variety of sources, be it professional photogrammetry or photos taken with common consumer-grade cameras or even video clips. Compared with traditional 3D modelling techniques such as laser scanning or manual modelling, SFM does not need to obtain detailed scene information or camera parameters in advance, which significantly reduces the complexity and cost of 3D reconstruction. In addition, SFM greatly reduces the need for human intervention. Traditional methods usually rely on professional operations and complex parameter configuration, while SFM effectively improves work efficiency and scalability by automating the process of feature extraction and matching. This automation not only accelerates the speed of data processing but also makes 3D modelling techniques more popular, bringing revolutionary impact to academic research, industrial applications, and consumer markets. Therefore, this paper aims to use the SFM algorithm to reconstruct the model of the test object quickly and efficiently. Firstly, the feature points of each input image were extracted, and the feature point descriptor was calculated.

Then, the feature points are matched and adjusted one by one to obtain the last suitable image. In this process, the camera's internal and external parameters and the 3D information of the actual scene can be obtained so as to obtain the 3D model of the sparse point cloud. Then, the model was constructed and optimized based on multi-view photometric consistency and Multi-view Stereo (MVS). Tang et al. [10] respectively used COLMAP, OpenMVG and OpenMVS for 3D modelling of the same building. Although OpenMVS was added in the MVS stage, OpenMVG+OpenMVS finally obtained a higher accuracy 3D model than COLMAP, but in the SFM process, COLMAP outperforms COLMAP+OpenMVG. Therefore, in this paper, COLMAP is combined with OpenMVS to ensure that the SFM stage can obtain a more accurate coefficient point cloud model while ensuring that the subsequent MVS can output a more complete and accurate dense point cloud model and grid model. In order to verify that COLMAP+OpenMVS is a modelling method that can achieve high-precision modelling in a short time, OpenMVG+OpenMVS and COLMAP+CMVS-PMVS are compared. In the field of medical imaging, Wang et al. [11] developed an innovative method called IDEAS for multi-energy computed tomography (CT) reconstruction, which leverages sparsity to enhance image quality. Their method integrates local sparsity, nonlocal self-similarity, and spectral correlation through nonlocal low-rank Tucker decomposition and multi-task tensor dictionary learning. This approach significantly improved reconstruction quality across numerical simulations and physical experiments, showcasing substantial advancements over existing methods. Wang et al. [12] developed the Hybrid-Domain Integrative Transformer Iterative Network (HITI-Net) for enhancing image reconstruction. This method effectively combines model-based and deep learning approaches to improve image quality and accuracy in sparse-view CT imaging, showing superior performance over existing techniques.

In this paper, we focus on using information from dense point clouds to improve the detail and accuracy of 3D models. For the use of dense point clouds, In the realm of document image binarization, there are also fruitful results. Nguyen [13] effectively combined the advantages of 2D image processing and 3D point cloud data, we provide a more robust solution to the challenge of severely degraded text binarization.

II. RELEVANT THEORIES

SFM is to extract features from a series of ordered or unordered images, estimate camera pose and 3D point position, and then reconstruct the 3D structure of an object or scene using Multi View Stereo (MVS). This section will conduct research and analysis based on the relevant principles of SFM and MVS.

A. STRUCTURE FROM MOTION

SFM (Structure from Motion) is one of the core technologies of 3D reconstruction based on UAV mapping [14]. It mainly includes three stages:(1) extracting frames and matching frames according to the common features of two adjacent frames; (2) camera trajectory and pose estimation; (3) Recovering the 3D structure of the target scene based on the information obtained in the previous two steps. Incremental SFM is often used as a benchmark for comparison between different algorithms due to its robustness and high accuracy in 3D model reconstruction. Since this paper focuses on the accuracy of 3D models, incremental SFM is selected for the research work.

B. MULTI VIEW STEREO

After SFM has completed the sparse reconstruction of a specific object or scene, it can obtain the internal and external parameters of the camera, the sparse 3D points, and the corresponding 2D points of the image. However, this information is only represented in the 3D space by the sparse discrete point cloud, and the result can not represent the object or scene completely and continuously. Hence, it needs to be converted into a dense representation of the object or scene. Then, the fine 3D reconstruction is completed. However, MVS is a process of stereo-matching the scene by using the internal and external parameters of the camera estimated by

SFM to find the points with photometric consistency in the space [15].

MVS algorithms are usually classified into voxel-based algorithms, point cloud diffusion-based algorithms, and depth map fusion-based algorithms according to the representation of the scene. We employ COLMAP (a generic motion recovery architecture and multi-view stereo pipeline.) OpenMVS is an open-source multi-view geometric reconstruction tool that combines information from multiple image views to produce a 3D model. Both use the MVS algorithm based on depth map fusion. Firstly, the domain image is selected for each image to construct a stereo image pair, and then the depth map of each image is estimated, and then the depth map of each image is fused to extract the object surface [16]. The three methods differ in the details of their implementation. COLMAP employs a block-matching approach [17], where the image is divided into small blocks (usually rectangular areas of fixed size), and then stereo matching is performed on these blocks. This block-matching method can capture the smaller scale depth variation so that the depth map can capture the details of the object more accurately. When selecting domain image pairs, OpenMVS should consider that the images have sufficient similarity, and match feature points as much as possible. The angle should be significant to ensure the accuracy of the reconstruction of the 3D model. The region growing method is used to establish a priority queue according to the confidence of the reconstruction. Then, the depth is estimated from the initial sparse feature points, and the beam adjustment optimization is carried out for each seed point. After the optimization, each point is judged. If there is no depth value in the field and the confidence of the current pixel is higher than a specific range of the field pixel, its field pixel is added to the queue as a seed point. Finally, a complete dense 3D point cloud model was obtained by fusing all point cloud data [18]. CMVS (Cluster Multi-View-Stereo) -patch-based method PMVS (Patch Multi-View-Stereo) is a point cloud diffusion-based method. Firstly, the CMVS algorithm is used to group camera views into multiple groups. Each group contains camera views with overlapping regions. The goal is to decompose large-scale multi-view scenarios into smaller subproblems to reduce computational complexity and memory requirements. Subsequently, the PMVS algorithm was used for each group. The local patch information in the image was used to calculate the depth relationship between the patches through feature matching and geometric consistency verification. The local dense point cloud was generated [19].

C. SURFACE MODELING

The basic principle of surface modelling is to approximate and reconstruct continuous geometric surfaces through mathematical algorithms and numerical methods according to the geometric relationship of discrete data points or grids [20]. Depending on the different forms and characteristics of the data, a variety of surface modelling methods can be used, including triangulation and implicit functions. Among them, triangulation is based on the spatial distribution of discrete data points, and a continuous triangular mesh is formed by connecting the points into triangles. The implicit function method uses the implicit function to represent the surface, in which the zero value face of the function corresponds to the surface, and the implicit function method can reconstruct the surface through interpolation, statistical learning, and other techniques [21].

III. METHODOLOGY

A. EXPERIMENTAL PROCEDURE

In our research, an image acquisition strategy combining a ground-based handheld device with a UAV was used to achieve a comprehensive 3D reconstruction of the study object. Ground data were collected using an iPhone 8 with a built-in Sony IMX315 CMOS sensor. The acquisition of each image is strictly controlled at 1 meter from each other, and the operator maintains the camera at a stable horizontal height to reduce the potential impact of perspective differences on the accuracy of the 3D model. Aerial image acquisition was performed by a DJI M300 RTK drone equipped with a Chansi P1 sensor. The sensor has a high resolution of 45 million pixels and a large full-frame size of 35.9 × 24 mm, which significantly improves the detail capture ability of aerial images. To ensure adequate image data acquisition, the flight path of the UAV is set by accurate pre-planning software, where the heading overlap rate is set to 80% and the side overlap rate is set to 55%. This high overlap rate configuration ensures that sufficient image data is obtained from multiple angles to support efficient matching and 3D processing of subsequent images.

In the image processing stage, the SIFT algorithm [22] automatically identifies matching feature points from multi-view images. The algorithm extracts features based on local interest points, which can extract local features invariant to rotation, scale scaling, and brightness change from images. In addition, these features exhibit a certain degree of stability to viewpoint changes, affine transformations, and noise.

In the actual data acquisition process, the size of the objects in the image may be different due to the camera distance, shooting angle, and other factors. To deal with the different scales of features that may be present in an image, such as large and small objects, feature points need to be detected at multiple scales. The Gaussian pyramid generates a series of images of different scales by applying Gaussian blur to the image layer by layer, which facilitates the detection of features at various scales and ensures the scale invariance of the detected feature points. The image scale space is represented by L(x,y,σ), as in Equation (1):

$$L(x, y, \sigma) = G(x, y, \sigma) * I(x, y) \tag{1}$$

where σ denotes the spatial scaling factor and represents the degree of continuity of the image, its value is proportional to the degree of continuity of the image function. Here, I(x,y) denotes the input image. G(x,y,σ) represents the Gaussian kernel, and the equation is given in (2):

$$G(x, y, \sigma) = \frac{1}{2\pi\sigma^2} e^{-\frac{x^2+y^2}{2\sigma^2}} \tag{2}$$

Furthermore, all the extreme points in the image are calculated. This process needs to focus on the edges and corners of the image, which are the locations where the important feature points often appear. The Difference of Gaussians (DoG) operation [23] can effectively highlight edges and corners in an image, and DoG is approximately Laplacian of Gaussian (LoG). By detecting each pixel in the scale space of the DoG and comparing it with 26 neighboring points, the algorithm can effectively improve its accuracy. Thus, the detection of extreme points is realized. This ensures that extreme values are detected in both scale space and 3D space, simplifying computations while preserving feature points. The calculation equation is as follows:

$$D(x, y, \sigma) = [G(x, y, k\sigma) - G(x, y, \sigma)] * I(x, y) \tag{3}$$

where k is the scaling factor.

The feature points with strong edge response change sharply on an edge but change less in the vertical direction, which is not conducive to stable feature matching. If these unstable feature points are retained, they will lead to an increased probability of subsequent mismatching, inaccurate positioning, and poor robustness of the model. Therefore, it is necessary to eliminate such points to obtain more stable feature points, and curve fitting of the DoG function in scale space (Equation (4)) is needed.

$$\frac{\operatorname{trace}(\mathbf{H})}{\operatorname{Det}(\mathbf{H})} < \frac{(r+1)^2}{r}, r = 10 \tag{4}$$

where H is the Hessian matrix and r is the threshold parameter.

In order to determine the position L(x,y,σ) of a feature point in scale space, a keypoint descriptor must be generated. In addition, the magnitude direction m(x,y) and Angle θ(x,y) of the influence caused by an area of radius 3 × 1.5σ centered at the feature point need to be calculated to ensure rotation invariance of the feature point:

$$m(x, y) = \sqrt{[L(x+1,y)-L(x-1,y)]^2+[L(x, y+1)-L(x, y-1)]^2} \tag{5}$$

$$\theta(x, y) = \alpha \cdot \arctan \frac{L(x, y+1) - L(x, y-1)}{L(x+1,y) - L(x-1,y)} \tag{6}$$

Feature descriptors are extracted in the neighborhood of keypoints for image matching. The descriptor captures the texture information around the key points, which gives the feature points strong discrimination ability and ensures that they still have robust discrimination even in the presence of illumination changes and perspective changes. A 16×16 window is extracted centered on the main direction. The gradient magnitude and gradient direction are calculated for each pixel within the window. Subsequently, eight orientation gradient histograms are generated for each 4 × 4 subwindow. Each feature point consists of 16 seed points to form a keypoint. Finally, 4×4×8=128 dimensional SIFT feature vectors were generated. Through Euclidean distance calculation, all the extracted feature points are matched exactly, as shown in Equation (7):

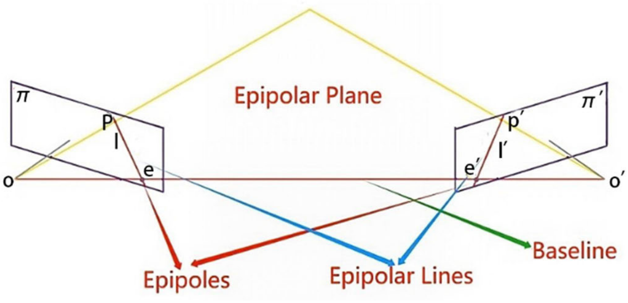

FIGURE 1. Schematic diagram of the Epipolar geometry model.

$$\operatorname{dist}(\mathbf{d}_1, \mathbf{d}_2) = \sqrt{\sum_{i=1}^{128} (\mathbf{d}_{1,i} - \mathbf{d}_{2,i})^2} \tag{7}$$

where dist ($\mathbf{d}_1$, $\mathbf{d}_2$) is the distance between descriptors $\mathbf{d}_1$ and $\mathbf{d}_2$, and $\mathbf{d}_{1,i}$ and $\mathbf{d}_{2,i}$ represent the ith component of descriptors $\mathbf{d}_1$ and $\mathbf{d}_2$, respectively. For each feature point descriptor, find the two closest descriptors: the nearest neighbor $\mathbf{d}_{2,\text{nearest}}$ and the second nearest $\mathbf{d}_{2,second\text{nearest}}$. The distance ratio of the nearest neighbor and the second nearest neighbor is calculated by equation (8), and the distance of the two closest feature descriptors is compared to screen out reliable matches.

$$\frac{\operatorname{dist}\left(\mathbf{d}_1, \mathbf{d}_{2,\text{nearest}}\right)}{\operatorname{dist}\left(\mathbf{d}_1, \mathbf{d}_{2,\text{secondnearest}}\right)} \tag{8}$$

RANSAC [24] geometric verification is performed on the feature points. Firstly, k sample points are randomly selected from the sample set, and the current model parameters are estimated according to the mathematical model of sample mapping. The error threshold t is set, and the matching error of all sample points is calculated according to the current model estimate value. The sample points whose error is less than the threshold t is judged as interior points. Otherwise, they are exterior points, and the interior and interior points under the current model estimate value are recorded. Then, the maximum number of inner points is updated, and the above steps are repeated until the number of iterations exceeds M, or the maximum number of inner points exceeds the specified threshold of the number of inner points, and the loop is terminated. Finally, the model parameters were re-estimated for the interior point set corresponding to the most significant interior point, and the optimal model estimation value was obtained.

The test object's sparse point cloud is obtained by sparse reconstruction using the results obtained after feature matching and geometric verification. The Epipolar geometry model is used to estimate the camera pose and motion trajectory [25]

(Figure 1). Assuming that π and π' are two images captured by the camera, the point P in space is imaged in both captured images, p and p' are the projections of P on π and π', respectively, and the camera optical center of the two images are o and o'. Connecting the camera optical center o, o', the resulting concatenation is the baseline. Between the baseline and the space point P forms the Epipolar plane, and the Epipolar line l, l' between the Epipolar plane and the image plane of the image π, π' becomes the Epipolar line.

The homonym of a point P in 3D space that matches in two images must fall on a certain pole line, so the search for the correspondence between two images is narrowed from 2D to 1D under the pair of Epipolar constraint, making it much less difficult.

To remove the bias in this process and optimize the estimation of camera pose and 3D point cloud with higher accuracy and precision, the Motion Recovery Structure (SFM) incorporates bundle adjustment (BA) for each image [26]. The objective equation of BA is as follows.

$$\min_{\mathbf{P}_j,\mathbf{M}_i} \sum\nolimits_i^n \sum\nolimits_j^m \|\mathbf{m}_{ij} - f\left(\mathbf{p}_j, \mathbf{M}_i\right)\| \tag{9}$$

where $\mathbf{m}_{ij}$ denotes the ith observation on the jth image of an image point observation in the 2D phase plane, which is associated with the camera matrix $\mathbf{p}_j$ of the jth image, $\mathbf{p}_j = K[\mathbf{R}_j \mid \mathbf{t}_j]$.

In order to improve the details and accuracy of the 3D model, the sparse point cloud is transformed into a dense point cloud by the processing of CMVS-PMVS and OpenMVS. The CMVS-PMVS algorithm [27] optimizes the density and quality of point clouds by segmenting a large-scale image set and applying stereo-matching techniques in each subset. The algorithm generates dense point clouds using a point cloud diffusion method without using depth maps. In contrast, COLMAP and OpenMVS obtain high-resolution dense point clouds by finding photometric consistency depth map estimation algorithm [28] to generate depth maps based on disparity information.

The dense point cloud is finely meshed using the Poisson reconstruction method and Delaunay triangulation. Poisson modelling [29] uses the normal vector information in the point cloud to construct the Poisson equation and then solves the equation to obtain the 3D surface. The gradient field is defined for the normal vector corresponding to each input point cloud $\mathbf{p}_i$ as follows.

$$\mathbf{V}(\mathbf{x}) = \sum_i \delta\left(\mathbf{x}-\mathbf{p}_i\right)\mathbf{n}_i \tag{10}$$

where V(x) is the gradient field at position x; $\delta(\mathbf{x}-\mathbf{p}_i)$ is the Dirac delta function, which means that it has infinite value at point $\mathbf{p}_i$ and zero elsewhere.

The smooth implicit function f (as shown in Equation (11)) is reconstructed from the gradient field V(x), and by solving this Poisson equation, the implicit function f can be obtained, whose zero isosurface is the reconstructed surface.

$$\nabla \cdot \mathbf{V} = \nabla^2 \mathbf{f} \tag{11}$$

where ∇·V is the divergence of the gradient field V; $\nabla^2\mathbf{f}$ f is the Laplacian of the implicit function f. This method generates a smooth 3D surface model from discrete point cloud data.

The Delaunay criterion requires that any triangle in the triangulation's circumcircle not contain points from other point clouds. Therefore, Delaunay triangulation [30] ensures the mathematical optimality and geometric consistency of the mesh. Finally, the parameters of 3D models generated by different reconstruction techniques are compared to evaluate the effect and applicability of each method.

The methodology and workflow diagram of this study (Figure 2) details the whole process from data collection to final model generation, showing the technical implementation and achievement presentation of each stage.

B. DATASET AND MAJOR ASSESSMENT INDEX

1) THE ETH3D DATASET

The ETH3D dataset [31] is a dataset containing image sequences of multiple scenes and 3D reconstruction results associated with them. These image sequences were acquired in different environments, including scenes such as indoor, outdoor, and urban streets. The dataset provides 3D reconstruction results such as camera pose estimation, dense point clouds and surface meshes for each scene. The evaluation tools and metrics that come with the ETH3D dataset are used to evaluate and compare the performance of different 3D reconstruction algorithms, which can help researchers in accurate performance evaluation and facilitate algorithm comparison and analysis.

2) THE MAJOR ASSESSMENT INDEX

(1) RMSE [32] is the root mean square error, which measures the difference between the predicted value and the actual value of the data, and is calculated as follows:

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(\hat{y}_i - y_i)^2}{n}} \tag{12}$$

where $\hat{y}_i$ is the predicted value; y_i is the observed value; n is the number of observations.

(2) Mean track length [33] is an evaluation metric used in object tracking or trajectory prediction tasks to measure the difference between the predicted trajectory and the actual trajectory. It calculates the average distance between the predicted trajectory point and the actual trajectory point at each time step.

The location data of the device or camera is collected during the scanning process. This usually involves the spatial coordinates (x, y, z) of each point in the time series and computing the distance di between neighboring points:

$$d_i = \sqrt{(x_{i+1} - x_i)^2 + (y_{i+1} - y_i)^2 + (z_{i+1} - z_i)^2} \tag{13}$$

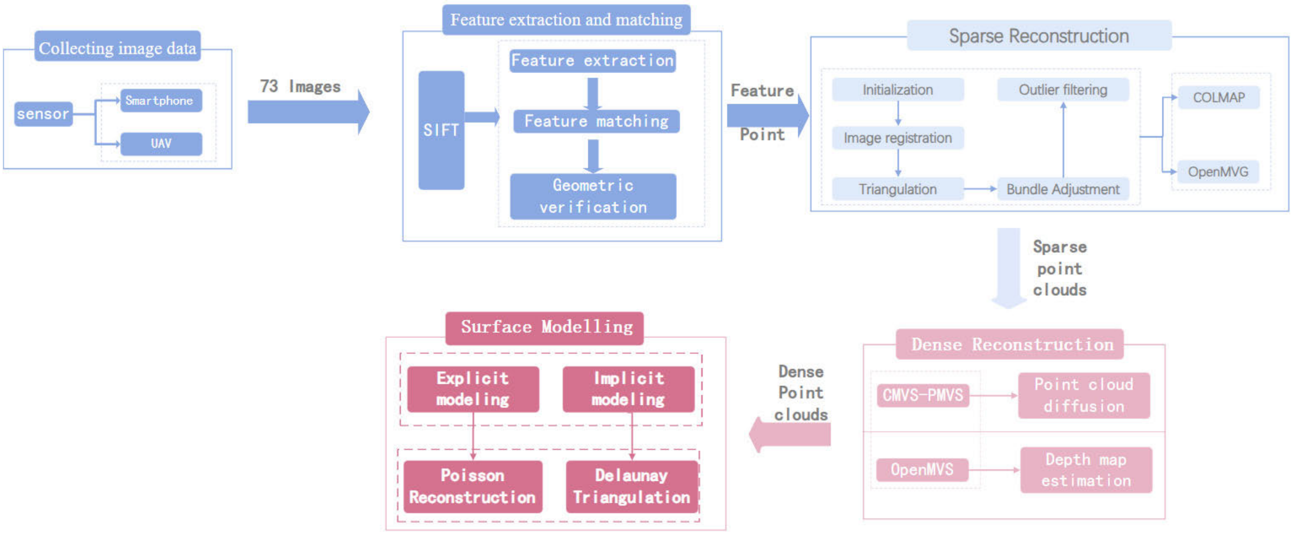

FIGURE 2. Flow chart of the complete work.

Then combining them to calculate the average:

$$MTL = \frac{\sum_{i=1}^{n-1} d_i}{n-1} \tag{14}$$

where n is the number of time steps, and $\mathrm{d_i}$ is the distance between the predicted trajectory point and the true trajectory point at the ith time step.

(3) The reprojection error [30] is the distance between the position of the 3D point projected onto the image plane and the actual observed position. A smaller reprojection error indicates a higher accuracy of 3D reconstruction.

$$\mathrm{MRE} = \left(\frac{1}{\mathrm{N}}\right) \star \Sigma \left|\left|x_i - x'_i\right|\right| \tag{15}$$

where N is the total number of 3D points; x_i is the actual observed pixel coordinate of the ith point; x'_i is the reprojected pixel coordinate of the ith point. ||.|| is Euclidean distance (modular length).

(4) SSIM (Structural Similarity Index) [34] is a metric used to measure the similarity between two images. It is a measure based on visual perception used to compare the similarity of two images. It takes into account three aspects: brightness, contrast, and structure alterations, which are key factors affecting human visual perception. In 3D modelling, SSIM comparison between the model image generated by rendering and the actual image (or a high-quality reference image) can provide a visual assessment of the accuracy of the model. SSIM values range from [0, 1] in real cases, where closer to 1 indicates that the model is highly similar to the original image. SSIM is calculated as follows:

$$SSIM\ (x, y) = \left(\frac{2\mu_x\mu_y + C_1}{\mu_x^2 + \mu_y^2 + C_1}\right) \cdot \left(\frac{2\sigma_{xy} + C_2}{\sigma_x^2 + \sigma_y^2 + C_2}\right) \cdot \left(\frac{\sigma_{xy} + C_3}{\sigma_x\sigma_y + C_3}\right) \tag{16}$$

Among them: μ_x and μ_y are the average values of the images x and y, respectively. σ_x and σ_y are the standard deviations of the images x and y, respectively. σ_{xy} is the covariance between images x and y.

(5) The F1 score [35] is the harmonic mean of accuracy and completeness, and a higher score indicates a better performance of the algorithm. The F1 score is calculated as follows:

$$F1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall}) \tag{17}$$

where precision is the precision and Recall is the recall ratio.

(6) Mean distance from mesh to point cloud [36] measures how accurately the reconstructed grid covers the actual point cloud data. For each point P_i in the point cloud, the distance to its closest point on the grid is calculated. If Q is a set of points on a grid, then for each point P_i in the point cloud P, the distance from the grid $d(p_i, Q)$ is calculated as follows.

$$d\ (p_i, Q) = \min_{q \in Q} \parallel p_i - q \parallel \tag{18}$$

The mean distance from the point cloud to the mesh is defined as follows.

$$\mathrm{AverageAccuracy} = \frac{1}{|P|} \sum_{p_i \in P} d\ (p_i, Q) \tag{19}$$

where$|P|$ is the number of points in the point cloud.

(7) Mean distance from point cloud to mesh [37] measures the coverage degree of the original point cloud data in the reconstructed model, that is, the completeness. That is the average of the distance from each point in the original point cloud to the closest point in the point cloud is transformed by the reconstructed model.

For each point q_j in the point cloud, the distance to its closest point on the grid is calculated. If Q is a set of points on a grid, then for each point q_j in the point cloud P, the distance

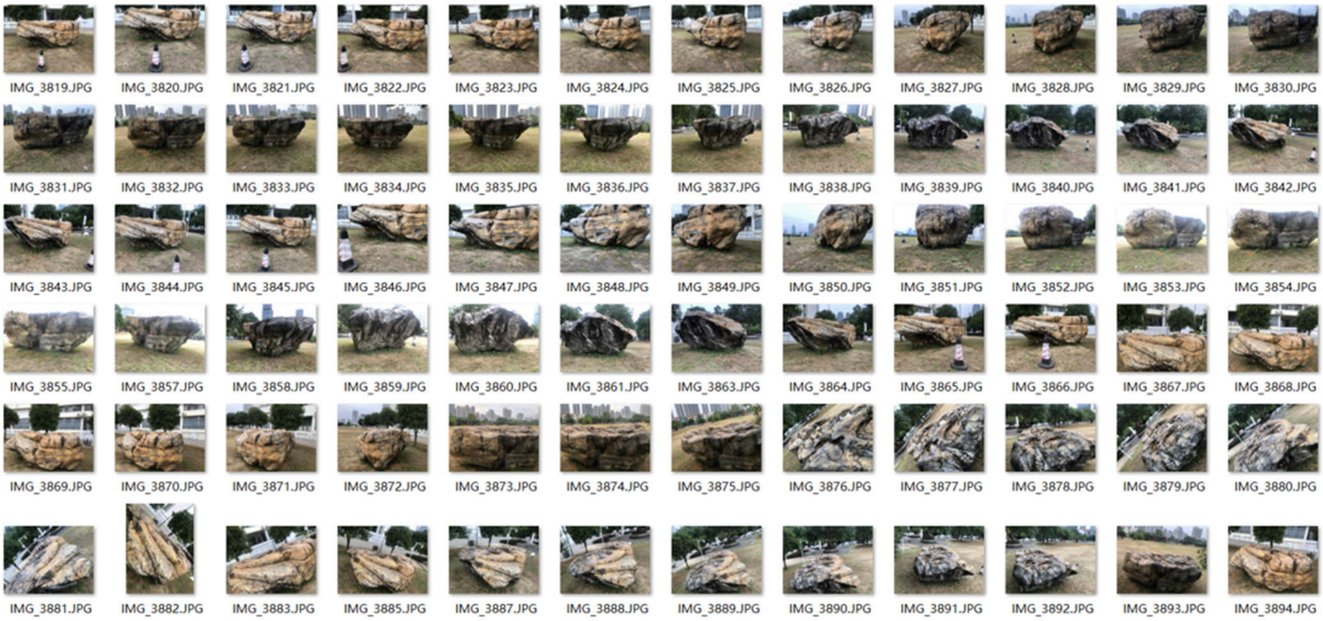

FIGURE 3. Part of the image data.

from the grid $d(q_j, P)$ is calculated as follows.

$$d\left(q_j, P\right) = \min_{p_i \in P} \parallel q_j - p_i \parallel \tag{20}$$

The mean distance from the mesh to the point cloud is defined as follows.

$$\text{Average Completeness} = \frac{1}{|Q|} \sum_{q_j \in Q} d\left(q_j, P\right) \tag{21}$$

where $|Q|$is the number of points on the mesh.

The average distance from the point cloud to the grid and the average distance from the grid to the point cloud can be obtained from Equations (19) and (20), and the accuracy and completeness of the model can be calculated, as shown in Equations (22) and (23). In this paper, the maximum distance is 0.5m.

$$\begin{aligned}&\text{Accuracy}\% \\ &= \left(1 - \frac{\text{Average distance from mesh to point} - \text{cloud}}{\text{Max distance in dataset}}\right) \\ &\qquad \times 100\end{aligned} \tag{22}$$

$$\begin{aligned}&\text{Completeness}\% = \\ &\left(1 - \frac{\text{Average distance from point} - \text{cloud to mesh}}{\downarrow \text{Max distance in dataset}}\right) \\ &\qquad \times 100\end{aligned} \tag{23}$$

C. PROPOSED METHOD

In this paper, we propose a 3D reconstruction method that combines the technologies of COLMAP and OpenMVS, effectively addressing the limitations of traditional reconstruction methods in maintaining model topological consistency. This method meticulously leverages the efficient capabilities of COLMAP in feature extraction and matching, which accurately identifies and matches complex features from multiple viewpoints, thereby providing a solid data foundation for subsequent 3D reconstruction. Simultaneously, by incorporating OpenMVS's optimized multi-view stereo reconstruction technology, our approach further enhances the process from dense point clouds to fine meshing, particularly in automatically filling and smoothing common holes and discontinuities in point cloud data, greatly improving the model's integrity and reliability. Moreover, this integration of technologies ensures a more efficient and

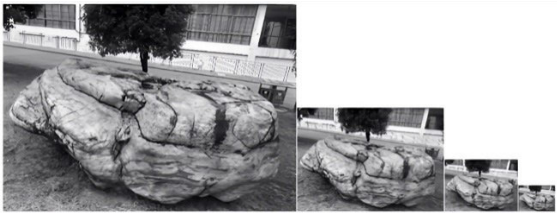

FIGURE 4. Image of Gaussian pyramid.

FIGURE 5. Image of Laplacian pyramid.

FIGURE 6. The main direction for the feature points.

coherent data flow throughout the reconstruction process, optimizes resource allocation, reduces computation time, and significantly enhances the final model's quality and practicality. The complete process is shown in Figure 2.

IV. RESULTS

A. DATA ACQUISITION

We selected a natural stone block as the subject of our case study. The block has an approximate area of 18 m^2. To capture detailed image data of the object, we used a mobile phone for surround photography within the study area. For the top area, a UAV was employed. The mobile phone used in the study is equipped with a Sony IMX315 12.2M 1/2.93 1.22um stacked CMOS camera sensor. The UAV, on the other hand, is equipped with a Zenith P1 sensor, which has a sensor size of 35.9 × 24 mm and 45 megapixels of effective pixels. To ensure comprehensive coverage, we set the heading overlap rate to 80% and the side overlap rate to 55%. Some of the acquired images are depicted in Figure 3.

B. 3D RECONSTRUCTION

1) FEATURE EXTRACTION AND MATCHING

Gaussian pyramid scale Spaces of different scales are obtained by Gaussian blurring through the Gaussian pyramid, as shown in Figure. 4.

To extract features, the DoG (difference of Gaussian) operation is performed to highlight features at essential positions such as edges and corners of the image data, as shown in Figure 5.

Calculating the gradient direction and amplitude determines the main direction of the key point, so the descriptor is consistent in the case of rotation. Figure 6. shows a schematic diagram of the main directions of random feature points.

The neighborhood of each keypoint, it is divided into 4 × 4 grids, and the gradient histograms of 8 directions are computed for each grid to form a feature vector of 4 × 4×8=128 dimensions. Figure 7 shows the diagram of the feature vector of the keypoint descriptor.

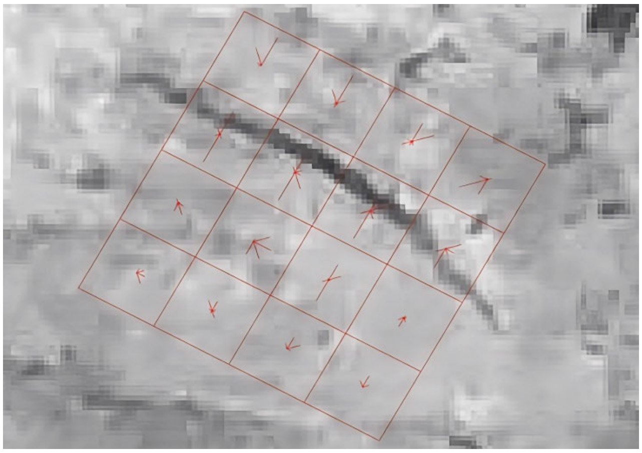

FIGURE 7. The main direction for the feature points.

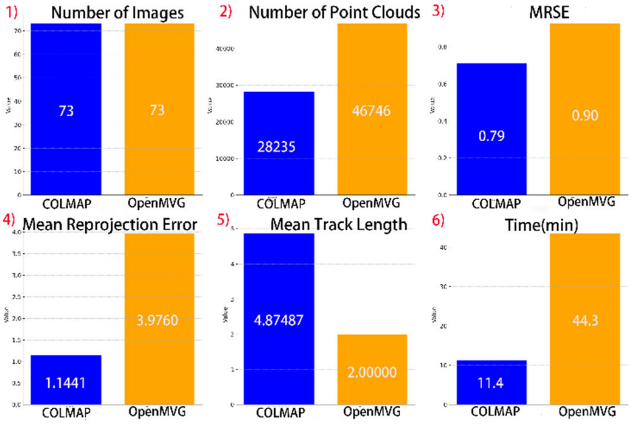

FIGURE 8. 1) Number of images processed; 2) number of point clouds; 3) RMSE error; 4) Average reprojection error; 5) average track length; 6) Total time (min).

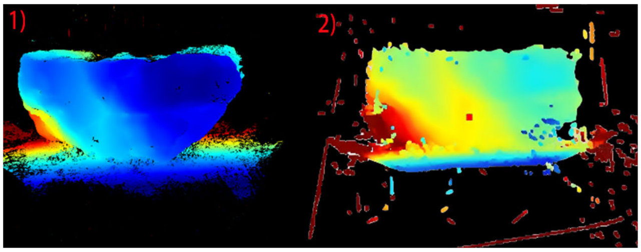

FIGURE 9. 1)The depth map generated by COLMAP; 2) The depth map generated by OpenMVG+OpenMVS.

2) SPARSE RECONSTRUCTION

In the process of sparse reconstruction, there are several factors that can affect the results. These factors include: 1) The reverse projection of 3D spatial points can introduce bias to the matching reconstruction. 2) External conditions, such as the inaccuracy of measuring instruments or human factors,

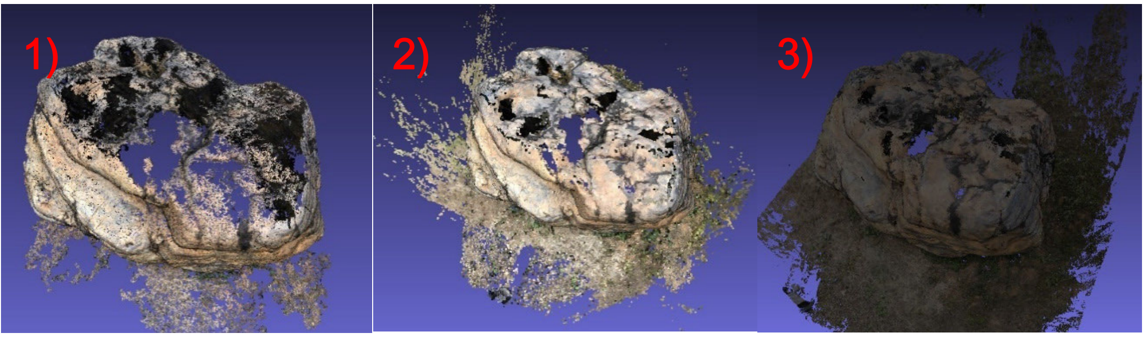

FIGURE 10. 1)Dense point cloud scene generated by CMVS-PMVS; 2) Dense point cloud scene generated by COLMAP; 3)Dense point cloud scene generated by OpenMVG+OpenMVS.

can lead to biased measurement results. 3) The number of observations is often greater than the minimum required for determining quality, resulting in redundant observations and variations among them. After the adjustment optimization of the beam method, the point cloud was triangulated. Finally, the iterative global adjustment optimization of the beam method was carried out to optimize the pose of the existing camera and the 3D sparse point cloud coordinates. See Figures 8 for sparse reconstruction parameters for COLMAP and OpenMVG.

In Figure 8, sparse point cloud refers to the number of 3D points obtained by feature matching and initial triangulation, which reflects the number of feature points identified in the preliminary reconstruction. A higher number means more features are captured. The number of point clouds generated by COLMAP is 28,235, far less than the 46,746 generated by OpenMVG. The mean reprojection error of COLMAP in the SFM stage is 1.1441, which is significantly better than 3.9760 of OpenMVG. The mean track length of COLMAP is 4.87487, which is much higher than that of OpenMVG of 2.0000. The RMSE of COLMAP is 0.79, while OpenMVG is 0.90, which means that the reconstruction result of the latter is very different from the actual data. In terms of time, COLMAP takes much less time than OpenMVG in the SFM process due to its powerful global optimization framework and GPU acceleration of key steps.

3) DENSE RECONSTRUCTION

Depth maps of COLMAP and OpenMVS were generated respectively on the basis of two coefficient point cloud data (Figure 9).

However, CMVS-PMVS generated dense point clouds by point cloud diffusion, so there is no depth map. COLMAP and OpenMVG+OpenMVS visualized the estimated depth map of each image in the MVS stage Figure 9 (1) demonstrates that COLMAP generates a more complete depth map with fewer noise, holes, and better continuity. This method also produces accurate depth estimates. On the other hand, Figure 9 (2) shows a depth map with more noise and poor edge continuity. This can be attributed to the fact that OpenMVS outputs redundant dense point clouds, which increases the noise in the depth map. The various methods for densely reconstructing the scenario are depicted in Figure 10. A careful examination of Figure 10 reveals that the CMVS-PMVS algorithm results in numerous gaps in the weakly textured region located at the top of the object under study. These gaps occur due to the algorithm's tendency to halt expansion when encountering points with low confidence intervals in weakly and repetitively textured regions. In contrast, both COLMAP and OpenMVG + OpenMVS exhibit higher scene completeness, particularly in weakly textured areas. Although the dense point cloud scene generated by OpenMVG+OpenMVS contains more redundant points, the most comprehensive 3D point cloud scene can be obtained by manually eliminating the surrounding clutter.

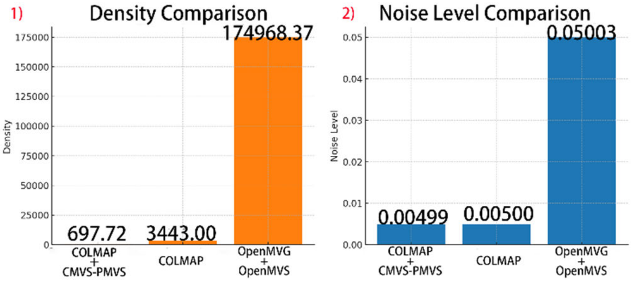

FIGURE 11. Point cloud density as well as noise level plot.

The point cloud density is obtained by counting the number of point clouds in a unit volume using the total number of point clouds. At the same time, the noise level is calculated by the distance of the point to a reference surface or fitted model. Figure 11 illustrates the point cloud density as well as the noise level for generating dense point clouds in different ways. The point cloud density of CMVS-PMVS is 697.7234, which is the lowest density among the three, indicating the worst point cloud coverage. This is because the algorithm splits the original image set into smaller subsets during the clustering process. While this reduces the number of images that need to be processed for a single PMVS instance, it also limits the number of viewpoints that can be used for reconstruction. Reducing the number of viewpoints may lead to insufficient coverage of some spatial regions, which affects

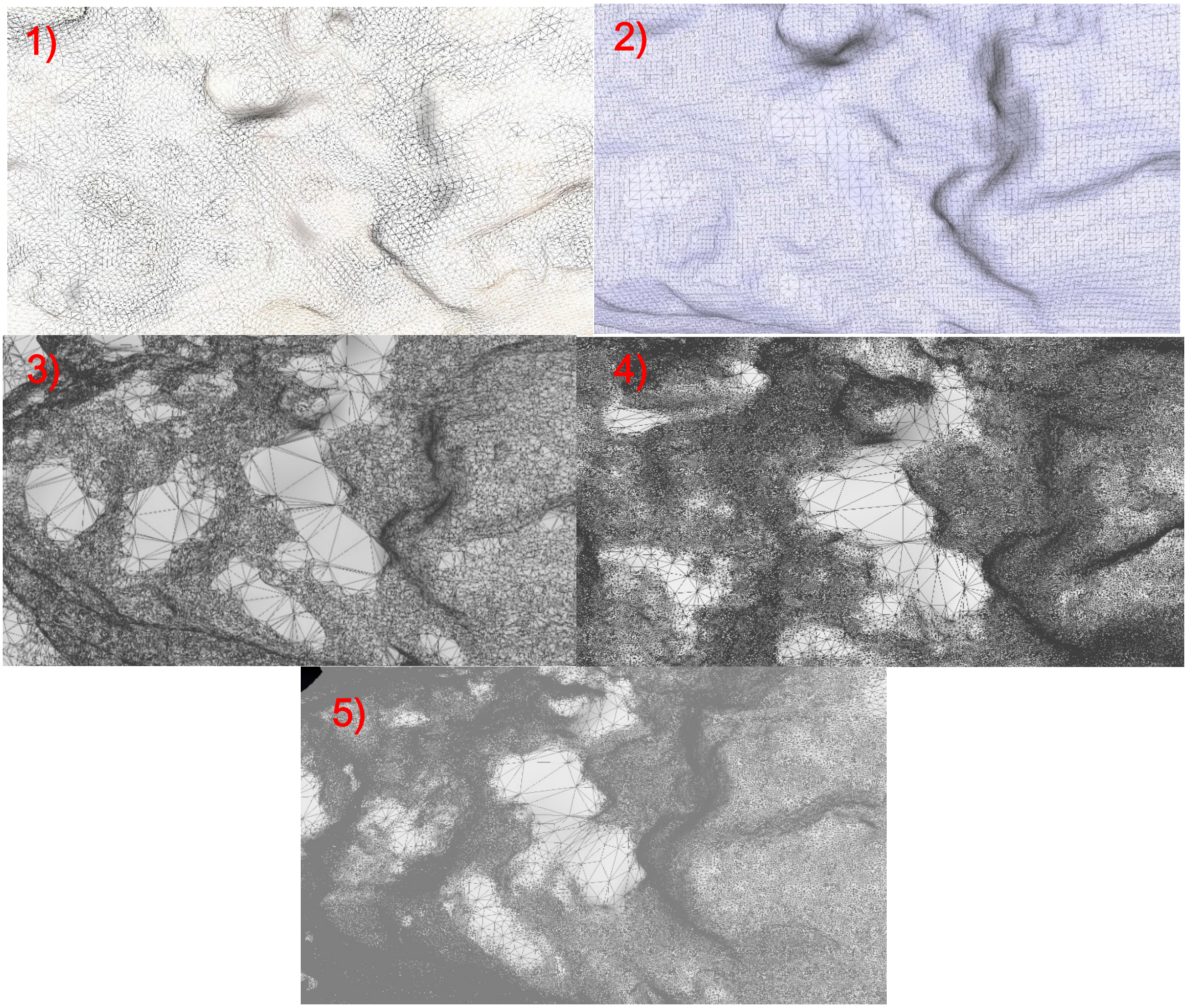

FIGURE 12. Mesh models of different method.

the density of the point cloud. Second, clustering can handle images that are spatially adjacent but assigned to different groups. This segmentation may lead to loss of information in areas close to the group boundary due to lack of sufficient view overlap during the reconstruction process, which in turn affects the continuity and density of the point cloud.

Although the data sparsity strategy adopted by CMVS-PMVS can significantly improve processing speed and reduce resource requirements, the feature matching process of this strategy is carried out independently within each cluster and may not utilize global information across groups. This localized approach may miss important global features, especially those that span multiple clustering regions. This will not only affect the point cloud density but also reduce the matching quality and increase the error in the reconstruction process. Therefore, its generated point cloud has a low density. Although the density is low, the noise level is 0.0049, which indicates that the generated point cloud is of high quality and less error.

3D reconstruction using COLMAP yielded a good point cloud density (3443.0) and a low noise level (0.0049). That is, COLMAP can accurately match the corresponding pixels in the image, which not only increases the density of the reconstructed point cloud but also ensures low error and low noise. Even in the dense reconstruction phase, COLMAP uses beam adjustment to optimize the overall model and reduce errors to achieve high-quality reconstruction results.

Since OpenMVG mainly focuses on how to extract features from the image to the greatest extent in the SFM stage. Although this can better generate a very dense point cloud, this strategy will lead to the extraction of more low-quality feature points, which will affect the subsequent OpenMVS input feature matching errors in the MVS process. These errors will be amplified in the depth map generation process,

TABLE 1. Table of parameters for Poisson modelling.

Methods	Number of vertices	Number of meshes	Number of vertices after optimization	Number of meshes after optimization	Accuracy	Completeness	Time (min)
COLMAP	211142	421216	211142	421216	80.6%	83.0%	77.4
COLMAP+CMVS-PMVS	79233	157571	79233	157571	39.2%	19.0%	80.1
COLMAP+OpenMVS	1017714	2023196	1019416	2032864	97.6%	99.8%	88.3
OpenMVG+OpenMVS	5980772	11955540	735761	1390491	95.6%	97.2%	228.0

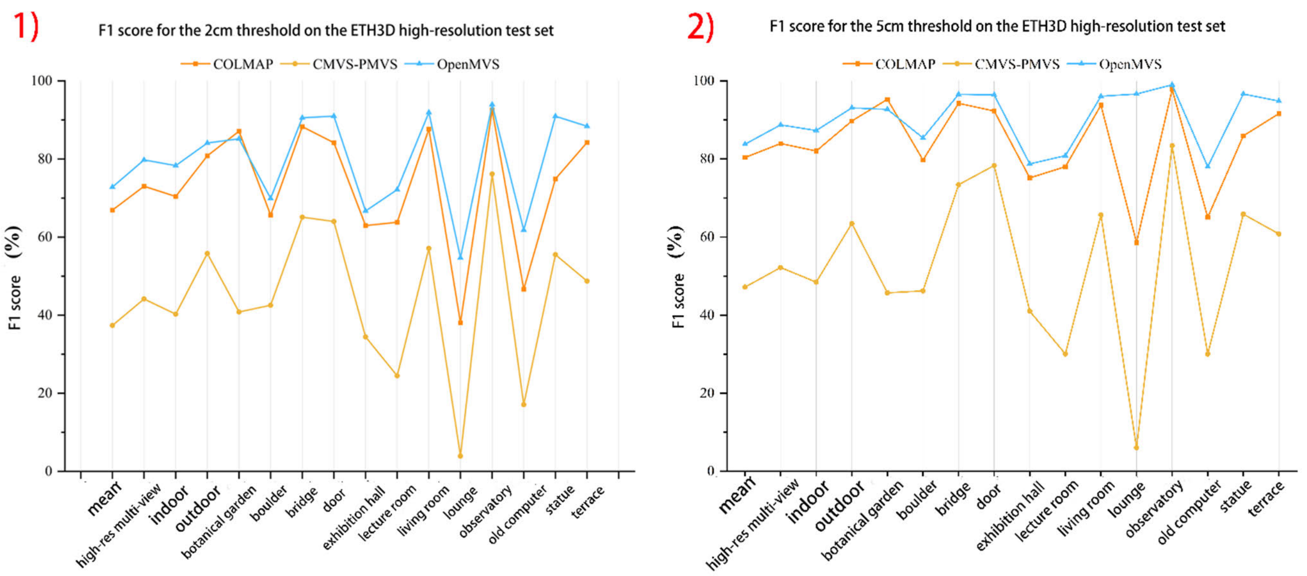

FIGURE 13. F1 score on the ETH3D high-resolution test set.

resulting in a much higher noise level than the previous two methods.

4) SURFACE MODELING

Figure 12 (1) and (2) illustrate the 3D models generated by COLMAP and CMVS-PMVS Poisson Surface Reconstruction, respectively. The stone surface model produced through the Poisson surface reconstruction method exhibits a precise fit to the model surface, resulting in a smooth reconstructed surface with a more uniform distribution of triangular mesh elements. According to Table1, the Poisson modelling approach generates a lower number of vertices and meshes, which may potentially affect the quality of subsequent texture mapping.

Figure 12 (3) shows the 3D model generated by COLMAP using Delaunay triangulation. Both Delaunay triangulation and Poisson modelling can approximate the original point cloud data and generate the whole stone mesh model. However, the details can be observed, and the mesh reconstructed by Delaunay triangulation is not evenly distributed in the weak texture area. Figure 12 (4) and Figure 12 (5) show the 3D models generated by OpenMVS using Delaunay triangulation in the dense point clouds generated by COLMAP and OpenMVG, respectively. Given that OpenMVS has the surface refinement function to re-optimize the established surface mesh, Therefore, it can better avoid the phenomenon of over-fitting caused by Delaunay triangulation. This function iteratively adjusts the position of the point cloud and the topology of the triangular mesh to improve the consistency between the point cloud and the reconstructed mesh, generating more accurate and continuous results. Figure 12 shows that whether based on COLMAP or OpenMVG, the 3D mesh optimized by OpenMVS is smoother and more detailed and makes great improvement in weak texture regions. The texture mapping of all white 3D models was completed to obtain the final 3D model results, and the 3D modelling parameters of the test object between different algorithms were obtained (Table 1). COLMAP and COLMAP+CMVS-PMVS do not

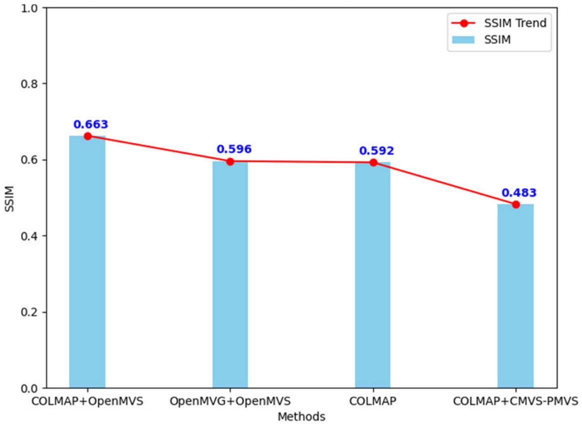

FIGURE 14. **SSIM values of 3D modelling in natural scenes by different methods.**

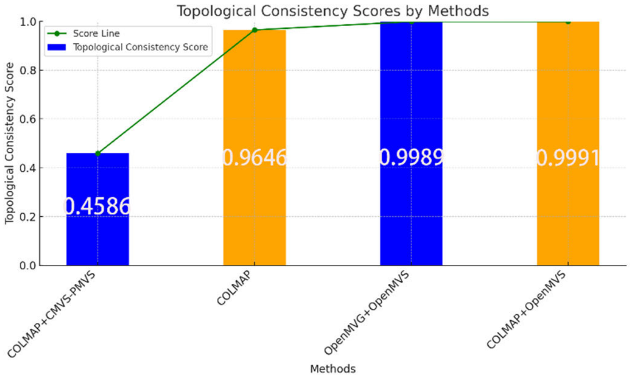

FIGURE 15. **Topological Consistency Score of 3D modelling in natural scenes by different methods.**

perform vertex and mesh optimization operations, and the data remains unchanged after optimization. The accuracy of the 3D model completed by the COLMAP method is 80.6%, which indicates that the geometry of the 3D model reconstructed by the COLMAP method is close to the geometry of the test object. The completeness is 83.0%, and indicating that there are many points in the point cloud that are poorly aligned with the mesh surface, which is caused by overly simplified or coarse meshes and perform the worst among all methods. Although the CMVS-PMVS algorithm can reduce the amount of computation through clustering, making it more efficient to deal with large amounts of data, the view data in this study only has 73 images, and the clustered model cannot accurately match most of the details of the original data. COLMAP+OpenMVS increases the number of vertices and meshes after optimization, indicating that the point cloud, vertices, and meshes obtained by COLMAP in the early stage of this method have high accuracy, and there is no need to delete redundant vertices and meshes. OpenMVS generates a large number of dense point clouds during MVS, so the OpenMVG+OpenMVS method generates a large number of vertices and meshes, but OpenMVS clears 87.70% of vertices and 88.37% of meshes during mesh optimization.

It indicates that the point cloud initially generated by the method is very dense and contains a large number of repeated or approximately overlapping vertices. These vertices contribute little to the visual quality of the final model but also significantly increase the processing complexity and resource consumption. At the same time, in the process of mesh generation, in order to avoid affecting the final accuracy and aesthetics of the model, a large number of low-quality vertices and meshes are deleted. That is, the efficiency of this method is extremely low. In terms of accuracy and completeness, COLMAP+OpenMVS achieves 97.6% accuracy and 99.8% completeness, which indicates that there is a good correspondence between the point cloud and the mesh, and the mesh generated by this method is of high quality. However, OpenMVG+OpenMVS performs slightly worse than COLMAP+OpenMVS but significantly better than COLMAP+CMVS-PMVS. In terms of speed, COLMAP implements parallel processing and makes efficient use of computing resources. In terms of accuracy, COLMAP has higher accuracy and robustness than OpenMVG. Unlike OpenMVG's exhaustive method, which compares all possible pairs of feature points based only on their similarity, COLMAP also considers geometric relationships between feature points, which leads to mismatching in repetitive or low-texture regions. Although the number of sparse point cloud generated by OpenMVG is larger, any small error will be amplified in the subsequent dense reconstruction process of OpenMVS. The research by Tang et al. [10] indicates that the OpenMVG+OpenMVS approach, while achieving good dense reconstruction accuracy, suffers from algorithmic instability, warranting further optimization. Wang et al.'s [37] study reveals that in the field of medical image 3D reconstruction, COLMAP outperforms in the sparse reconstruction stage, while OpenMVS excels in the MVS stage. Our experimental results align with these findings, further confirming the advantages of the COLMAP+OpenMVS combination.

V. CONCLUSION AND FORESIGHT

A. *ETH3D BENCHMARK DATASET EVALUATION*

The ETH3D dataset is a dataset containing image sequences of multiple scenes and 3D reconstruction results associated with them. These image sequences were acquired in different environments, including scenes such as indoor, outdoor, and urban streets. The dataset provides 3D reconstruction results such as camera pose estimation, dense point clouds and surface meshes for each scene. The evaluation tools and metrics that come with the ETH3D dataset are used to evaluate and compare the performance of different 3D reconstruction algorithms, which can help researchers in accurate performance evaluation and facilitate algorithm comparison and analysis.

Figure 12 illustrates the F1 scores calculated by COLMAP, CMVS-PMVS, and OpenMVS with a threshold of 2cm and 5cm, respectively, in the test dataset. It can be seen from Figures 12 (1) and 12 (2) that under the threshold of 2cm and 5cm, OpenMVS has the highest score in different scenarios, followed by COLMAP, but the gap with OpenMVS is not obvious. However, CMVS-PMVS has the lowest score for each scenario among the three. This is because CMVS-PMVS lacks effective point cloud data in weak texture regions and stops point cloud diffusion, so it is less flexible, which makes it unable to fit complex scenes well.

B. NATURAL SCENARIO DATA EVALUATION

In this study, all models were compared with the original image data under the same fixed viewing angle, and the highest SSIM value was calculated. As shown in Figure 13, the SSIM values of 3D models generated by all methods are slightly lower than the original image data. This is because this method requires strict control of the model viewing angle to be consistent with the image data, even with advanced algorithms and accurate data processing, due to the diversity of viewpoints captured by the original images and the inherent limitations of model rendering. These minor differences in viewpoint and detail can affect the structural similarity of the image, resulting in lower SSIM scores. Therefore, when evaluating 3D models, it is important to understand and acknowledge that SSIM values may be lower than expected due to viewpoint consistency issues. However, the results demonstrate that even with such inherent limitations, the 3D model generated by OpenMVG+OpenMVS is still the closest to the actual test object.

As previously mentioned, SSIM has limitations in evaluating the quality of 3D models, particularly in capturing geometric and topological features. Therefore, this paper introduces topological consistency as an alternative metric to assess 3D model quality more comprehensively. This study calculates a topological consistency score by normalizing the number of holes, connected components, and boundary edges. A score closer to 1 indicates better topological consistency of the model. Figure 14 illustrates the varying topological consistency scores among the four different methods. COLMAP+CMVS-PMVS has the lowest score (0.4586), suggesting it is prone to generating more topological defects when processing 3D models, possibly due to its limitations in handling multi-view geometric information. COLMAP achieves a higher score (0.9646), demonstrating better topological consistency. Methods incorporating OpenMVS perform even better: OpenMVG+OpenMVS achieves a near-perfect score (0.9989), while COLMAP+OpenMVS has the highest score (0.9991). This indicates that combining the efficient computation of COLMAP with the high-precision reconstruction of OpenMVS significantly improves the topological quality of the model.

We employ different software combinations for 3D reconstruction and compare their performance in the SFM and MVS phases. The analysis results show that in the SFM stage, the number of sparse point clouds generated by OpenMVG is 46,746, which is 1.66 times that of 28,235 generated by COLMAP. Although OpenMVG captured features from more subjects at this stage, its processing pipeline was more complex, resulting in a significant increase in total processing time of 43.57 minutes compared to 11.25 minutes for COLMAP. In addition, the average reprojection error of COLMAP in the SFM stage is 1.1441, which is significantly better than that of OpenMVG+OpenMVS of 3.9760, indicating its obvious advantage in reconstruction accuracy. Meanwhile, the average track length of COLMAP is 4.87487, which is much higher than that of OpenMVG+OpenMVS(2.000). The experimental results show that COLMAP has excellent advantages in capturing more view information, which is conducive to generating 3D models with high integrity.

In the MVS stage, the number of dense point clouds generated by OpenMVS reaches 14,514,039, which is 6.2 times that generated by COLMAP. Although the number is enormous, it also leads to the generation of much noise, which increases the subsequent computation time and memory requirements. In contrast, CMVS-PMVS generated the least number of dense point clouds with 1519,825 points. Although the processing efficiency is improved, the model details still need to be included. The 3D model obtained by COLMAP+CMVS-PMVS is rough and cannot meet the requirements of high-precision modelling. Although COLMAP performs well in point cloud generation alone, the final generated 3D model performs mediocre in visualization due to insufficient texture computation.

In contrast, OpenMVS generated 3D models performed the best and were able to more accurately reflect the structure of real objects. COLMAP+OpenMVS method has the highest accuracy, completeness, SSIM value and topological consistency, indicating that the reconstructed model generated by this method has the highest similarity with the reference image, and retains more image details and structural information. Secondly, although the accuracy, completeness, SSIM value and topological consistency of the OpenMVG+OpenMVS method are slightly lower than those of the COLMAP+OpenMVS method, the OpenMVG+OpenMVS method still shows high similarity, indicating that the reconstructed model generated by this method has better performance in quality. The topological consistency values of the COLMAP method are relatively close to those of the previous two methods. The completeness, accuracy, SSIM value and topological consistency value of COLMAP+CMVS-PMVS method are the lowest, which indicates that the reconstructed model generated by this method has the lowest similarity with the reference image, and the structure preservation effect is the worst. Compared with OpenMVG+OpenMVS, the combination of COLMAP and OpenMVS improves the reconstruction accuracy by 2% and the completeness by 2.6%.

In summary, COLMAP+OpenMVS can obtain a dense and fine 3D point cloud in 88.32 minutes, maintain a low reprojection error and a long average track length, and generate a high-precision, detailed, and realistic 3D model.

C. SUMMARY AND FORESIGHT

Based on the existing 3D reconstruction technology, this paper proposes an improved reconstruction method by combining the application of UAV and mobile phones for data acquisition and using COLMAP and OpenMVG algorithms for image processing. This approach not only optimizes the flexibility of data acquisition but also significantly improves efficiency and accuracy through an automated reconstruction process. On resource-constrained devices, such as smartphones or low-power drones, which often struggle to handle complex data processing tasks, this study shows its superiority. In addition, through multi-sensor fusion and deep learning-based stereo vision methods, the proposed technique can effectively reconstruct scenes with poor texture or highly repetitive scenes, alleviating the limitations of traditional methods in these scenes. Future work will continue to study improving the efficiency of the algorithm, enhancing the generalization ability and robustness, to adapt to more kinds of practical application scenarios, especially those applications with high requirements for real-time processing and high-precision models. At the same time, the multi-view stereo vision method based on deep learning has achieved good results in 3D reconstruction. Multi-sensor fusion and deep learning methods will carry out subsequent research work based on this paper, trying to solve the practical application problem of a 3D reconstruction algorithm based on multi-view photos in real scenes.

Nonetheless, current approaches still face challenges of scalability and computational complexity when dealing with large-scale or highly dynamic scenarios. Especially on resource-constrained devices, efficient data processing and algorithm optimization are particularly critical. In addition, scenes with poor texture or high repetition are still difficult to handle and may lead to poor reconstruction quality. Future research is needed to explore new sensor technologies or new ways to mix existing technologies while ensuring the efficiency and accuracy of the algorithm. In addition, the generalization ability and robustness of the algorithm are also the focus of future research, expecting to achieve accurate and reliable 3D reconstruction in different environments and on different devices.

REFERENCES

[1] Y. Hou, J. Song, and L. Wang, "P-2.27: Application of 3D reconstruction technology in VR industry," in *SID Symp. Dig. Tech. Papers*, vol. 54, 2023, pp. 588–590, doi: 10.1002/sdtp.16227.

[2] N. Tanabi, A. M. Silva, M. A. O. Pessoa, and M. S. G. Tsuzuki, "Robust algorithm software for NACA 4-Digit airfoil shape optimization using the adjoint method," *Appl. Sci.*, vol. 13, no. 7, p. 4269, Mar. 2023, doi: 10.3390/app13074269.

[3] D. Li, B. Zhang, Q. Wu, Q. She, and C. Zhong, "A lightweight semantic map construction method for robot target search," in *Proc. 34th China Conf. Control Decision Making*, 2022, pp. 73–77.

[4] C. P. Bachiller and E. Upegui, "Low cost 3D reconstruction of cave paintings for the conservation of Colombian historical memory: Case study indigenous rock art of the sacred place 'Piedras de Tunjo,'" *Int. Arch. Photogramm., Remote Sens. Spatial Inf. Sci.*, vol. XLVIII-2/W1-2022, pp. 251–256, Dec. 2022.

[5] L. Liu, H. Cai, M. Tian, D. Liu, Y. Cheng, and W. Yin, "Research on 3D reconstruction technology based on laser measurement," *J. Brazilian Soc. Mech. Sci. Eng.*, vol. 45, no. 6, 2023, Art. no. 297, doi: 10.1007/s40430-023-04231-9.

[6] C.-S. Yang, Q. Zhang, Q. Xu, C.-Y. Zhao, J.-B. Peng, and L.-Y. Ji, "Complex deformation monitoring over the Linfen–Yuncheng basin (China) with time series InSAR technology," *Remote Sens.*, vol. 8, no. 4, p. 284, Mar. 2016, doi: 10.3390/rs8040284.

[7] X. Liu, P. Ren, X. Sun, C. Xu, and Q. Zhou, "Review of 3D digitization methods for ancient Chinese architecture," *J. Shanxi Univ., Natural Sci. Ed.*, no. 3, pp. 592–603, 2023, doi: 10.13451/j.sxu.ns.2022112.

[8] B. Yu, G. Chen, M. Duan, F. Cao, and X. Zhang, "Application of UAV remote sensing in 3D reconstruction of large immovable cultural relics," *Surveying Mapping Bull.*, no. 5, pp. 43–46 and 61, 2017, doi: 10.13474/j.cnki.11-2246.2017.0151.

[9] S. Jiang, "Study on key technologies for efficient SfM reconstruction from UAV oblique imagery," Ph.D. dissertation, Dept. Remote Sens. Inf. Eng., Wuhan Univ., Wuhan, China, 2018.

[10] N. Tang, J. Yu, H. Xu, Z. Liang, and P. Rui, "Comparison of point cloud model reconstruction techniques based on image matching," *Eng. Invest.*, vol. 49, no. 6, pp. 62–67, 2021.

[11] S. Wang, H. Yu, Y. Xi, C. Gong, W. Wu, and F. Liu, "Spectral-image decomposition with energy-fusion sensing for spectral CT reconstruction," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1–11, 2021, doi: 10.1109/TIM.2021.3078555.

[12] Y. Wang, J. Ren, A. Cai, S. Wang, N. Liang, L. Li, and B. Yan, "Hybrid-domain integrative transformer iterative network for spectral CT imaging," *IEEE Trans. Med. Imag.*, vol. 73, 2024, Art. no. 4504513, doi: 10.1109/TIM.2024.3379388.

[13] Q.-V. Dang and G.-S. Lee, "Utilizing 3D information from point clouds to support document image binarization," *EasyChair*, 2022. Accessed: Mar. 18, 2024. [Online]. Available: https://easychair.org/publications/preprint/9141

[14] E. H. H. Siong, M. F. M. Ariff, and A. F. Razali, "The application of smartphone based structure from motion (SFM) photogrammetry in ground volume measurement," *Int. Arch. Photogramm., Remote Sens. Spatial Inf. Sci.*, vol. XLVIII-4, pp. 145–152, Feb. 2023.

[15] R. Zhang, X. Yi, H. Li, L. Liu, G. Lu, Y. Chen, and X. Guo, "Multiresolution patch-based dense reconstruction integrating multiview images and laser point cloud," *Int. Arch. Photogramm., Remote Sens. Spatial Inf. Sci.*, vol. XLIII-B2-2022, pp. 153–159, 2022, doi: 10.5194/isprs-archives-XLIII-B2-2022-153-2022.

[16] J. Liao, "Research on high-precision 3D reconstruction of complex scenes based on multi-view photography," Ph.D. dissertation, Dept. Remote Sens. Inf. Eng., Wuhan Univ., Wuhan, China, 2021.

[17] H. Liu, "Research on 3D reconstruction technology using feature pyramid networks," M.S. thesis, Dept. Inf. Sci. Eng., Yunnan Univ., Kunming, China, 2022.

[18] X. Meng, "3D modeling of buildings with feature preservation based on enhanced 3D point cloud models," Ph.D. dissertation, Dept. Geogr. Sci., Nanjing Normal Univ., Nanjing, China, 2021.

[19] Q. Kong, L. He, L. Yuan, and B. Liu, "Improved PMVS three-dimensional reconstruction point cloud filtering algorithm," *Comput. Appl. Softw.*, no. 4, pp. 215–219 and 270, 2021.

[20] X. Sun, "Research on surface reconstruction algorithms for large-scale scene point clouds," M.S. thesis, Dept. Inf. Sci. Eng., Northeastern Univ., Shenyang, China, 2019.

[21] J. Lang, "Research on variable-scale 3D modeling technology based on point clouds," M.S. thesis, Dept. Educ. Sci., Nanjing Normal Univ., Nanjing, China, 2021.

[22] A. Xu, Y. Hua, C. Xia, and S. Chen, "An improved feature point matching algorithm based on SIFT," *Software*, vol. 43, no. 9, pp. 83–86 and 119, 2022.

[23] X. Liu, X. Zhao, Z. Xia, Q. Feng, P. Yu, and J. Weng, "Secure outsourced SIFT: Accurate and efficient privacy-preserving image SIFT feature extraction," *IEEE Trans. Image Process.*, pp. 4635–4648, 2023, doi: 10.1109/TIP.2023.3295741.

[24] M. Wang and W. Liu, "Stereo image matching algorithm based on improved SIFT features," *Comput. Eng. Appl.*, vol. 49, pp. 203–206, 2013.

[25] B. Alsadik and N. A. Abdulateef, "Epipolar geometry between photogrammetry and computer vision—A computational guide," *ISPRS Ann. Photogramm., Remote Sens. Spatial Inf. Sci.*, vol. V-5-2022, pp. 25–32, May 2022.

[26] Y. He and J. Yue, "Research and implementation of CMVS/PMVS multi-view dense matching method," *Surveying Mapping Geograph. Inf.*, vol. 38, pp. 20–23, 2013, doi: 10.14188/j.2095-6045.2013.03.004.

[27] V. Sebestyén, M. Bulla, Á. Rédey, and J. Abonyi, "Data-driven multilayer complex networks of sustainable development goals," *Data Brief*, vol. 25, Aug. 2019, Art. no. 104049, doi: 10.1016/j.dib.2019.104049.

[28] X. Chen, "Research on 3D reconstruction algorithms based on multiple views," M.S. thesis, Dept. Inf., North China Univ. Technol., Beijing, China, 2024.

[29] R. Song, "Research on large complex surface reconstruction methods based on point cloud data," M.S. thesis, Dept. Mech. Elect. Eng., Univ. Electron. Sci. Technol., Chengdu, China, 2020.

[30] Y. Chengming, "Parallel construction algorithm of Delaunay triangulation network," Ph.D. dissertation, Xinyang Normal Univ., Xinyang, China, 2022.

[31] M. Cao, S. Li, W. Jia, and X. Liu, "A review of feature tracking methods in structure from motion technology," *J. Comput. Res. Develop.*, vol. 41, pp. 2536–2565, 2018.

[32] Z. Liang, Y. Yu, W. Hou, and X. Xu, "Crane main girder deformation identification method based on UAV 3D reconstruction," *Hoisting Conveying Mach.*, no. 17, pp. 25–30, 2023.

[33] L. Jia, C. Wang, and H. Wang, "Comparison and analysis of UAV oblique photography 3D modelling software," *Geomatics Spatial Geo-Inf.*, vol. 47, no. 5, pp. 176–179, 2024.

[34] H. Xu, W. Guo, D. Li, J. Hao, and G. Xu, "Automatic identification and decryption method of sensitive targets in real scene 3D model textures," *Bull. Surveyingv Mapping*, no. 12, pp. 153–158, 2023, doi: 10.13474/j.cnki.11-2246.2023.0376.

[35] H. Zhang, Y. Ding, and D. Li, "Structural surface crack detection method based on 3D reconstruction," *Ind. Construct.*, vol. 54, no. 5, pp. 60–67, 2024, doi: 10.13204/j.gyjzG22102611.

[36] F. Hu, "Research on 3D reconstruction method of cable-stayed bridge structure perception based on deep learning," Ph.D. dissertation, Harbin Inst. Technol., Harbin, China, 2022.

[37] S. Wang, H. Xue, Y. Zhang, and Z. Yao, "Research on key frame extraction and 3D reconstruction methods for endoscopic video," *J. Chongqing Technol. Bus. Univ., Natural Sci. Ed.*, pp. 1–11, Mar. 2024.

HAOTIAN MING is currently pursuing the master's degree with Guangxi University of Science and Technology. His main research interests include remote sensing mapping and image recognition.

QI LI is currently a Senior Engineer with Guangxi University of Science and Technology. His main research interests include remote sensing mapping, renewable energy, unmanned aerial vehicle, solar energy evaluation, and prediction.

HAIBO XIA is currently pursuing the master's degree with Guangxi University of Science and Technology. His main research interests include remote sensing mapping and image recognition.

PENG LI is currently an Associate Professor with the Department of Public Health, Faculty of Medicine, Guangxi University of Science and Technology. Her main research interests include solar energy, regional climate models, atmospheric environmental change, urban disaster prevention and reduction, and public health.