Multiplatform Bundle Adjustment Method Supported by Object Structural Information

Jianchen Liu and Wei Guo

Abstract—**The registration and integration of data from different platforms are becoming more and more important for real-scene three-dimensional (3-D) reconstruction. In urban areas, the integration of unmanned aerial vehicle images and terrestrial images can compensate for geometric distortions and texture blurring in models generated from single-platform images. However, it remains a question of how to maintain a high accuracy while accounting for the discrepancies between various platforms. Bundle adjustment is a crucial step in building a detailed 3-D model. However, traditional bundle adjustment is usually applied to a single platform. In the case of multiplatform data with significant differences in resolution, flight height, or viewing angles, it can lead to the issues of instability and low accuracy in solving bundle adjustment problems. This article innovatively proposes a multiplatform bundle adjustment method, which is supported by object structural information. First, the method performs patch-based matching of images from different platforms and obtains cross-platform tie points. Second, refined patches obtain object structural information by calculating the depth values and ground sampling distances of image points. Finally, multiplatform bundle adjustment is conducted using weights calculated for both object and image points based on factors obtained in the second step. The experimental results show that, in general, the proposed method can achieve the accuracy level required for practical applications. Compared with the bundle adjustment method without object structural information, the improvement of accuracy is significant, with an average improvement of 53.38% across the four datasets.**

Index Terms—**Bundle adjustment (BA), image matching, multiplatform, object structural information.**

I. Introduction

IN RECENT years, the construction of digital cities has increased in pace. The three-dimensional (3-D) visualization, air pollution analysis, disaster analysis, and scene simulation are some fields where it finds extensive application. However, the production process of photorealistic 3-D models typically requires significant manual intervention, although it serves as the foundation of digital urban spatial data [1]. Oblique photogrammetry based on airborne oblique imagery systems and unmanned aerial vehicle (UAV) have made great progress in recent years, the 3-D mesh model can be automatically generated from the images obtained by oblique photography, and this method is gradually replacing the traditional pipeline based on orthophoto and digital elevation model (DEM) [2]. Fig. 1(a) illustrates that oblique photogrammetry based on UAV often results in incomplete information acquisition in urban areas due to occlusion, narrow flight zones, and large camera tilt angles, causing holes, blurs, or distortions in the 3-D model [1]. Terrestrial images can provide a complementary perspective and offer complete building profile information; however, the top of the building is lost in the resulting 3-D model [see the example in Fig. 1(b)]. Therefore, the integration of different platforms helps to enrich the detailed information of 3-D models and realize the realistic 3-D modeling of urban areas.

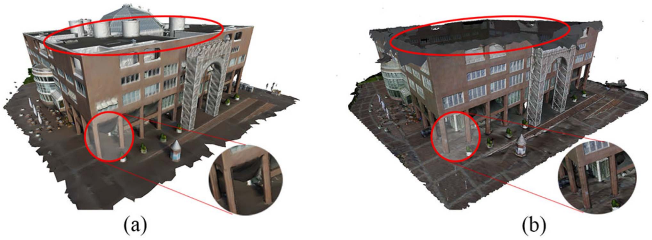


Fig. 1. Three-dimensional mesh models generated from different platforms. (a) Aerial images. (b) Terrestrial images.

However, it is difficult to build a 3-D model using photos taken from multiple platforms for the following reasons: First, since different platforms capture an object from different angles, traditional methods of image matching cannot accommodate the drastic perspective distortion. As a result, an insufficient number of tie points can be used for cross-platform bundle adjustment (BA) [3], [4], Second, the traditional BA is suitable usually for single-platform images with quadrilateral footprints and similar scales. It ignores the object structural information of the tie point, such as the orientation of the local surface and the distance to the camera, which can significantly affect the stability and accuracy of BA. In addition, it assumes the same precision for all observations and applies identical weights to different parts of the observations from different directions [5]. For multiple platforms, the traditional BA hypothesis is invalid. Therefore, how to solve the impact of differences in different platforms on the BA process is the key to solving the problem.

In this article, a multiplatform combined BA method is proposed based on the object structure information, which contains two main parts: the position and direction of local planes. Its significant contribution lies in the transformation of the abstract

Manuscript received 28 August 2023; revised 10 November 2023; accepted 30 November 2023. Date of publication 5 December 2023; date of current version 14 December 2023. This work was supported in part by the National Natural Science Foundation of China under Grant 42171439, and in part by the Qingdao Science and Technology Demonstration and Guidance Project under Grant 22-3-7-cspz-1-nsh. *(Corresponding author: Wei Guo.)*

The authors are with the College of Geodesy and Geomatics, Shandong University of Science and Technology, Qingdao 266590, China (e-mail: liujianchen @sdust.edu.cn; 17685542660@163.com).

Digital Object Identifier 10.1109/JSTARS.2023.3339293

concept of differences between multiple platforms into concrete computable object structural information. This transformation facilitates their controlled incorporation within the BA process. The proposed method consists of the following two main elements: First, patch-based matching: create patches based on the sparse point clouds acquired from aerial images and terrestrial images separately and optimize the patches by variational patch refinement to make them closer to the real surface of objects, and then perform patch-based matching to obtain cross-platform tie points; second, reweighting based on object structure information: the object structure information is calculated based on the result of variational patch refinement. The ground sampling distance (GSD) of the pixel can be calculated by the orientation of the refined patches, and the depth value can be calculated by the position of the refined patches. The above factors are combined to reweight the image points. This proposed process categorizes all tie points into two groups: internal platform tie points (Its visible image points belong to the image acquired from the same platform) and cross-platform tie points (Its visible image points belong to images acquired from different platforms). Additionally, cross-platform tie points establish connections across each platform to improve consistency. Although they account for only a small proportion of all tie points, they are of crucial importance. Diverse weights are assigned based on the tie point degree and the ratio to total tie points.

The rest of this article is organized as follows. Section II describes the existing methods of multiplatform combined data processing briefly. Section III describes the proposed method and its key steps in detail. In Section IV, the performance of the proposed method is evaluated using different sets of aerial and terrestrial images. Finally, Section V concludes this article.

II. Related Work

Early research has primarily focused on multiplatform data processing through the merging of ground-based and airborne laser scanning information to enable several applications. These applications include point cloud data registration [6], [7], the evaluation of sea cliff changes, utilizing both terrestrial and airborne LiDAR technology [8], object detection, and 3-D urban modeling [9], [10]. In addition, UAV-based photogrammetry and terrestrial laser scanning (TLS) have been combined by various researchers, leading to applications, including the seamless mapping of river channels [11], the generation of DEM [12], the 3-D mapping and monitoring of open-pit mine areas [13], and the environmental management [14]. The development of photogrammetry and computer vision has significantly improved the automatic 3-D reconstruction pipelines based on images. Currently, some researchers are exploring the integration of aerial and terrestrial imagery. The integration of aerial photogrammetry and terrestrial photogrammetry has resulted in high-resolution 3-D models, enabling effective applications in cultural heritage protection [15], high fidelity urban models reconstruction [16], and urban environment studies [17]. It is important to note that reliable and freely available datasets are critical for testing and comparing new algorithms and procedures [18].

The research articles on the integration of aerial and terrestrial photogrammetry mainly include two aspects: tie point matching between aerial and terrestrial images and combined BA for multiplatforms.

A. Image Matching

In recent years, many research articles have made great progress in the aerial-to-terrestrial image matching; there are mainly two kinds of methods: direct image matching and rectified image matching.

Direct image matching: This kind of method tries to find the tie point between these two platforms directly. Since Lowe [19] proposed the groundbreaking invention of the scale-invariant feature transform (SIFT) algorithm, many researchers have improved various aspects of SIFT based on it, such as dimensionality [20], calculation speed [21], affine invariance [22], and perspective invariance [23]. However, the SIFT-like features are not essentially affine invariant, and the performance degrades significantly when the translation tilt exceeds 25° [24]. In the case of aerial-to-terrestrial image matching, a large part of the tilt between image pairs is more than 25°, even more than 60°, then point match outliers appear. Lin et al. [25] proposed an outlier filtering method based on matching consistency that can be applied in the cases mentioned above. While the authors employ a hypothetical point matching training bilateral function to detect point matching outliers, it does not mitigate the critical challenge of linking aerial and terrestrial images: perspective distortion. Great progress has also been made in deep-learning-based methods, such as DFM [26]. DFM performs well for scenes in which the target object is nearly planar, but performance drops when the surface of the object is uneven. SuperPoint [27] and SuperGlue [28] have also made great progress in multiview image matching. SuperPoint outperforms the traditional algorithms in feature extraction, but it is generally not possible to make the network more rotationally invariant and light invariant due to the tradeoff between invariance and discrimination [29]. SuperGlue has been improved based on SuperPoint, which can greatly improve the performance of feature matching.

Rectified image matching: This kind of method aims to transform an image acquired from one viewpoint into another in order to eliminate any perspective differences. In [30], the terrestrial images are warped to the perspective of aerial by depth-based warping, then the rectified images and the aerial images are matched by SIFT. In [1], aerial and terrestrial images are projected to the base planes, which are based on building facades, and then the rectified images with similar perspective and scale characteristics are matched. The two methods mentioned above eliminate perspective distortions, which can be challenging to use for matching, by rectifying the image from one platform to another. However, this rectification method has the disadvantage of not considering the actual shape and orientation of the object surface, which can lead to potential errors in matching. The problem is addressed in this article by proposing a patch-based matching method that utilizes sparse point clouds generated from aerial and terrestrial imagery to establish patches. These patches are then optimized based on the local object surfaces

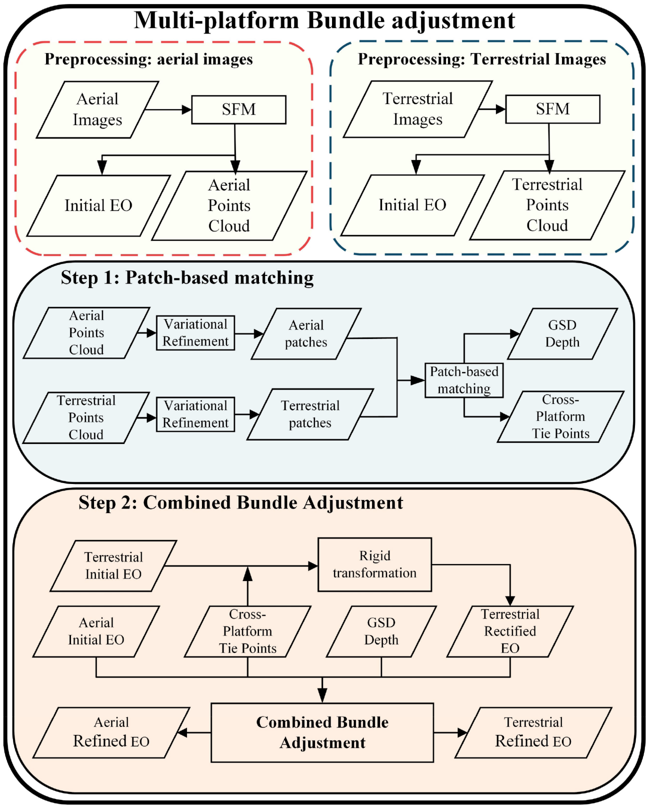


Fig. 2. Overall workflow of the proposed approach.

using variational refinement. By projecting the points onto these patches, the perspective distortion and scale differences are reduced.

B. Bundle Adjustment

However, although great progress has been made in the aerial-to-terrestrial image matching, the precision and accuracy of the matching still do not meet the requirements in the practical application. Fortunately, the BA optimizes the camera's position and orientation, as well as the 3-D points obtained during image matching [31]. This results in the elimination of reprojection errors caused by mismatching during the BA stage. However, the conventional BA assigns equal and symmetrical weight to all observations based on the assumption that their importance is identical. In practice, however, this is not the case. The uncertainty of the automatically matched points is between 0.1 and 0.5 pixels, whereas the uncertainty of the manually measured image points is assumed to be 0.5 pixels [32]. As shown in [33], for onboard GPS/IMU measurements, the positioning error is between 5 and 10 cm at best, and the uncertainty of orientation for roll and pitch is typically about 0.005° and about 0.008° for heading. In fact, for the case of different significance of observations, researchers have proposed a lot of ways to measure their priority or their contribution to the block. Gerke [34] suggests assigning a higher weight to the constraints between points (horizontal and vertical) than other observations. Moreover, Xie et al. [5] assigned higher weights to observations with higher resolution and penalized those with lower resolution according to the GSD and the height of platform to the ground at oblique photogrammetry. However, the above method assumes a flat ground surface and attributes GSD solely to the platform's tilt angle and height, making it inapplicable to regions with complex structures, such as intricate buildings. In multiple platforms, the importance of varying observations differs significantly, which is further complicated by the difficulty of establishing a positional relationship between different photogrammetric platforms due to their differing heights and positions. Environmental and terrain factors, such as image scales and lighting conditions, can significantly impact the accuracy of orientation and photogrammetric products [35].

One simple and direct approach to address this issue is to reassign observation weights based on the data characteristics of different platforms during the BA process. The proposed method performs reweighting based on two aspects: image-point observations and cross-platform tie points. The method penalizes the observation values of image points with large perspective angle deviations and the cross-platform tie points with low connectivity degrees, effectively solving the problem of differences in image data caused by different platforms, viewpoints, and resolutions, thus improving the accuracy and stability of the computation.

III. Methodology

A. Overview of Proposed Approach

The framework of the multiplatform combined BA supported by the object structural information method is shown in Fig. 2. To obtain the initial elements of exterior orientation (EO) parameters for both the aerial and terrestrial platforms, the structure from motion (SFM) pipeline is run separately for each platform using available auxiliary data, such as global navigation satellite system (GNSS) and ground control points (GCPs). This process generates sparse point clouds for each platform. Sparse point clouds are utilized to construct patches, which are refined using the proposed variational patch refinement method. These refined patches are then used for patch-based matching, allowing for the acquisition of cross-platform tie points. The sparse point cloud of the terrestrial platform is aligned with the aerial one via cross-platform tie points to obtain a set of rigid transformation parameters. This set of transform parameters is applied to the initial EO parameters of the terrestrial images to obtain the rectified EO parameters; this step eliminates the discrepancy of the EO parameters between the two platforms. Then, the GSD and depth values of the image-point projection onto the real surface of the building are calculated according to the generated patches. Finally, the initial EO parameters of the aerial images and the rectified EO parameters of the terrestrial images are used as initial values for the combined BA. A reweighting based on the cross-platform tie points and the object structural information is performed during the process.

This article rectifies the terrestrial images for the following reasons: First, due to the limitations of the terrestrial platform perspective, the GCPs are often not all visible on the terrestrial images, and the GNSS signals of the terrestrial platform may

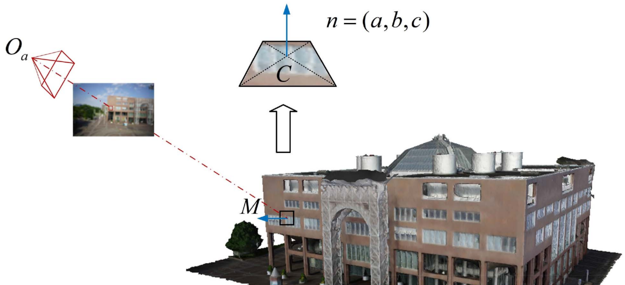


Fig. 3. Graphical representation of the patch.

be obscured, resulting in poor accuracy of the SFM results of the terrestrial images alone. Second, the aerial images usually have a wider perspective, and the GNSS signal remains relatively stable, and the two platforms do not share a unified coordinate system after SFM alone.

B. Variational Patch Refinement Method

The patch is constructed from the coordinates of the point cloud, and it consists of two parts: the normal vector $n = (a, b, c)$ and the centroid C (as shown in Fig. 3). Since the normal vector n can be equivalently replaced by two azimuths (α, β), a patch can be expressed as (α, β, d), where d is the fourth element of the planar standard equation.

Assuming that images are represented by I_i and I_j, patches are represented by P. In the variational patch refinement process, the patch is projected onto I_i and I_j. The degree of similarity is measured by the normalized cross correlation (NCC), and a larger NCC means that the patch is closer to the real surface of the object. The proposed method constructs an energy function to iteratively calculate the position and orientation of the best patch such that the NCC of the patch projection on the two images is maximized. The energy function is constructed as follows:

$$E_{i,j}(P) = \sum_{m_i \in I_i \cap I_j^{P,i}} h\left(I_i, I_j^{P,i}\right)(m_i) \tag{1}$$

where $I_j^{P,i}$ denotes the reprojection of image I_i on image I_j through the patch P. h denotes the decreasing function of the photoconsistency measure of images I_i and I_j at the image point m_i of I_i, specifically: $h(I_i, I_j) = 1 - \text{NCC}$. Also, $E_{i,j}(P)$ describes the discrepancy between the patch projection region and the image pair.

To minimize $E_{i,j}(P)$, this method solves for the partial derivatives of $E_{i,j}(P)$ with respect to the three parameters (α, β, d) of patch P, which can be expressed as the following equation:

$$\frac{dE_{i,j}(P)}{dP(\alpha, \beta, d)} = \frac{dh\left(I_i, I_j^{P,i}\right)(m_i)}{dI_j^{P,i}(m_i)} \cdot \frac{dI_j(m_j)}{dm_j} \cdot \frac{dH'}{dP(\alpha, \beta, d)} \tag{2}$$

where m_i and m_j denote the coordinate points of image I_i and image I_j, respectively, and H' denotes the normalization of the homographic matrix H.

The derivative of the first photoconsistency can be computed as a gradient by assigning the derivative to the pixel intensity $I_j^{P,i}(m_i)$ of image $I_j^{P,i}$. The derivative of the pixel intensity with respect to the image point m_i coordinates can be expressed as the pixel gradient of the corresponding image point m_j in image I_j, i.e., $dI_j(m_j)/dm_j$. Assuming that the projection matrix for image I_j is denoted as M_j, X_j represents the object point corresponding to image points m_i and m_j, with T being the transformation matrix that maps image points to the object point on the patch. This matrix comprises variables α, β, and d. The following equations can be derived:

$$\begin{aligned} H \cdot \tilde{m}_i &= M_j \cdot X_j \\ X_j &= T \cdot \tilde{m}_j. \end{aligned} \tag{3}$$

The correspondence between H and T can be obtained by the above equation, where H can be calculated from the projection matrix of the two images so that only T contains the unknowns related to the patches. In the above equation, $\tilde{m}_i$ and $\tilde{m}_j$ represent the homogeneous coordinates of the image points m_i and m_j.

In the above iterative process of solving for the optimal positions and orientations of the patches, the initial values of the positions and orientations of the patches have a significant impact on the time consumption of the iterations. The initial parameters of the patch are computed from the points cloud and the position of the images, where the centroid C of the patch is the coordinates of each point cloud, and the normal vector n is parallel to the line connecting the average of the optical centers of all reference images to the centroid of the patch.

C. Patch-Based Tie Points Matching Method

The process of image matching provides the initial value of observations for BA procedure, so the precision of BA is directly related to the accuracy of image matching. There are many factors that affect the accuracy of image matching, such as illumination differences, perspective distortion, and scale variations. Among them, perspective distortion and scale variations are the most common problems among multiplatform data. As shown in Fig. 4, O_1, O_2, and O_3 represent the different projection centers, and their observations of the same object point are indicated by arrows. It is evident that the GSD of the same matched image pair may vary greatly when projected onto different platforms. When captured by the ground platform, the GSD of the pixel is approximately equal in the x and y directions, and the shape of the pixel footprint is close to a square. When captured by the aerial platform, the pixel footprint has a large difference in the x and y directions, resulting in a trapezoidal shape. In this article, the object structure information is obtained by the result of variational patch refinement, specifically, by the center of refined patch and the normal vector to calculate the depth value and GSD.

Since each platform is at a different distance from the building, this results in a variation in the proportion of the building

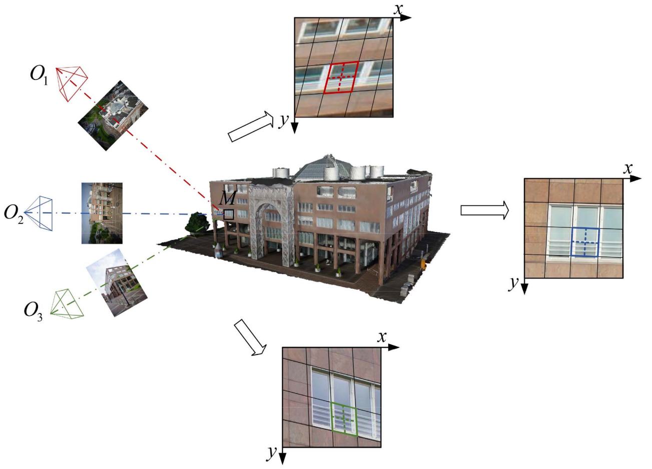


Fig. 4. Pixel footprint of the same matched image pair on different perspectives.

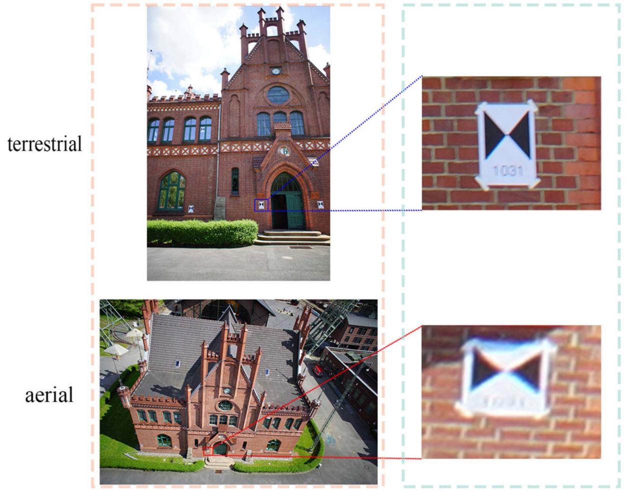


Fig. 5. Difference in resolution of the same landmark on different platforms.

on each platform's image. This phenomenon is visualized in the resolution, as shown in Fig. 5. The terrestrial platform is generally closer to the building, so the landmarks on the image have a higher resolution and are less prone to mismatching. The aerial platform is generally farther away from the building, so the landmarks that correspond to the ground platform have a lower resolution on the image and are prone to mismatching.

By variational patch refinement, the perspective distortion and scale difference caused by the drastic change of image perspective are eliminated (as shown in Fig. 6). The refined patch center is taken as the feature point for patch-based matching using SIFT descriptors. Initial matching was performed using fast library for approximate nearest neighbors (FLANNs), followed by the elimination of outliers using bidirectional matching and saliency detection using nearest neighbor distance ratio (NNDR).

D. Reweighting Based on Object Structure Information

The object structural information in this article consists of two parts: the GSD and the image-point depth value, both of which can be computed based on the optimized patch. Among them, the depth value is calculated based on the distance between the projection center and the centroid of the patch, and the calculation principle of GSD is shown as follows.

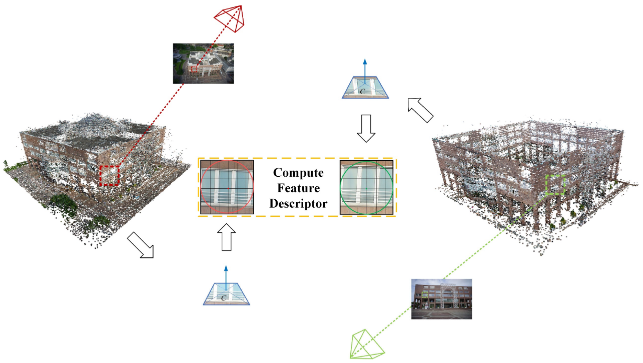


Fig. 6. Patch-based tie points matching.

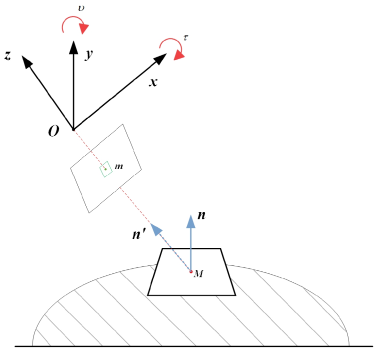


Fig. 7. Calculation of the GSD for pixels.

Fig. 7 shows that the square pixel m projected on patch M has GSDs that can be represented by (4) in both x and y directions. Here, n is the normal vector of the refined patch

$$\begin{aligned}\mathrm{GSD}_x &= \frac{D}{f\cos\tau}\\ \mathrm{GSD}_y &= \frac{D}{f\cos\upsilon}\end{aligned} \tag{4}$$

where D is the distance from projection center O to patch, and f is the focal length (mm). The orientation of refined patch n is obtained from the initial value of n' after rotating τ angles around the x-axis and υ angles around the y-axis

$$\begin{aligned}P_y &= \frac{\mathrm{dep}_0}{\mathrm{dep}}\cdot\frac{\mathrm{GSD}_x}{\mathrm{GSD}_y}\\ P_x &= \frac{\mathrm{dep}_0}{\mathrm{dep}}\cdot\frac{\mathrm{GSD}_y}{\mathrm{GSD}_x}.\end{aligned} \tag{5}$$

Formula (5) is the weighting part of the image point in which P_x and P_y represent the weights of the image points in the horizontal and vertical directions, separately. The dep_0 represents the average depth of the image, where the image point is located, and the dep represents the depth value of the image

TABLE I
Primary Parameters of the Experimental Dataset Used in This Article

Datasets	Verwaltung		Lohnhalle		Center Hall		Hospital	
Platforms	Aerial	Terrestrial	Aerial	Terrestrial	Aerial	Terrestrial	Aerial	Terrestrial
Sensor	DJI_S800(Sony Nex-7)	Sony Nex-7	DJI_S800(Sony Nex-7)	Sony Nex-7	DJI_S800(Sony Nex-7)	Sony Nex-7	DJI_Phantom 4(FC6310S)	Sony a7
Image resolution (pixel)	6000 × 4000	6000 × 4000	6000 × 4000	6000 × 4000	6000 × 4000	6000 × 4000	5472 × 3648	6000 × 4000
Number of images	166	97	229	169	178	172	361	949
Number of GCP/CP	4/30		3/30		4/36		6/30	
Number of cross-platform tie points	88 620		57 883		60 348		74 614	

point

$$P_{\text{objPt}} = G \cdot \min\left(\frac{C_1}{C_2}, \frac{C_2}{C_1}\right) \cdot \frac{N_0}{N}. \quad (6)$$

Formula (6) is the weighting part of the object points in which P_{objPt} denotes the weight of the cross-platform tie point. The N_0 denotes the total number of tie points and N denotes the number of cross-platform tie points. The C_1 and C_2 denote the number of images of the two platforms where the tie point is visible, and G denotes the degree of cross-platform tie point ($G = C_1 + C_2$).

In the method proposed in this article, the reweighting of the two parts, image point and object point, is realized on the basis of the normal equation. This is shown in the following equation:

$$\begin{bmatrix} \dot{B}_{ij}^T P_{ij} \dot{B}_{ij} & \dot{B}_{ij}^T P_{ij} \ddot{B}_{ij} \\ \ddot{B}_{ij}^T P_{ij} \dot{B}_{ij} & \ddot{B}_{ij}^T P_{ij} \ddot{B}_{ij} + P_j \end{bmatrix} \begin{bmatrix} \dot{\delta}_i \\ \ddot{\delta}_j \end{bmatrix} = \begin{bmatrix} \dot{B}_{ij}^T P_{ij} l_{ij} \\ \ddot{B}_{ij}^T P_{ij} l_{ij} \end{bmatrix} \quad (7)$$

$$P_{ij} = \begin{bmatrix} P_x & \\ & P_y \end{bmatrix} \quad (8)$$

where P_{ij} is the 2∗2 weight matrix about the image point, and the weighting of the image point proposed in this method is achieved by modifying the element values of the weight matrix. $B = [\ \dot{B} \quad \ddot{B}\ ]$ denotes the partial derivatives of image EO parameters and coordinates of object point, respectively. The subscripts *i* and *j*, respectively, denote the id of the image and the object point. *l* indicates the observations, and $[\ \dot{\delta} \quad \ddot{\delta}\ ]^T$ denotes the correction values of the six image EO parameters and the three coordinate parameters of object point, respectively.

In this experiment, the reweighting of the object points is adding the diagonal weight matrix of the same dimensions to the lower right matrix block, as shown in (7) for P_j (P_{object}).

IV. Experimental Results and Analysis

A. Description of Experimental Data

To verify the performance of the proposed method, systematic experiments and analyses are performed in this article using UAV images and terrestrial images provided by International Society for Photogrammetry and Remote Sensing (ISPRS) and EuroSDR [18] and the university hospital data of the Shandong University of Science and Technology (see Fig. 8). The processing platform for this experiment is equipped with an Intel 10th i5 CPU, a NVIDIA 3050 GPU, and 16 GB of RAM. The primary parameters of the experimental dataset are shown in Table I.

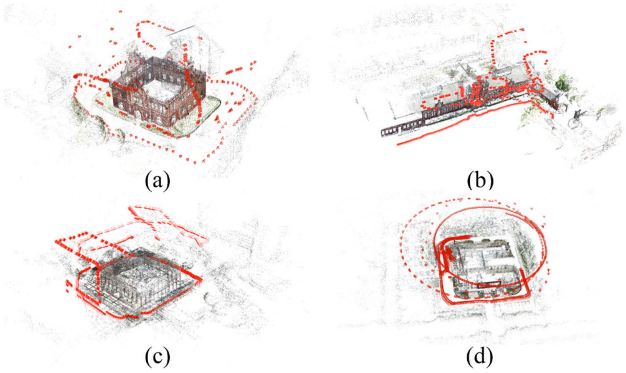


Fig. 8. Aerial and terrestrial images reconstructed using Colmap [36]. All datasets: (a) Verwaltung; (b) Lohnhalle; (c) center hall; and (d) hospital.

The classic BA method mentioned in this article refers to the equal weight BA implemented using Visual Studio in the Windows environment, which does not consider the structural information of the object.

B. Result and Analysis of Patch-Based Matching

The patch-based matching method proposed in this article is compared with traditional algorithms through experimental trials. In order to showcase the superiority of the proposed algorithm, the article will juxtapose it with the affine-SIFT [22] (ASIFT) algorithm and GMS [37] algorithm. In this article, the FLANN method is used to match descriptors, and the outliers are eliminated by saliency detection using NNDR, with the ratio set to 0.6. This article additionally introduces a deep-learning method SuperGlue [28] to compare with the proposed method. Five pairs of images, characterized by substantial perspective variations, have been chosen. The matching outcomes are depicted in Fig. 9, while Table II presents the count of matched points and the count of inlier points. The accuracy of the matching results is checked by reprojection error.

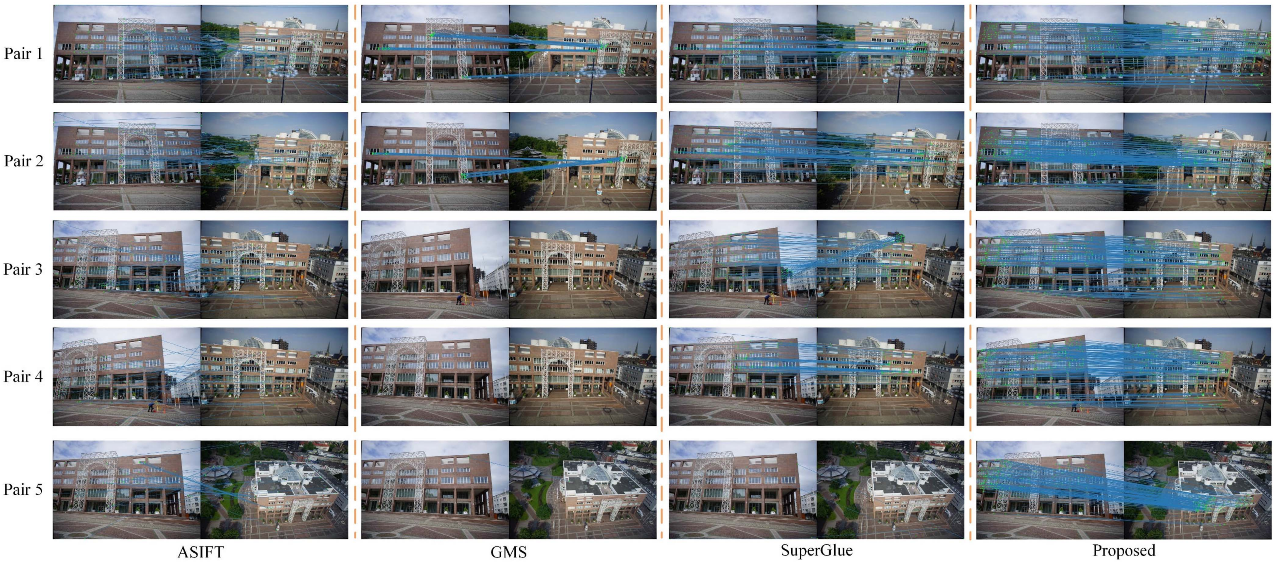


Fig. 9. Comparison of two matching methods for five image pairs.

TABLE II
COMPARISON OF MATCHING RESULTS

Method Inlier/all	ASIFT	GMS	SuperGlue	Proposed method
Pair 1	59/82	80/96	165/408	329/338
Pair 2	58/78	64/76	165/317	279/321
Pair 3	8/69	0/53	109/120	235/287
Pair 4	1/38	0/21	122/129	245/259
Pair 5	0/33	0/16	0/4	135/157

C. Result and Analysis of Rigid Transformation

When GNSS information is unavailable, differences between the two platforms can be eliminated through spatial similarity transformations. After patch-based matching, the cross-platform tie points can be obtained. Since the feature points extracted by the SFM process are the same as those of the patch-based matching, the correspondence of points between two platform sparse point clouds can be obtained. It is assumed that the point P_1 in the first cluster point cloud and the point P_2 in the second cluster point cloud have a correspondence and can be rigidly transformed from P_1 to P_2 by the following equation:

$$P_2 = \lambda R P_1 + T \tag{9}$$

where λ is the global scale factor of the two platforms, R is the global rotation matrix, and T is the 3-D translation vector. It is easy to know that (9) contains seven rigid transformation parameters (three translation parameters, three rotation parameters, and one scale factor), which need at least three pairs of points between the sparse point clouds to be solved. The solved rigid transformation parameters can be used to align the EO parameters of the two platforms by the following equation:

$$\begin{aligned} R' &= R R_0 \\ T' &= \lambda R T_0 + T \end{aligned} \tag{10}$$

where (R_0, T_0) and (R', T') denote the original and transformed rotational and positional EO parameters of the images, respectively.

Two control points are visible in the terrestrial images of the Verwaltung dataset, while all four control points are visible in the aerial images. Aligning the EO parameters of the terrestrial platform with those of the aerial platform using rigid transformation can significantly enhance the accuracy of the terrestrial EO parameters. Fig. 10(a) illustrates that the point clouds acquired from different platforms exhibit minimal differences after this step, almost overlapping convincingly. Fig. 10(b) displays the point cloud of Lohnhalle dataset obtained following the rectification of the terrestrial platform to the aerial platform, revealing a significant reduction in discrepancies between different platforms after the rigid transformation. The dataset of the hospital exhibits a high degree of accuracy in its initial EO parameters for both platforms, leading to close point clouds generated by each platform. The center hall dataset is unique as only the images of the terrestrial platform contain visible markers. Images captured via the UAV platform lack any visible marker on the building surface. To eliminate the discrepancies between the two platforms, a transformation is necessary from the aerial platform to the terrestrial platform. Due to the unavailability of checkpoint data from the aerial platform, it is not possible to compare the accuracy before and after transformation in this dataset. In conclusion, Fig. 10 illustrates the visual effect of the rigid transformation, demonstrating a significant reduction in differences between the two platforms after transformation.

Fig. 11 illustrates the Euclidean distance of the point cloud before and after the rigid transformation of one point cloud to the other, obtained by computing the average of all the points with correspondences.

The limitations of SFM alone, such as perspective restriction and obstruction of GNSS signal, contribute to the poor accuracy of the terrestrial images. Table III presents the residuals of checkpoints before and after the transformation of terrestrial images.

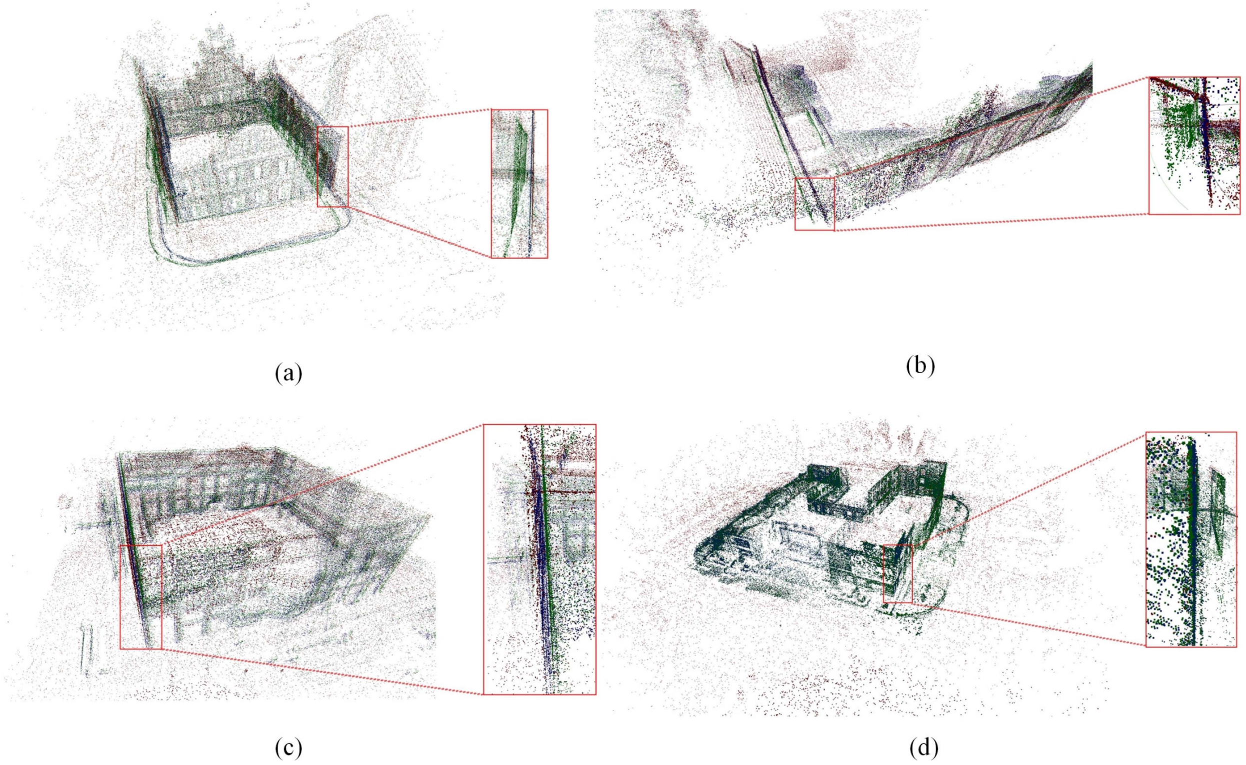


Fig. 10. Schematic of point cloud rigid transformation results. Green: Terrestrial; Red: Aerial; Blue: Transformed from one to the other. All datasets: (a) Verwaltung; (b) Lohnhalle; (c) center hall; and (d) hospital.

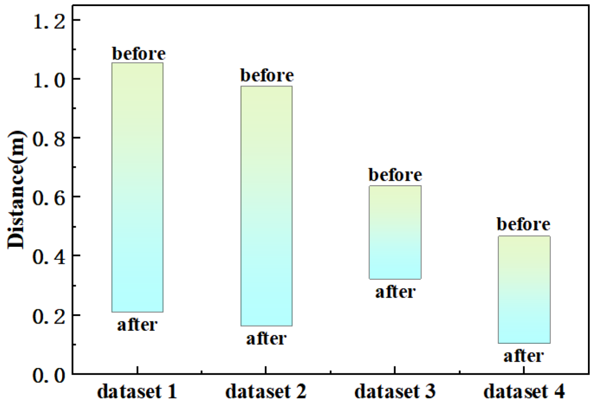


Fig. 11. Euclidean distance before and after point cloud transformation. Before: Platform1- platform2; After: Platform1- transformed2.

On the two public datasets from ISPRS (center hall dataset with no checkpoints in the aerial imagery) and the School Hospital dataset, the rigidly transformed ground platforms showed various levels of improvement in accuracy compared with the pretransformation ones.

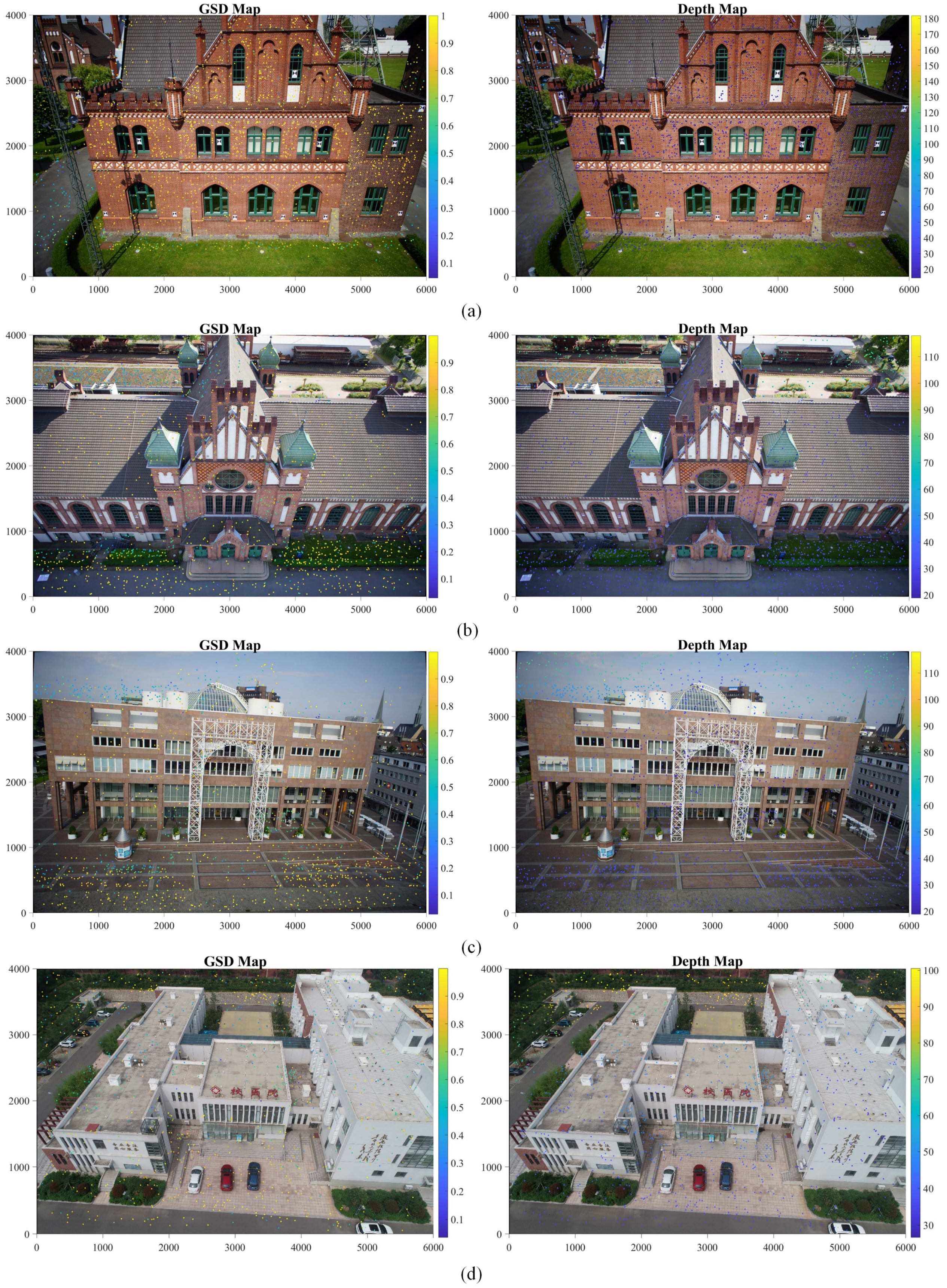
D. Result and Analysis of GSD and Depth Value

Fig. 12 displays the GSD and depth values of each image point obtained through the patch refinement method. The left of Fig. 12(a) reveals that the GSDs of the image points on the building facade are roughly equal in both x and y directions, while the GSDs of the image points on the ground display more disparity in both directions. The right of Fig. 12(a) depicts a larger depth value for the image points near the bottom of the building, consistent with the actual scenario.

TABLE III
RESIDUAL STATISTICS OF CHECKPOINTS BEFORE AND AFTER RIGID TRANSFORMATION PROCESSING OF TERRESTRIAL IMAGES

Dataset	RMSE(m)	X	Y	Z	Total
Verwaltung	Before transformation	0.144	0.051	0.250	0.293
	After transformation	0.066	0.048	0.188	0.204
	Improvement	54.17%	5.88%	24.80%	30.38%
Lohnhalle	Before transformation	0.080	0.105	0.150	0.199
	After transformation	0.047	0.061	0.053	0.094
	Improvement	41.25%	41.90%	64.67%	52.76%
Hospital	Before transformation	0.185	0.095	0.162	0.264
	After transformation	0.137	0.082	0.149	0.218
	Improvement	25.95%	13.68%	8.02%	17.42%

Fig. 12(b) demonstrates the GSD and depth values on one of the images in the Lohnhalle dataset. As depicted in the left of Fig. 12(b), the GSD of the image point projected onto the roof exhibits similarity close to 1 in both x and y directions, while that of the image point on the building side is less than 1 since the



(a)

(b)

(c)

(d)

Fig. 12. GSD and depth values of all image points on one image obtained using patch refinement. All datasets: (a) Verwaltung; (b) Lohnhalle; (c) center hall; and (d) hospital.

TABLE IV
ACCURACY RESULT OF COMBINED BA OF ALL DATASETS

Dataset	RMSE(m)	X	Y	Z	Total
Verwaltung	Classical BA	0.106	0.124	0.027	0.166
	Proposed	0.011	0.052	0.026	0.059
	Improvement	89.62%	58.07%	3.70%	64.24%
Lohnhalle	Classical BA	0.107	0.191	0.098	0.240
	Proposed	0.051	0.071	0.056	0.104
	Improvement	52.34%	62.83%	42.86%	56.67%
Center Hall	Classical BA	0.053	0.059	0.015	0.081
	Proposed	0.026	0.016	0.015	0.034
	Improvement	50.94%	72.88%	0%	58.03%
Hospital	Classical BA	0.261	0.132	0.186	0.347
	Proposed	0.187	0.039	0.122	0.227
	Improvement	28.35%	70.45%	34.41%	34.58%

camera viewpoint is near to the frontal view of the roof during photography. In the right of Fig. 12(b), corresponding depth values of the image points projected onto both the roof and the side are observed to be approximately 20 m, consistent with the actual situation. Those depths for the image points projected on the distant ground are approximately 60 m.

Fig. 12(c) presents the GSD and depth values of every image point in the center hall dataset based on the patch refinement method. As seen on the left, the GSD ratio of image points projected onto the facade of the building is almost 1. In contrast, the GSD ratio of the image points projected on the ground is less than 1 and decreases as the distance from the camera increases. On the right, after excluding the impact of some depth values, image points projected onto the building facade are shown to have similar depth values. Conversely, the depth values of image points projected onto the ground are proportional to the distance from the camera, accurately reflecting the real-world scenario.

Fig. 12(d) represents the GSD and depth values of each image point on an image obtained from the hospital dataset of the Shandong University of Science and Technology. In the left graph of Fig. 12(d), the GSD of the image point projected on the ground in front of the main entrance is close to 1 in the ratio of x and y directions, while the GSD of the image point on the roof of the building in the far distance has a ratio of less than 1. In the right graph of Fig. 12(d), the depth value of the image point gradually increases with the distance from the camera, which is in line with the real situation.

E. Results and Analysis of Combined BA

Table IV demonstrates the residual errors of various methods. On the three publicly available datasets from ISPRS, the proposed method improves 64.24%, 56.67%, and 58.03%, respectively, compared with the classical BA method. The average improvement is 64.30%, 64.59%, and 23.28% in the x, y, and z directions, respectively. On the dataset of hospital of Shandong University of Science and Technology, the proposed method improves 34.58% with respect to the classical BA method. It improves 28.35%, 70.45%, and 34.41% in the x, y, and z directions, respectively. The above data show that the overall accuracy of the proposed method can still be maintained at a high level even when other platforms are added for combined BA.

V. Discussion

This article innovatively introduces a multiplatform combined BA method based on object structural information, effectively quantifying the influence factor of weight in the BA process for local surface information of target objects. The main contribution of this article is the transformation of the abstract concept of differences between multiple platforms into concrete and computable object structural information. This transformation aids in controlling them during the BA process.

The patch-based matching method proposed in this article is based on the variational patch refinement method to create feature descriptors, providing sufficient cross-platform tie points for multiplatform combined BA and calculating specific numerical values for object structural information.

The method proposed in this article still has the following limitations. Regarding the patch-based matching process: First, compared with the traditional matching methods, patch-based matching methods require the construction of patches, which necessitates obtaining a sparse point cloud first through traditional reconstruction pipelines, such as SFM, adding computational time to the algorithm. Second, while this method can effectively handle perspective distortions between image pairs, significant differences in image scale can considerably impact its performance due to resolution differences. For the multiplatform combined BA process: In the context of 3-D reconstruction in urban areas, this method has a clear advantage over the traditional method. However, in regions, where the GSD and depth value changes are not significant, the improvements achieved by this method may not be as pronounced.

VI. Conclusion

In order to improve the accuracy and stability of combined multiplatform data processing, this article proposed a multiplatform combined BA method based on object structural information. The method determines the weights of each image point based on the GSDs (in both x and y directions) and depth value, and the weights of each object point based on the degree of cross-platform tie points and the ratio to all tie points. Four blocks of images provided by the ISPRS base dataset and the dataset of hospital in the Shandong University of Science and Technology are used for experimental validation and comparative analysis. The results demonstrate that the accuracy of the proposed method is better than that of the classical BA method in practical applications. In the experiments on all datasets, the overall average improvement in accuracy is 53.38%.

The significance of the proposed method is that it combines aerial oblique images with terrestrial images for BA, which solves the problems of unstable settlement and poor accuracy that easily occur in traditional methods when facing different platforms' data. Moreover, the terrestrial platform is aligned to the aerial platform through the common tie point matched by the patch-based matching. It solves the problem of large difference of initial EO parameters between different platforms and the problem of tracking the loss of GNSS of terrestrial platform in urban environment. By combining BA between different platforms, the accuracy of integrated 3-D modeling is improved, which provides an effective solution for high-fidelity and high-precision integrated modeling in urban environment.

References

[1] B. Wu, E. Xie, H. Hu, Q. Zhu, and E. You, “Integration of aerial oblique imagery and terrestrial imagery for optimized 3D modeling in urban areas,” *ISPRS J. Photogramm. Remote Sens.*, vol. 139, pp. 119–132, 2018.

[2] F. Remondino and M. Gerke, “Oblique aerial imagery: A review,” in *Photogrammetric Week*. Stuttgart, Germany: Wichmann/VDE Verlag, 2015, pp. 75–83.

[3] M. Gerke, F. Nex, and P. Jende, “Co-registration of terrestrial and UAV-based images—Experimental results,” *Int. Arch. Photogramm., Remote Sens. Spatial Inf. Sci.*, vol. 40, pp. 11–18, 2016.

[4] Q. Shan, C. Wu, B. Curless, Y. Furukawa, C. Hernandez, and S. M. Seitz, “Accurate geo-registration by ground-to-aerial image matching,” in *Proc. 2nd Int. Conf. 3D Vis.*, Tokyo, Japan, 2014, pp. 525–532.

[5] L. Xie, H. Hu, J. Wang, Q. Zhu, and M. Chen, “An asymmetric re-weighting method for the precision combined bundle adjustment of aerial oblique images,” *ISPRS J. Photogramm. Remote Sens.*, vol. 117, pp. 92–107, 2016.

[6] L. Cheng, L. Tong, M. Li, and Y. Liu, “Semi-automatic registration of airborne and terrestrial laser scanning data using building corner matching with boundaries as reliability check,” *Remote Sens.*, vol. 5, no. 12, pp. 6260–6283, 2013.

[7] L. Cheng, L. Tong, Y. Wu, Y. Chen, and M. Li, “Shiftable leading point method for high accuracy registration of airborne and terrestrial LiDAR data,” *Remote Sens.*, vol. 7, no. 2, pp. 1915–1936, 2015.

[8] A. P. Young et al., “Comparison of airborne and terrestrial lidar estimates of seacliff erosion in Southern California,” *Photogramm. Eng. Remote Sens.*, vol. 76, pp. 421–427, 2010.

[9] J. Boehm and N. Haala, “Efficient integration of aerial and terrestrial laser data for virtual city modeling using LASERMAPs,” in *Proc. Int. Arch. Photogramm., Remote Sens. Spatial Inf. Sci.-Arch.*, 2005, pp. 192–197.

[10] M. Kedzierski and A. Fryskowska, “Terrestrial and aerial laser scanning data integration using wavelet analysis for the purpose of 3D building modeling,” *Sensors*, vol. 14, no. 7, pp. 12070–12092, Jul. 2014.

[11] C. Flener et al., “Seamless mapping of river channels at high resolution using mobile LiDAR and UAV-photography,” *Remote Sens.*, vol. 5, no. 12, pp. 6382–6407, 2013.

[12] M. M. Ouédraogo, A. Degre, C. Debouche, and J. Lisein, “The evaluation of unmanned aerial system-based photogrammetry and terrestrial laser scanning to generate DEMs of agricultural watersheds,” *Geomorphology*, vol. 214, pp. 339–355, 2014.

[13] X. Tong et al., “Integration of UAV-based photogrammetry and terrestrial laser scanning for the three-dimensional mapping and monitoring of open-pit mine areas,” *Remote Sens.*, vol. 7, no. 6, pp. 6635–6662, 2015.

[14] S. W. Son, D. W. Kim, W. G. Sung, and J. J. Yu, “Integrating UAV and TLS approaches for environmental management: A case study of a waste stockpile area,” *Remote Sens.*, vol. 12, no. 10, 2020, Art. no. 1615.

[15] C. Balletti, F. Guerra, V. Scocca, and C. Gottardi, “3D integrated methodologies for the documentation and the virtual reconstruction of an archaeological site,” *Int. Arch. Photogramm., Remote Sens. Spatial Inf. Sci.*, vol. 40, pp. 215–222, 2015.

[16] W. Wahbeh, G. Muller, M. Ammann, and S. Nebiker, “Automatic image-based 3D reconstruction strategies for high-fidelity urban models—Comparison and fusion of UAV and mobile mapping imagery for urban design studies,” *Int. Arch. Photogramm., Remote Sens. Spatial Inf. Sci.*, vol. 43, pp. 461–468, 2022.

[17] M. Rumpler et al., “Evaluations on multi-scale camera networks for precise and geo-accurate reconstructions from aerial and terrestrial images with user guidance,” *Comput. Vis. Image Understanding*, vol. 157, pp. 255–273, 2017.

[18] F. Nex, M. Gerke, F. Remondino, H.-J. Przybilla, M. Baumker, and A. Zurhorst, “ISPRS benchmark for multi-platform photogrammetry,” *ISPRS Ann. Photogramm., Remote Sens. Spatial Inf. Sci.*, vol. 2, pp. 135–142, 2015.

[19] D. G. Lowe, “Distinctive image features from scale-invariant keypoints,” *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, 2004.

[20] K. Mikolajczyk et al., “A comparison of affine region detectors,” *Int. J. Comput. Vis.*, vol. 65, pp. 43–72, 2005.

[21] H. Bay, A. Ess, T. Tuytelaars, and L. van Gool, “Speeded-up robust features (SURF),” *Comput. Vis. Image Understanding*, vol. 110, no. 3, pp. 346–359, 2008.

[22] J.-M. Morel and G. Yu, “ASIFT: A new framework for fully affine invariant image comparison,” *Soc. Ind. Appl. Math. J. Imag. Sci.*, vol. 2, no. 2, pp. 438–469, 2009.

[23] G.-R. Cai, P.-M. Jodoin, S.-Z. Li, Y.-D. Wu, S.-Z. Su, and Z.-K. Huang, “Perspective-SIFT: An efficient tool for low-altitude remote sensing image registration,” *Signal Process.*, vol. 93, no. 11, pp. 3088–3110, 2013.

[24] X. Gao, S. Shen, Z. Hu, and Z. Wang, “Ground and aerial meta-data integration for localization and reconstruction: A review,” *Pattern Recognit. Lett.*, vol. 127, pp. 202–214, 2019.

[25] W.-Y. D. Lin, M.-M. Cheng, J. Lu, H. Yang, M. N. Do, and P. Torr, “Bilateral functions for global motion modeling,” in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 341–356.

[26] U. Efe, K. G. Ince, and A. A. Alatan, “DFM: A performance baseline for deep feature matching,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops*, 2021, pp. 4279–4288.

[27] D. DeTone, T. Malisiewicz, and A. Rabinovich, “SuperPoint: Self-supervised interest point detection and description,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops*, 2018, pp. 337–33712.

[28] P.-E. Sarlin, D. DeTone, T. Malisiewicz, and A. Rabinovich, “SuperGlue: Learning feature matching with graph neural networks,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 4937–4946.

[29] R. Pautrat, V. Larsson, M. R. Oswald, and M. Pollefeys, “Online invariance selection for local feature descriptors,” in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 707–724.

[30] X. Gao, L. Hu, H. Cui, S. Shen, and Z. Hu, “Accurate and efficient ground-to-aerial model alignment,” *Pattern Recognit.*, vol. 76, pp. 288–302, 2018.

[31] J. L. Schonberger and J.-M. Frahm, “Structure-from-motion revisited,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Las Vegas, NV, USA, 2016, pp. 4104–4113.

[32] A. Gruen, “Development and status of image matching in photogrammetry,” *Photogramm. Rec.*, vol. 27, no. 137, pp. 36–57, 2012.

[33] A. Ip, N. El-Sheimy, and M. Mostafa, “Performance analysis of integrated sensor orientation,” *Photogramm. Eng. Remote Sens.*, vol. 73, no. 1, pp. 89–97, 2007.

[34] M. Gerke, “Using horizontal and vertical building structure to constrain indirect sensor orientation,” *ISPRS J. Photogramm. Remote Sens.*, vol. 66, no. 3, pp. 307–316, 2011.

[35] E. M. Farella, A. Torresani, and F. Remondino, “Refining the joint 3D processing of terrestrial and UAV images using quality measures,” *Remote Sens.*, vol. 12, no. 18, 2020, Art. no. 2873.

[36] J. L. Schönberger, E. Zhang, J.-M. Frahm, and M. Pollefeys, “Pixelwise view selection for unstructured multi-view stereo,” in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 501–518.

[37] J. Bian, W.-Y. Lin, Y. Matsushita, S.-K. Yeung, T.-D. Nguyen, and M.-M. Cheng, “GMS: Grid-based motion statistics for fast, ultra-robust feature correspondence,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 2828–2837.

Jianchen Liu was born in Jiamusi, China, in 1987. He received the M.S. degree in geomatics engineering from the Shandong University of Science and Technology, Qingdao, China, in 2013, and the Ph.D. degree in photogrammetry and remote sensing from the School of Remote Sensing and Information Engineering, Wuhan University, Wuhan, China, in 2017.

He is currently an Associate Professor with the College of Geodesy and Geomatics, Shandong University of Science and Technology. His research interests include unmanned aerial vehicle photogrammetry, computer stereovision, and 3-D modeling by oblique images.

Wei Guo was born in Heze, China, in 1999. He received the B.S. degree in geomatics engineering in 2021 from Shandong University of Science and Technology, Qingdao, China, where he is currently working toward the master’s degree in digital photogrammetry.