remote sensing

MDPI

Article

Robust Fusion of Multi-Source Images for Accurate 3D Reconstruction of Complex Urban Scenes

Yubin Liang *, Yang Yang, Yifan Mu and Tiejun Cui

School of Geographic and Environmental Sciences, Tianjin Normal University, Tianjin 300387, China; 2110080035@stu.tjnu.edu.cn (Y.Y.); 2210080039@stu.tjnu.edu.cn (Y.M.); cuitiejun@tjnu.edu.cn (T.C.)
* Correspondence: ybliang@tjnu.edu.cn

Abstract: Integrated reconstruction is crucial for 3D modeling urban scenes using multi-source images. However, large viewpoint and illumination variations pose challenges to existing solutions. A novel approach for accurate 3D reconstruction of complex urban scenes based on robust fusion of multi-source images is proposed. Firstly, georeferenced sparse models are reconstructed from the terrestrial and aerial images using GNSS-aided incremental SfM, respectively. Then, cross-platform match pairs are selected based on point-on-image observability. The terrestrial and aerial images are robustly matched based on the selected match pairs to generate cross-platform tie points. Thirdly, the tie points are triangulated to derive cross-platform 3D correspondences. The 3D correspondences are refined using a novel outlier detection method. Finally, the terrestrial and aerial sparse models are merged based on the refined correspondences, and the integrated model is globally optimized to obtain an accurate reconstruction of the scene. The proposed methodology is evaluated on five benchmark datasets, and extensive experiments are performed. The proposed pipeline is compared with a state-of-the-art methodology and three widely used software packages. Experimental results demonstrate that the proposed methodology outperforms the other pipelines in terms of robustness and accuracy.

Keywords: 3D reconstruction; terrestrial–aerial integration; structure from motion; cross-platform image matching; outlier detection

Citation: Liang, Y.; Yang, Y.; Mu, Y.; Cui, T. Robust Fusion of Multi-Source Images for Accurate 3D Reconstruction of Complex Urban Scenes. *Remote Sens.* **2023**, *15*, 5302. https://doi.org/10.3390/rs15225302

Academic Editors: Domenico Visintini and Filiberto Chiabrando

Received: 8 October 2023
Revised: 7 November 2023
Accepted: 7 November 2023
Published: 9 November 2023

1. Introduction

3D reconstruction of urban scenes provides a fundamental data source for many smart city researches and applications [1–3]. In recent years, airborne oblique photogrammetry has become one of the mainstream solutions for reconstructing photorealistic 3D models in urban scenes due to its high cost-effectiveness, high fidelity, and high accessibility of professional equipment [4–6]. Although airborne oblique photogrammetry is widely adopted for 3D modeling at the city scale, the bottom parts of reconstructed models are often unsatisfactory due to the occlusion of ground objects and large perspective distortion of aerial imagery, especially in complex urban scenarios. With the development of data acquisition techniques, the integration of aerial and terrestrial imagery is widely used for generating better 3D models in terms of completeness, accuracy, and fidelity [7–9].

One major challenge of terrestrial–aerial integrated 3D reconstruction is cross-platform image matching [10,11]. Establishing tie-points between images with large viewpoint and illumination variations is difficult for SIFT-like image matching methods [12]. Recent learning-based image matching methods can extract more distinctive features by using deep neural networks [13,14]. These learned features exhibit better performance on benchmark datasets. Although a few methodologies based on handcrafted and learning-based image matching algorithms have been proposed to improve the robustness of cross-platform image matching [15,16], the problem has not been fully resolved.

Another challenge is the accurate fusion of terrestrial and aerial models. Most studies reconstructed terrestrial and aerial models and merged the models via a 3D similarity

transformation. The similarity transformation is usually estimated using 3D correspondences derived from cross-platform tie points. However, previous studies have shown that the epipolar constraint cannot remove all mismatches. And the remaining cross-platform mismatches will introduce inaccurate observations to the estimation of similarity transformation and the global optimization of the integrated model. Although outlier detection methods have been proposed to filter remaining mismatches and incorrect 3D correspondences, the accuracy of model fusion can be further improved. Terrestrial and aerial models can also be merged based on point cloud registration [17,18]. However, these methods require accurate extraction and robust matching of common geometric features in two point clouds. The establishment of accurate correspondences in cross-platform point clouds is still challenging.

This paper presents a novel approach for integrated 3D reconstruction of urban scenes using aerial and terrestrial imagery. The main contributions of this paper are as follows. First, a robust image matching method is proposed to tackle the cross-platform image matching problem. Incremental Structure from Motion (SfM) with weighted Global Navigation Satellite System (GNSS) observations is used to reconstruct georeferenced terrestrial and aerial sparse models, respectively. Based on the sparse models, cross-platform match pairs are selected by projecting terrestrial points to the aerial images. Instead of matching rectified images or renderings, terrestrial and aerial images are directly matched based on the selected match pairs to generate cross-platform tie points. Second, an outlier detection algorithm is proposed to refine 3D correspondences between terrestrial and aerial models. The proposed algorithm is derived from the positioning uncertainty of photogrammetric reconstruction. The cross-platform tie points are robustly triangulated based on the terrestrial and aerial sparse models. Then, outliers are removed from the correspondences based on the statistics of positional differences. The similarity transformation from the terrestrial sparse model to the aerial sparse model is estimated based on the refined 3D correspondences. After merging the transformed terrestrial model with the aerial model, the integrated model is globally optimized.

The remainder of this paper is organized as follows. Section 2 reviews related research works. Section 3 elaborates on the proposed methodology, including reconstruction of terrestrial and aerial sparse models, robust matching of terrestrial and aerial images, and accurate fusion of terrestrial and aerial sparse models. Section 4 presents experimental results on five benchmark datasets, and the performance of the proposed methodology is demonstrated by comparative experiments and ablation studies. Section 5 discusses the experimental results and limitations of the proposed methodology. Finally, conclusions are made in Section 6.

2. Related Works

The core components of an image-based 3D reconstruction pipeline include image matching, image orientation, dense matching, and textured mesh construction. A traditional image matching procedure extracts feature points from images and finds initial matches between image pairs [19,20]. After mismatches are filtered based on the epipolar constraint with the random sample consensus (RANSAC) framework, tie points are established [21,22]. Then, the image orientation procedure estimates the optimal extrinsic parameters, intrinsic parameters, camera calibration parameters, and the sparse structure of a scene based on the tie points. The Structure from Motion (SfM) framework is the de facto standard for fully automatic image orientation [23–26]. The dense matching procedure establishes pixel-wise correspondences between matched images and generates dense depth maps, which can be used to derive the dense point cloud of the scene [27–29]. Based on the dense point cloud, the textured mesh construction procedure first builds the geometric model of the scene using a 3D triangular mesh and textures the mesh with images [30].

Robust cross-platform image matching is generally required when applying the above pipeline to terrestrial–aerial integrated reconstruction in complex urban scenarios. It is well known that SIFT is sensitive to viewpoint changes larger than 50 degrees [12]. ASIFT

improves the robustness of image matching by simulating all possible affine distortions and matches the simulated images using SIFT [31]. An image matching approach based on warping aerial images to ground was proposed for matching nadir and oblique aerial images [32]. The method reduced the viewpoint difference between the aerial and oblique images, and the warped images were robustly matched using SIFT. A similar approach was proposed for matching terrestrial and aerial images toward the reconstruction of ancient Chinese architecture [15]. This approach first conducted terrestrial and aerial sparse reconstruction, respectively. Then, terrestrial–aerial image pairs were selected based on co-visible mesh, and terrestrial images were warped to the perspectives of aerial images. The warped images were matched against the aerial images using SIFT. And the tie points were filtered and transferred to the original terrestrial images. Similarly, a rendering-based approach was proposed for cross-platform image matching [9]. This method detected building facades from the dense cloud derived from the aerial images and rectified each pair of images based on the detected facades. The method effectively increased the number of SIFT tie points between terrestrial and aerial images. A similar strategy was proposed to match terrestrial and aerial images based on rectifying images using textured mesh [9]. This method rendered the textured mesh derived from the aerial images to the perspectives of the terrestrial images and matched the renderings with the terrestrial images using SIFT. The method exhibited high robustness for cross-platform image matching on five benchmark datasets. An approach based on refined image patches was proposed to match cross-platform images [33]. Sparse point clouds were derived from aerial and terrestrial images, respectively. Image patches were built based on the point clouds and optimized to be close to the tangent plane of the object surface by variational patch refinement. The aerial and terrestrial image patches were matched using SIFT. Although these methods improved the robustness of matching cross-platform images, the dependence on the SIFT-like algorithms limits their capability in challenging urban scenarios.

In recent years, many learning-based methods have been proposed for robust image matching under challenging conditions [13,14,34]. These methods can extract more distinctive features using convolutional neural networks trained on benchmark datasets. Based on the Transformer framework, tie points can be established without using a feature detector [35]. These learning-based methods showed better adaptiveness to viewpoint and illumination variations than handcrafted methods on benchmark datasets [36]. These learning-based methods have been used for matching cross-platform images. A learning-based framework was proposed for matching terrestrial and aerial images [37]. A dense correspondence network was trained to learn consistent features among terrestrial and aerial images and generate dense correspondences. Then, sparse keypoints were extracted from each image, and tie points were established between each image pair based on the dense correspondences. The locations of the keypoints were further refined using the learned feature map to improve the quality of the tie points. A methodology based on the SuperGlue algorithm [38] was proposed for matching aerial, mobile mapping, and backpack images [8]. The method first generated a sparse point cloud from the aerial images. Then, the sparse point cloud was segmented, and facade planes were extracted from the segmented point cloud. Images acquired by different platforms were rectified onto the extracted facade planes, and the rectified images were matched using the SuperGlue algorithm. The method performed well on a challenging dataset. However, the point cloud segmentation results require manual checking and interactive improvements. Moreover, the SuperGlue algorithm extracts much fewer feature points from images with poor textures. The unevenly distributed tie points require manual adjustments. Although learning-based image matching methods have shown promising performance, recent studies have demonstrated that these methods do not have obvious advantages over handcrafted ones in conventional 3D reconstruction tasks [39–42]. Directly applying these learning-based methods to match cross-platform images and reconstruct 3D models of complex urban scenes remains challenging.

Accurate fusion of terrestrial and aerial models is also required for high-quality terrestrial–aerial integrated reconstruction. The estimation of accurate similarity transformation between terrestrial and aerial models requires precise 3D correspondences. These 3D correspondences are usually obtained by triangulating the cross-platform tie points. It is well known that outlier detection based on the epipolar constraint cannot eliminate all mismatches. Methods have been proposed to further remove the remaining mismatches. Mismatches were filtered by using thresholds on the variations of scale and principal orientation of SIFT features [15]. The affine transformation model with RANSAC loops was also used for filtering mismatches. Then, terrestrial–aerial tracks were triangulated to obtain 3D correspondences. A global bundle adjustment was performed to merge terrestrial and aerial point clouds, in which the Huber loss was introduced to deal with false 3D correspondences. An outlier detection method based on geometric constraints was proposed to filter mismatches [9]. The length, intersection, and direction constraints defined based on disparity vectors were used to remove outliers from initial matches. The remaining mismatches were filtered using the epipolar constraint. The established tie points were further refined by matching local patches in the original terrestrial and aerial images. A normalized correlation coefficient search was used to find initial matches, and the initial matches with a correlation score smaller than a threshold were pruned. A two-stage approach was proposed for outlier detection [33]. Outliers in initial matches were firstly filtered by cross-checking and saliency detection using the nearest neighbor distance ratio test. Then, a 3D similarity transformation between two sets of image patches was computed with the RANSAC framework to further remove outliers. The 3D similarity transformation was also used as an additional geometric constraint to limit the matching range of the image patches, which also improved the robustness of the approach. Geometric constraints were also proposed to filter mismatches in [8]. After mismatches were filtered using the epipolar constraint, a 3D point was calculated from each match pair. A match was considered an outlier if the corresponding 3D point was far from the facade plane or other 3D points calculated from matches on other images. After outlier detection, tie points were linked to build tracks. And tie points with short track lengths were further removed. The advantage of these outlier detection methods is that the constraints have clear geometric meanings and are easy to understand. However, setting threshold values for these constraints requires practical experience, which can be challenging for complex datasets.

In summary, recent studies have shown promising performances of learning-based image matching methods on benchmark datasets. However, the capabilities of these methods have not been effectively incorporated into the reconstruction pipeline. Furthermore, most outlier detection methods filter mismatches from the perspective of image matching. The positioning uncertainty of the terrestrial–aerial integrated reconstruction problem has not been fully exploited. Innovative methods need to be developed to robustly match cross-platform images and achieve accurate integrated reconstruction.

3. Methodology

3.1. Overview of Proposed Methodology

The workflow of the proposed methodology is illustrated in Figure 1. Firstly, image matching is performed separately on the terrestrial and aerial images of a scene. Georeferenced sparse models are reconstructed from the terrestrial and aerial images, respectively. Then, match pair selection is performed to determine match pairs between cross-platform images. Based on the selected match pairs, robust image matching is conducted to generate tie points that connect terrestrial and aerial images. Then, the cross-platform tie points are triangulated to derive 3D points from the terrestrial and aerial sparse models, respectively. Correspondences between the terrestrial and aerial 3D points are determined, and outliers are filtered. The terrestrial and aerial sparse models are merged based on the refined correspondences, and the integrated model is globally optimized to finally obtain an accurate reconstruction of the scene.

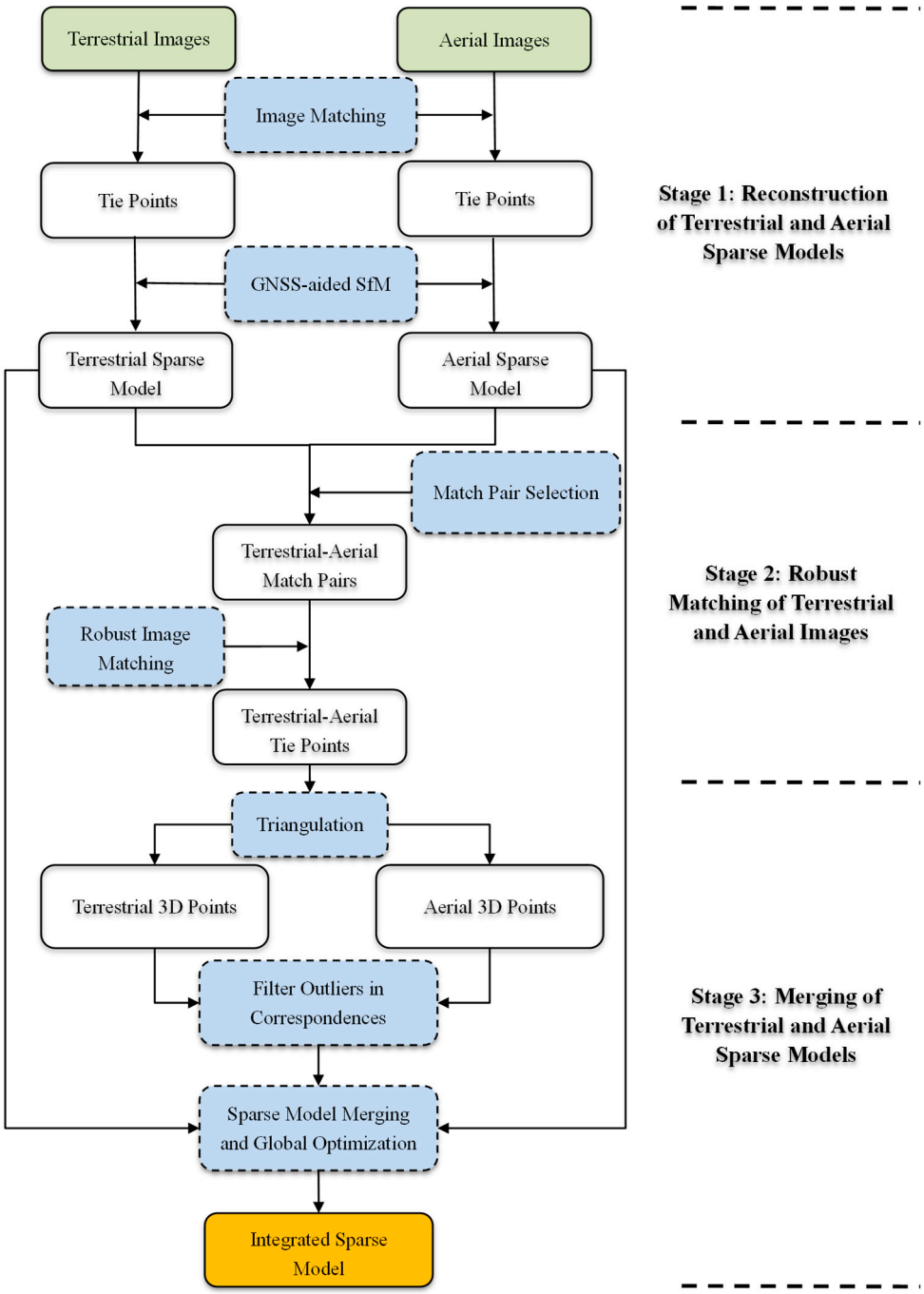


Figure 1. Workflow of the proposed methodology.

3.2. Reconstruction of Terrestrial and Aerial Sparse Models

The georeferenced sparse models are reconstructed from terrestrial and aerial images as follows. Image matching is first performed on the images. In the image matching process, RootSIFT [43] is used for feature point extraction and description. The feature points are matched using the approximate nearest neighbors (ANN) algorithm to determine initial matches. Then, the initial matches are verified based on the epipolar constraint with the RANSAC framework to generate geometrically consistent tie points. To speed up the matching process, match pairs are selected from K nearest neighbors (KNN) of each image. And image matching was only performed on the selected match pairs.

Based on the tie points, the GNSS-aided incremental SfM is used to reconstruct the georeferenced terrestrial and aerial sparse models in favor of its robustness and accuracy.

The incremental SfM procedure first selects an image pair and reconstructs an initial stereo model. Then, it grows the model by adding new images and globally optimizing all parameters in a loop. The object function for the global optimization is given by Equation (1).

$$E_1 = \sum_i \sum_j \rho_{ij} \left\| P(C_j, X_i) - x_{ij} \right\|^2 + p \sum_k \left\| M_k - S_k \right\|^2 \quad (1)$$

where P projects a 3D point X_i onto an image j, C_j represents camera parameters of image j, x_{ij} is an image observation, $\| \cdot \|$ denotes L2-norm, and ρ_{ij} is an indicator function. ρ_{ij} equals to 1 if X_i is visible to image j; otherwise, it equals 0. M_k is a position observation of image k, S_k is the estimation of the position, p is a weight that is calculated according to Equation (2).

$$p = \sigma_0^2 / \sigma_{GNSS}^2 \quad (2)$$

where σ_0 is the accuracy of image observations, σ_{GNSS} is the accuracy of the GNSS observations. After the GNSS-aided SfM, georeferenced terrestrial and aerial sparse models are obtained. In this study, a sparse model of a scene indicates the model reconstructed by the GNSS-aided SfM, which is composed of a sparse point cloud, exterior and interior orientations of images, and camera calibration parameters. The sparse point cloud is derived from geometrically consistent tie points. The exterior orientations of an image define the position and rotation of the image under the object coordinate system. The interior orientations include the focal length f and the offset of the principal point (cx, cy). The Brown's radial distortion model with three parameters (k1, k2, and k3) was used for camera calibration.

3.3. Robust Matching of Terrestrial and Aerial Images

Based on the terrestrial and aerial sparse models, the terrestrial and aerial images are robustly matched as follows. Firstly, the normal vector of each 3D point from the terrestrial sparse model is estimated. Based on the estimated normal vectors, the observability of each terrestrial point in the aerial images is determined. Terrestrial–aerial match pairs are selected by projecting terrestrial points to the aerial images in which they are observable. Based on the selected match pairs, the terrestrial and aerial images are robustly matched.

The normal vector of a point from the terrestrial point cloud is estimated by averaging its normalized observation vectors. The normal vector estimation is illustrated in Figure 2. In this top-view illustration, an estimated 3D point P on the facade of a building is observable in images I_1, I_2, I_3, and I_4. To calculate the normal vector of P, the observation vectors PS_1, PS_2, PS_3, and PS_4 are firstly calculated based on the estimated positions of P and respective perspective centers. Assume that the observation vectors are uniformly distributed in space, and the normal vector N of P is approximated by normalizing and averaging the observation vectors.

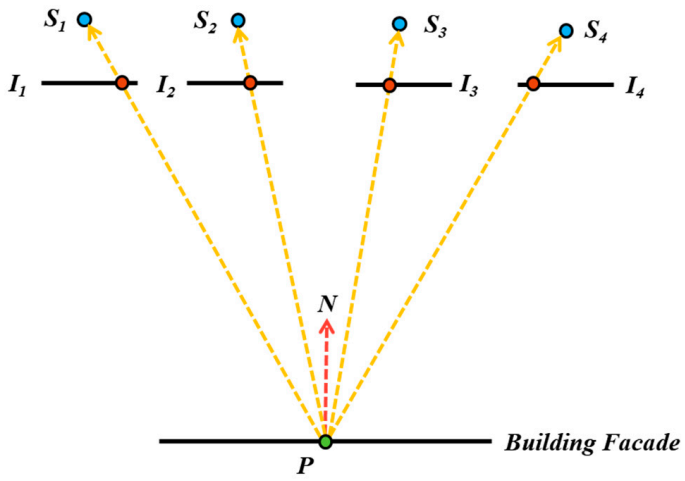


Figure 2. Illustration of normal vector approximation.

To improve the robustness and efficiency of the integrated reconstruction pipeline, cross-platform image matching is conducted on selected image pairs. Based on the normal vector estimation, match pairs between cross-platform images are selected as follows. Firstly, the procedure iterates through the terrestrial point cloud and calculates virtual observation vectors for each point. A virtual observation vector is the vector from a point to the perspective center of an aerial image. Then, for each point, the aerial images in which the point is observable are obtained by projecting the point onto the potential aerial images based on the Collinearity Equations. An aerial image is considered a potential aerial image of a point if the angle between the normal vector of the point and the corresponding virtual observation vector is smaller than a given threshold V_T. If the projection of the point is within the valid area of an aerial image, the point is considered observable in the image. If a point is observable both in an aerial image and a terrestrial image, the point is considered a common point between these two images. If the number of common points between a terrestrial image and an aerial image is larger than a given threshold N_T, these two images are considered as a valid match pair. For each terrestrial image, all aerial images that form a match pair with it can be determined.

Based on the selected match pairs, an image matching scheme is proposed to match cross-platform images. The relationship between the terrestrial and aerial images is illustrated in Figure 3. In the figure, the point on the facade of a building is observable in three terrestrial images and three aerial images. T_j is a terrestrial image. A_l, A_m, and A_n are selected aerial images that form match pairs with T_j. T_i and T_k are neighboring terrestrial images of T_j. In this study, the LoFTR algorithm [34] is used for matching the terrestrial and aerial images for its high robustness. It should be noted that image matching using LoFTR is directed, which means that the tie points generated by matching T_i against T_j are different from those by matching T_j against T_i. In this study, pairwise image matching is conducted as follows. For each terrestrial image T_j, it is matched against all the selected aerial images and its K nearest terrestrial images. The initial matches are verified based on the epipolar constraint with the RANSAC framework to obtain geometrically consistent tie points. The algorithm for robust cross-platform image matching, including the proposed match pair selection method and image matching scheme, is described by Algorithm 1.

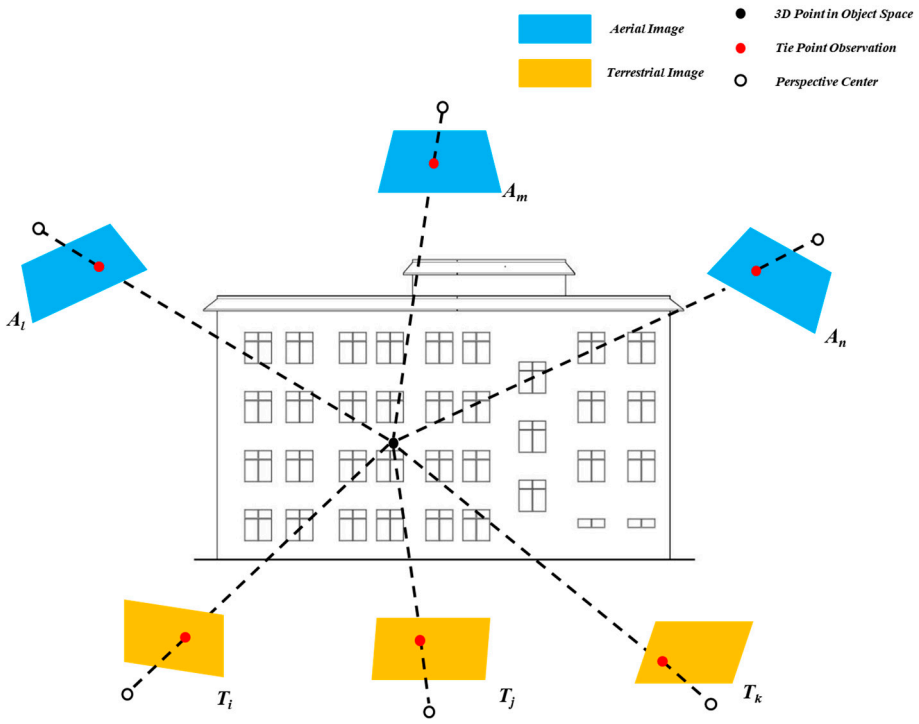


Figure 3. Illustration of the relationship between terrestrial and aerial images.

Algorithm 1 Robust cross-platform image matching

Input: terrestrial image set T, aerial image set A, terrestrial sparse point cloud $P = \{p_i | i \in \{1,2,3,\ldots,r\}\}$, threshold V_T to constrain the angle between a normal vector and a virtual observation vector, minimum number of common points N_T between a terrestrial image and an aerial image, K for searching nearest neighbors of a terrestrial image
Output: a set of tie points $G = \{t_k = \{I_i \rightarrow (u,v)\} | k \in \{1,2,3,\ldots,q\},\ i \in \{1,2,3,\ldots,m+n\}\}$ that record each group of tie points t_k including their position (u,v) in observable images $\{I_i\}$
Initialization: $G = \Phi$, $C = \{(T_i, A_j) \rightarrow c | i \in \{1,2,3,\ldots,m\},\ j \in \{1,2,3,\ldots,n\}\}$ which records the number of common points between a terrestrial image T_i and an aerial image A_j
for each point p_i in P
    find terrestrial images $VT = \{T_i\}$ in which p_i is observable
    calculate observation vectors of p_i
    calculate normal vector N_i of p_i by averaging observation vectors
    for each aerial image A_j in A
        calculate virtual observation vector V_{ij} of p_i
        if angle(N_i, V_{ij}) < V_T and projection of p_i is within the valid area of A_j
            for each image T_i in VT
                increment $C(T_i, A_j)$
            end for
        end if
    end for
end for
for each terrestrial image T_i in T
    for each aerial image A_i in A
        if $(CT_i, A_j) > N_T$
            build a match pair (T_i, A_j)
        end if
    end for
end for
organize match pairs to $M = \left\{ T_i \rightarrow \left\{ A_j \right\} \middle| i \in \{1,2,3,\ldots,m\},\ j \in \{1,2,3,\ldots,n\} \right\}$ which stores aerial images $\left\{ A_j \right\}$ that form a match pair with a terrestrial image T_i
for each terrestrial image T_i in T
    find terrestrial images $MT = \left\{ T_j \right\}$ that are K nearest neighbors of T_i
    match T_i against aerial images in $M(T_i)$
    match T_i against terrestrial images in MT
    add the tie points as a group to G
end for
return G

3.4. Accurate Fusion of Terrestrial and Aerial Sparse Models

The terrestrial and aerial sparse models are merged to generate an integrated model of the scene as follows. Firstly, the cross-platform tie points are separated into terrestrial and aerial groups. Then, the two groups of tie points are triangulated based on the terrestrial and aerial sparse models, respectively. Thirdly, correspondences between the two groups of triangulated 3D points are found, and outliers are filtered from the correspondences. Finally, the terrestrial and aerial sparse models are merged based on the refined correspondences, and the integrated model is globally optimized to further improve the accuracy of the reconstruction.

The terrestrial and aerial groups of tie points are robustly triangulated with the RANSAC framework. The triangulation of the tie points is illustrated in Figure 4. The triangulated 3D points are labeled with black points in the figure. These 3D points are triangulated from tie point observations of points A, B, and C. The tie point observations of the same point are labeled with the same color. As illustrated by Figure 4, the points triangulated from the aerial images do not coincide with those from the terrestrial images.

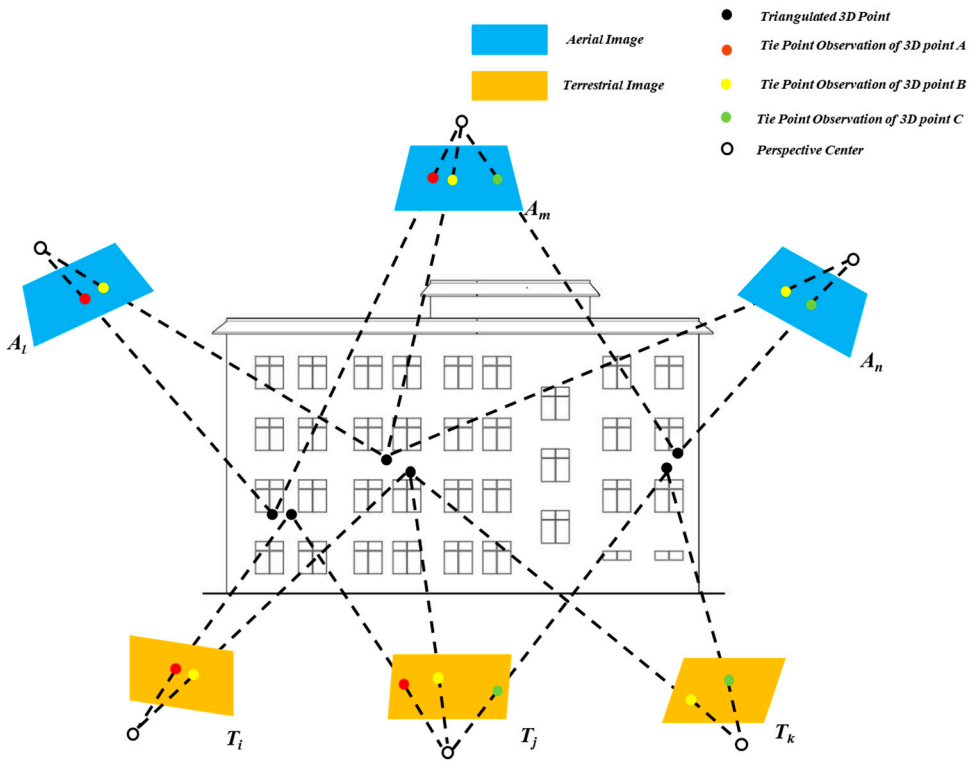


Figure 4. Triangulation of cross-platform tie points.

To precisely merge the terrestrial and aerial sparse models, a 3D similarity transformation from the terrestrial model to the aerial model is estimated based on the 3D correspondences that are found in the triangulated terrestrial and aerial 3D points. To improve the accuracy and robustness of the model fusion process, outliers in the correspondences are detected based on the following derivations. Assume that X_T^i and X_A^i are a pair of 3D correspondences that are triangulated based on respective terrestrial and aerial sparse models. Assume

$$X_T^i = X^i + e_T^i \tag{3}$$

$$X_A^i = X^i + e_A^i \tag{4}$$

where X^i is the true position of the object point corresponding to the correspondences, e_T^i and e_A^i are residual error vectors corresponding to terrestrial and aerial sparse models, respectively. Assume that e_T^i and e_A^i are subject to the three-dimensional normal distributions defined as follows.

$${e_T}^i \sim N(\mu_T, \Sigma_T) \tag{5}$$

$${e_A}^i \sim N(\mu_A, \Sigma_A) \tag{6}$$

where $N(\mu_T, \Sigma_T)$ and $N(\mu_A, \Sigma_A)$ define the positioning bias and accuracy of the terrestrial and aerial sparse models, respectively. Based on the above assumptions, the positional difference between X_T^i and X_A^i is subject to the three-dimensional normal distribution given by Equation (7).

$${X_T}^i - {X_A}^i \sim N(\mu_T - \mu_A, \Sigma_T + \Sigma_A) \tag{7}$$

Based on the above derivations, statistics of positional differences are used to detect and remove outliers from the correspondences. Specifically, mean and standard deviation values along the X, Y, and Z axes are calculated from the positional differences of all pairs of 3D correspondences. A pair of correspondences is considered an outlier as long as the positional difference along any axis is outside the range of the respective mean value plus and minus N_p times the standard deviation. After outliers are removed from the

correspondences, a 3D similarity transformation is estimated. For refined correspondences $\{P_i\}$ and $\{Q_i\}$, the similarity transformation is estimated by minimizing the object function given by Equation (8).

$$E_2 = \frac{1}{n}\sum_{i}^{n} h(\|Q_i - (\lambda R P_i + t)\|^2) \tag{8}$$

where R is the rotation matrix, t is the translation vector, λ is the scaling factor, and h is the Huber loss function, $\|\cdot\|$ denotes L2-norm. Then, the estimated similarity transformation is applied to the terrestrial sparse model. And the transformed terrestrial model is locally optimized with correspondences fixed to their positions in the aerial model. After merging the locally optimized terrestrial model with the aerial model, the integrated model is globally optimized by minimizing the object function given by Equation (1). The algorithm for robust sparse model fusion is described by Algorithm 2.

Algorithm 2 Robust sparse model fusion

Input: terrestrial sparse model M_T, aerial sparse model M_A, tie points G
Output: an integrated model M_{AT}
separate G to terrestrial tie points G_T and aerial tie points G_A
initialize terrestrial point set P_T and aerial point set P_A
for each group of tie points g_i in G_T
    robustly triangulate g_i to obtain a 3D point p_i based on M_T
    add p_i to P_T
end for
for each group of tie points g_j in G_A
    robustly triangulate g_j to obtain a 3D point p_j based on M_A
    add p_j to P_A
end for
find 3D correspondences $C = \left\{ c_i \to (p_j, p_k) \middle| i \in \{1,2,3,\ldots,m\},\ p_j \in P_T,\ p_k \in P_A \right\}$ between P_T and P_A
calculate positional differences of 3D correspondences, derive mean and standard deviation values along three axes
filter outliers in C based on the Three-Sigma Rule
estimate a 3D similarity transformation T based on refined correspondences
transform M_T to M_T' based on T, and locally optimize M_T'
merge M_T' and M_A to integrated model M_{AT}
globally optimize M_{AT}
return M_{AT}

4. Experimental Results

The proposed methodology was evaluated using five publicly available benchmark datasets. Firstly, the specifications of the datasets are detailed. Secondly, experimental results of sparse model reconstruction, terrestrial–aerial image matching, and terrestrial–aerial sparse model fusion are presented. Finally, the proposed methodology was compared with a state-of-the-art methodology and three software packages. The proposed methodology is implemented based on the open-source software OpenMVG (version 1.6) [44]. The LoFTR model from the Kornia library (version 0.7.0) [45] was used for matching the terrestrial and aerial images. The LoFTR model was pre-trained on the MegaDepth dataset [46]. The proposed algorithms were mainly implemented in the C++ programming language. And scripts for data preprocessing and LoFTR-based image matching were implemented in the Python programming language. All of the experiments were performed on a Dell Precision 7530 mobile workstation. The workstation is equipped with a Windows 10 Professional operating system, an Intel i9-8950HK CPU (6 cores, 2.9 GHz), an NVIDIA Quadro P3200 GPU, and 32 GB memory.

4.1. Specifications of Datasets

The datasets used for the experiments were downloaded from the website provided by the research team from Southwest Jiaotong University (SWTJU), China [9]. The Center and Zeche datasets were initially provided by the International Society for Photogrammetry and Remote Sensing (ISPRS) and the European SDR (EuroSDR) and acquired with the ISPRS scientific initiative in 2014 and 2015. The datasets were collected around two buildings in Dortmund, Germany. The SWJTU-LIB, SWJTU-BLD, and SWJTU-RES datasets were acquired and provided by the SWTJU research team. The SWJTU-LIB, SWJTU-BLD, and SWJTU-RES datasets were acquired around a library building, a research building, and a residential building, respectively. Specifications of the datasets are listed in Table 1.

Table 1. Dataset specifications.

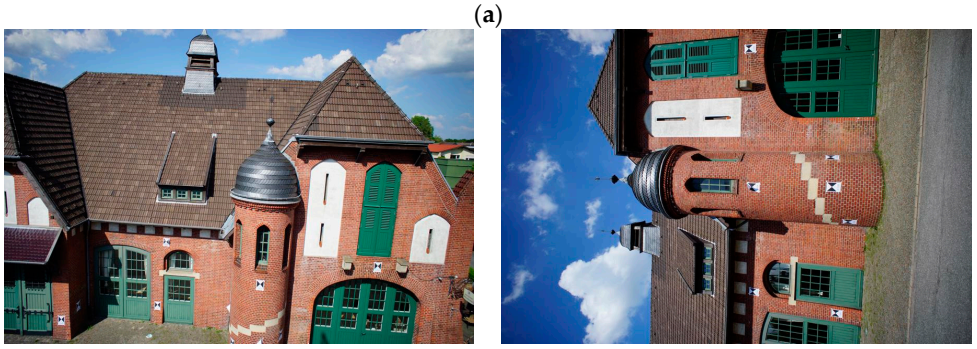
Dataset	Terrestrial Images	Aerial Images	Aerial Camera	Terrestrial Camera	Spatial Reference
Center	203	146	SONY Nex-7, 16 mm, 4000 × 6000		WGS 1984
Zeche	172	147	SONY Nex-7, 16 mm, 4000 × 6000		WGS 1984
SWJTU-LIB	78	123	SONY ILCE-5100, 40 mm, 4000 × 6000	Canon EOS M6, 19 mm, 4000 × 6000	WGS 1984/UTM
SWJTU-BLD	88	207	SONY ILCE-5100, 28/40 mm, 4000 × 6000	Canon EOS M6, 18 mm, 4000 × 6000	WGS 1984/UTM
SWJTU-RES	192	92	SONY ILCE-5100, 40 mm, 4000 × 6000	DJI spark, 6 mm, 3040 × 4056	WGS 1984/UTM

Each of the datasets is composed of several hundreds of aerial and terrestrial images. The aerial and terrestrial images of the ISPRS datasets were acquired using the same camera. The aerial and terrestrial images of the SWTJU datasets were acquired using two cameras. The image resolution of most cameras is 4000 by 6000 pixels. Image positioning observations, including latitude, longitude, and altitude defined under the World Geodetic System 1984 (WGS 1984), are provided in the ISPRS datasets as EXIF tags. Positioning observations under the WGS 1984, as well as the UTM (Universal Transverse Mercator) coordinate system, are provided in the SWTJU datasets. In this study, the East–North–Up (ENU) coordinate system is used as the object coordinate system for processing the ISPRS datasets. The UTM coordinate system is used as the object coordinate system for processing the SWTJU datasets. Sample images of the datasets are shown in Figure 5. The left column shows aerial images of five scenes, and the right column shows terrestrial images of the scenes.

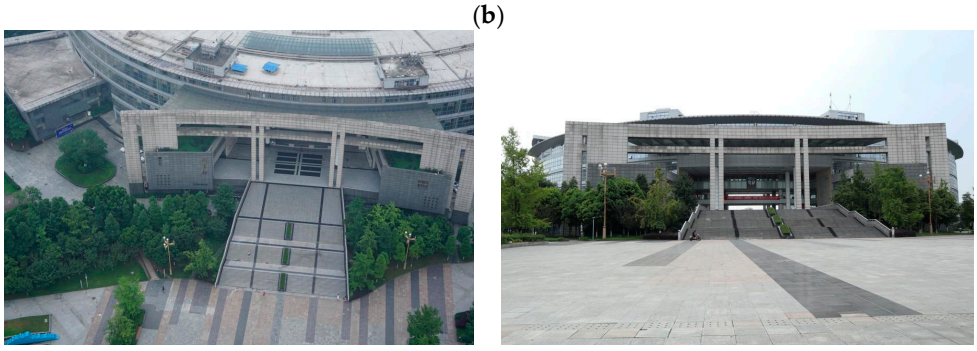
4.2. Terrestrial and Aerial Sparse Reconstructions

For each dataset, a terrestrial sparse model and an aerial sparse model were reconstructed from the terrestrial and aerial images, respectively. To match images efficiently, each image was matched against its ten nearest neighbors. For the GNSS-aided incremental SfM, the accuracy of the aerial GNSS observations was set to 0.1 m for all datasets. The accuracy of the positioning observations of the terrestrial images of the Center, Zeche, SWJTU-LIB, SWJTU-BLD, and SWJTU-RES datasets were set to 10 m, 10 m, 0.1 m, 0.1 m and 0.1 m, respectively. And the accuracy of image observations was set to 1 pixel. The statistics of the sparse reconstructions are listed in Table 2. All terrestrial and aerial images were registered during the sparse reconstructions for each dataset. Hundreds of thousands of 3D points were reconstructed from each dataset. The column of root-mean-squared error (RMSE) shows that all of the sparse reconstructions achieved subpixel accuracy. The experimental result shows high robustness and accuracy of the sparse reconstructions.

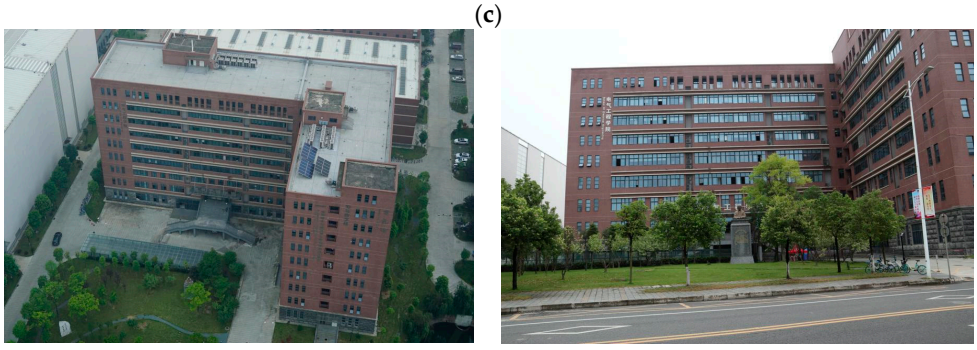
(**a**)



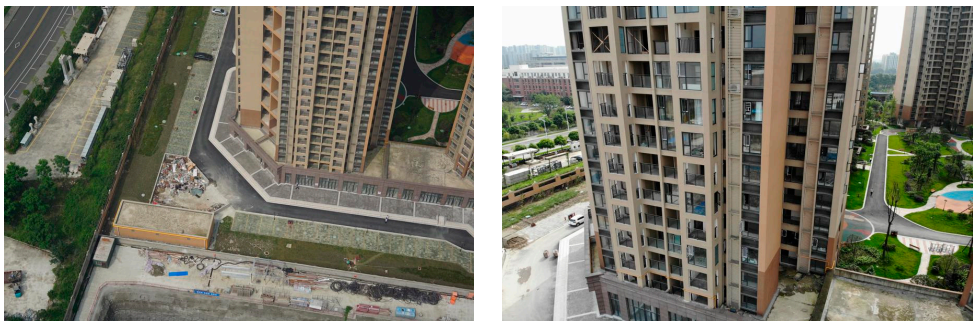
(**b**)



(**c**)



(**d**)



(**e**)

Figure 5. Sample images: (**a**) Center; (**b**) Zeche; (**c**) SWJTU-LIB; (**d**) SWJTU-BLD; (**e**) SWJTU-RES.

Table 2. Statistics of sparse reconstructions of five datasets.

Dataset	Reconstructed Points		RMSE (Pixel)	
	Aerial	Terrestrial	Aerial	Terrestrial
Center	612,504	349,588	0.56	0.59
Zeche	729,947	103,1828	0.45	0.41
SWJTU-LIB	153,397	40,200	0.67	0.78
SWJTU-BLD	143,623	99,669	0.56	0.72
SWJTU-RES	191,494	222,849	0.61	0.68

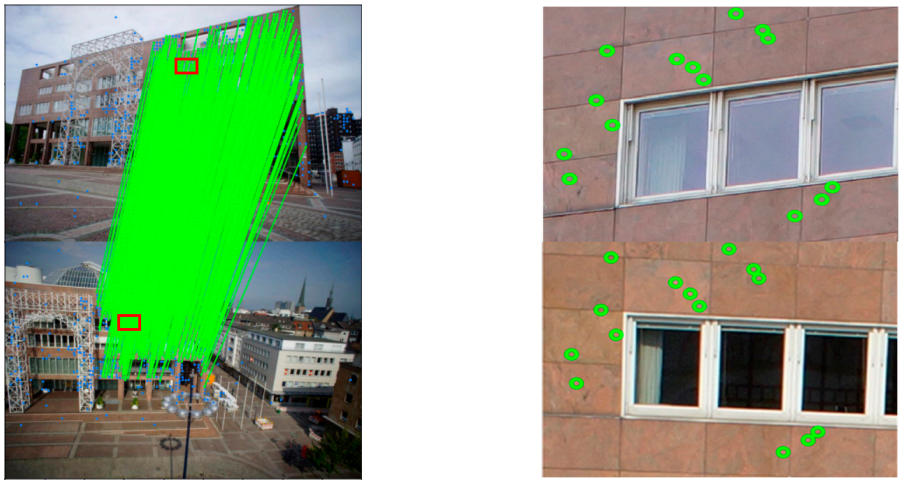
4.3. Cross-Platform Image Matching

The parameters for match pair selection between the terrestrial and aerial images were set as follows. The angle threshold V_T between the normal vector of a point and a virtual observation vector was set to 40 degrees. The threshold N_T for the validation of a terrestrial–aerial match pair was set to 300. The statistics of the selected terrestrial–aerial match pairs are listed in Table 3. The table shows that hundreds to thousands of match pairs were selected for the datasets. The maximum number of match pairs of a dataset shows the highest number of aerial images that overlap with a terrestrial image. Most of the minimum numbers are zero. It indicates that there exists at least one terrestrial image which does not overlap with any aerial images. The average number of match pairs correlates with the overlap ratio between the terrestrial and aerial images of a dataset. The standard deviation (STD) shows the variation of selected match pairs within a dataset.

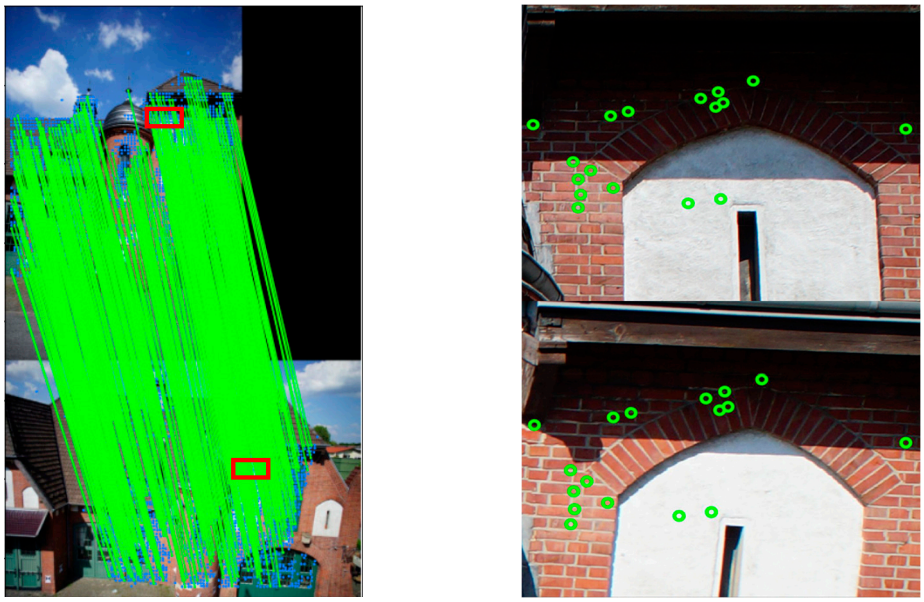
Table 3. Statistics of selected terrestrial–aerial match pairs.

Dataset	Total	Maximum	Minimum	Mean	STD
Center	2906	35	0	14.32	10.54
Zeche	3069	33	0	17.84	4.46
SWJTU-LIB	527	8	5	6.76	1.11
SWJTU-BLD	1349	28	0	15.40	7.06
SWJTU-RES	954	16	0	4.97	7.13

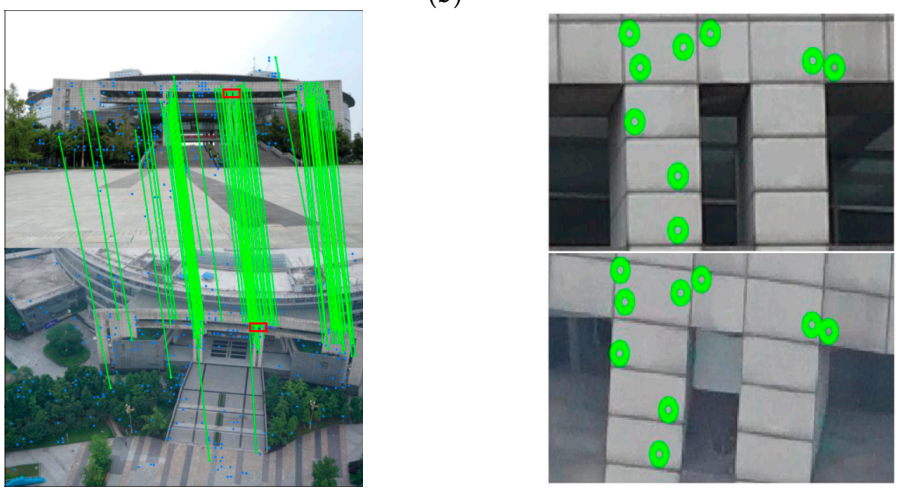
Each terrestrial image was matched against the aerial images that formed a match pair with it. And the terrestrial image was also matched against two neighboring terrestrial images. It should be noted that all images were subsampled to 600 by 900 pixels to make the cross-platform image matching process efficient. Then, the positions of the established tie points in the subsampled images were transferred back to the original images. Figure 6 shows matching results of the terrestrial and aerial images shown in Figure 5. For each dataset, the figure in the left column shows the geometrically verified tie points in two images that are connected using green lines. The figures in the right column are enlargements of the red rectangles in the left figure. The green circles in the figures on the right show the tie points, and the blue points in the figures on the left are outliers detected in initial matches. It can be seen from Figure 6 that most of the matches are visually correct, which demonstrates the robustness of the LoFTR model. It can also be seen that much fewer tie points were established between the cross-platform images of the SWJTU datasets than the ISPRS datasets, which indicates larger viewpoint and scale variations of the SWJTU datasets.



(**a**)

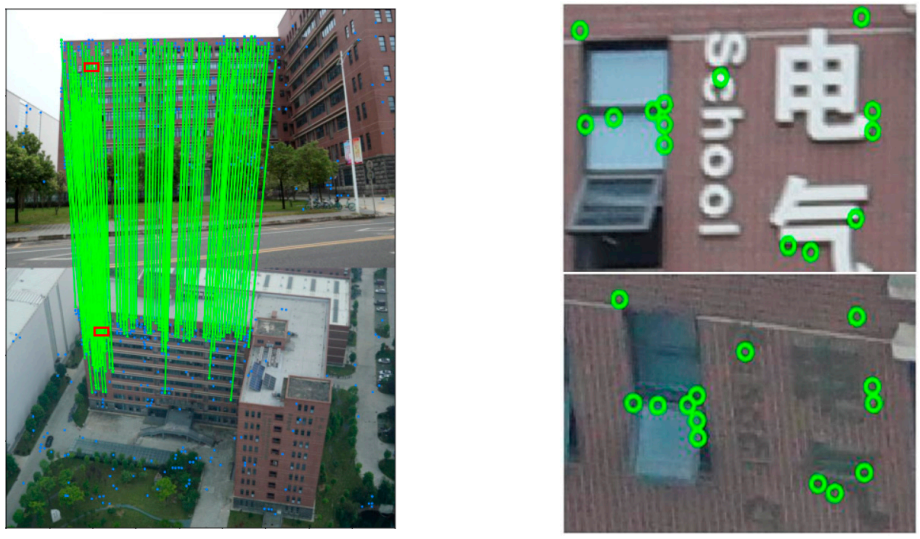


(**b**)

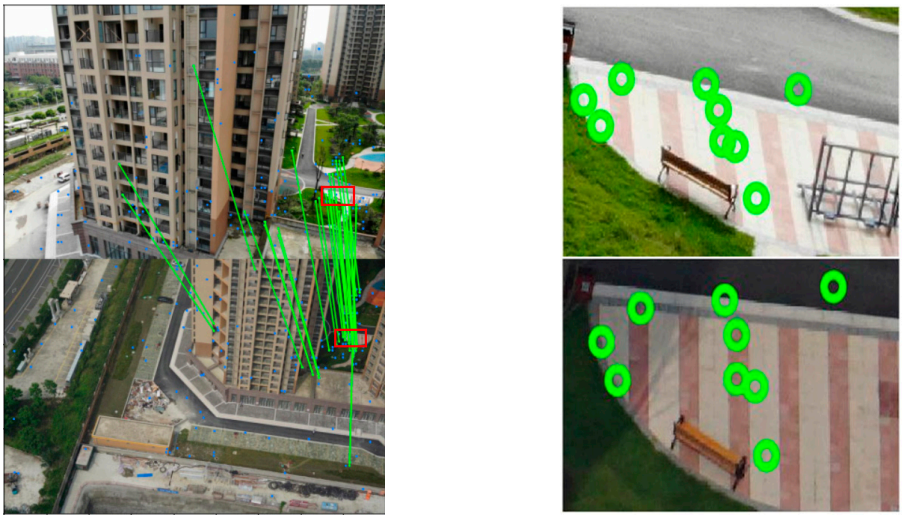


(**c**)

Figure 6. *Cont.*



(**d**)



(**e**)

Figure 6. Terrestrial–aerial image matching: (**a**) Center; (**b**) Zeche; (**c**) SWJTU-LIB; (**d**) SWJTU-BLD; (**e**) SWJTU-RES.

The statistics of terrestrial–aerial tie points are listed in Table 4. As can be seen from the table, tens to hundreds of tie points, on average, were established between a terrestrial–aerial image pair for the datasets. The average number of tie points of the SWJTU datasets is much lower than the ISPRS datasets, which indicates the difficulty of the SWJTU datasets. The standard deviation shows the variation of established tie points among the selected match pairs within a dataset.

Table 4. Statistics of cross-platform tie points.

Dataset	Minimum	Maximum	Mean	STD
Center	9	1415	99	161.65
Zeche	10	512	144	154.22
SWJTU-LIB	12	142	35	26.48
SWJTU-BLD	10	214	31	32.44
SWJTU-RES	9	98	12	15.59

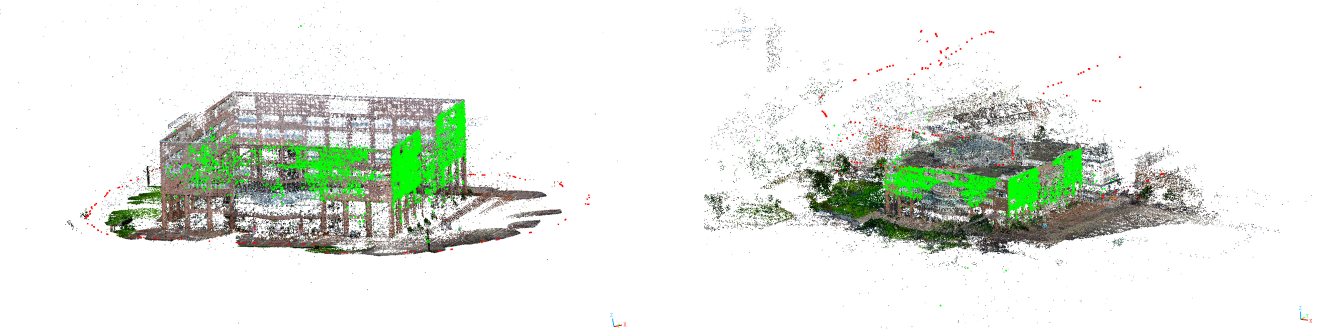
4.4. Triangulation of Tie Points and Sparse Model Fusion

The cross-platform tie points were triangulated based on the terrestrial and aerial sparse models, respectively. The number of triangulated points and 3D correspondences for each dataset is listed in Table 5. It can be seen that hundreds to tens of thousands of 3D points were triangulated for the datasets. The number of triangulated points in the ISPRS datasets is much higher than in the SWJTU datasets, as many more tie points are established in the ISPRS datasets. It is also found that the number of triangulated points from the aerial sparse model is higher than that from the terrestrial sparse model for each dataset. This is mainly because a terrestrial image was matched against 4.97 to 15.40 aerial images on average (cf. Table 3). And the same terrestrial image was matched against only two neighboring terrestrial images. It can be seen from the last column that hundreds of 3D correspondences were established in the SWJTU datasets. More 3D correspondences were obtained in the ISPRS datasets as more tie points were established.

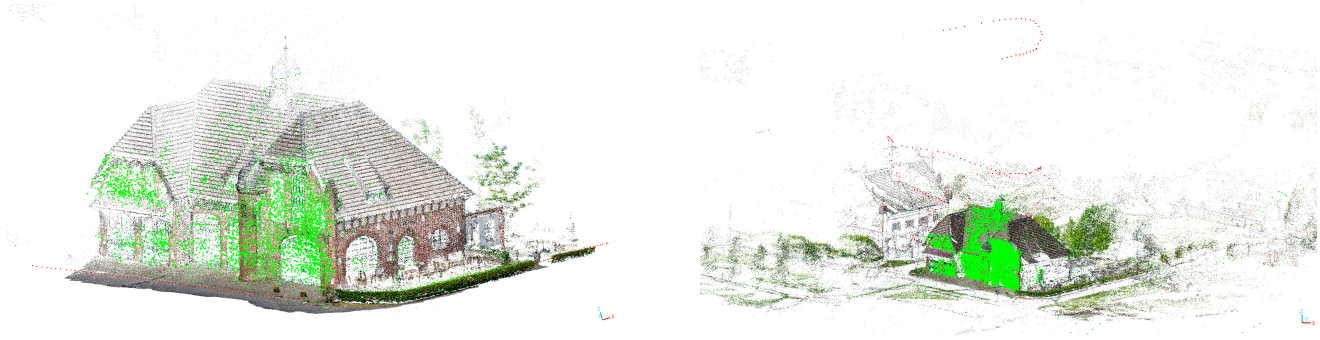
Table 5. Statistics of triangulated points.

Dataset	Aerial	Terrestrial	3D Correspondences
Center	26,701	13,008	10,225
Zeche	54,866	4986	1047
SWJTU-LIB	617	416	339
SWJTU-BLD	1941	700	459
SWJTU-RES	1754	307	166

Figure 7 shows the triangulated points overlaid on respective sparse models. The triangulated points are marked with green points. The left column shows the 3D points triangulated from the terrestrial sparse model of each dataset, and the aerial triangulated points are shown on the right. It can be seen that the triangulated points are mainly located on the facade of the buildings. The terrestrial triangulated points correspond well to the aerial triangulated points for each dataset. The figures also demonstrate the correctness of the terrestrial–aerial image matching results.



(**a**)



(**b**)

Figure 7. *Cont.*

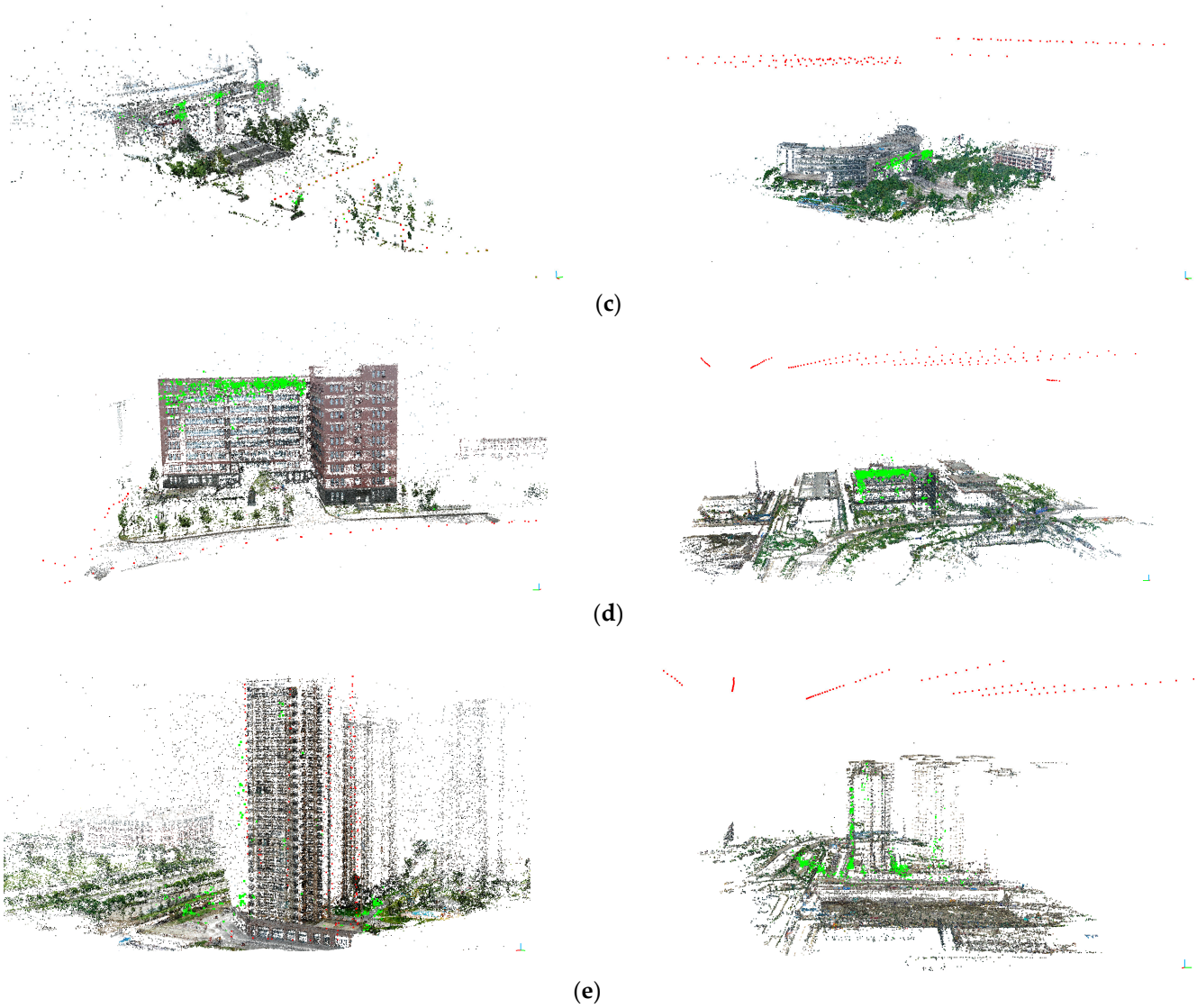


Figure 7. Triangulated 3D points overlaid on sparse models: (**a**) Center; (**b**) Zeche; (**c**) SWJTU-LIB; (**d**) SWJTU-BLD; (**e**) SWJTU-RES.

Before merging the reconstructed terrestrial and aerial sparse models, false correspondences were detected based on the proposed outlier detection method. Statistics of positional differences between the correspondences are listed in Table 6. The maximum and minimum values show the range of the positional difference along three axes. The *Mean* values of a dataset reflect the positional biases between the terrestrial and aerial sparse models of the dataset. The *STD* values of a dataset reflect the spatial proximity of the terrestrial and aerial sparse models in general. It can be seen from the table that the *Mean* values of the ISPRS datasets are larger than those of the SWJTU datasets, which indicates that there are larger positional biases between the terrestrial and aerial sparse models of the ISPRS datasets.

To detect outliers in the 3D correspondences, the threshold N_p was set to 3. The number of outliers detected in the 3D correspondences of the Center, Zeche, SWJTU-LIB, SWJTU-BLD, and SWJTU-RES datasets is 38, 16, 4, 21, and 5, respectively. Figure 8 shows tie point observations of an outlier detected in 3D correspondences of the SWJTU-RES dataset. Figure 8a,b show the tie point observations in two terrestrial images. Figure 8c,d show the tie point observations in two aerial images. The tie point observations are labeled with red circles in the terrestrial images and red plus signs in the aerial images. The figures show that the outlier escaped the outlier detection during both pairwise image matching and robust

triangulation. This type of outlier will reduce the accuracy of integrated reconstruction and even corrupt the terrestrial–aerial model fusion process.

Table 6. Statistics of positional difference between correspondences (Unit: Meters).

Dataset	Axis	Maximum	Minimum	Mean	STD
Center	X	543.00	−36.65	2.31	8.61
	Y	34.21	−248.90	−4.54	5.80
	Z	124.29	2.85	8.79	3.02
Zeche	X	21.88	−54.01	1.19	2.88
	Y	9.13	−109.34	−2.14	4.65
	Z	54.86	−9.30	−1.14	2.47
SWJTU-LIB	X	60.03	−18.61	0.03	6.32
	Y	33.21	−10.63	0.12	3.46
	Z	43.45	−14.47	0.61	4.71
SWJTU-BLD	X	20.09	−9.19	0.26	2.74
	Y	35.41	−43.42	0.80	4.69
	Z	10.48	−6.04	0.36	1.21
SWJTU-RES	X	10.12	−3.51	0.43	1.55
	Y	26.44	−58.44	−0.73	7.67
	Z	37.66	−2.26	0.83	3.07

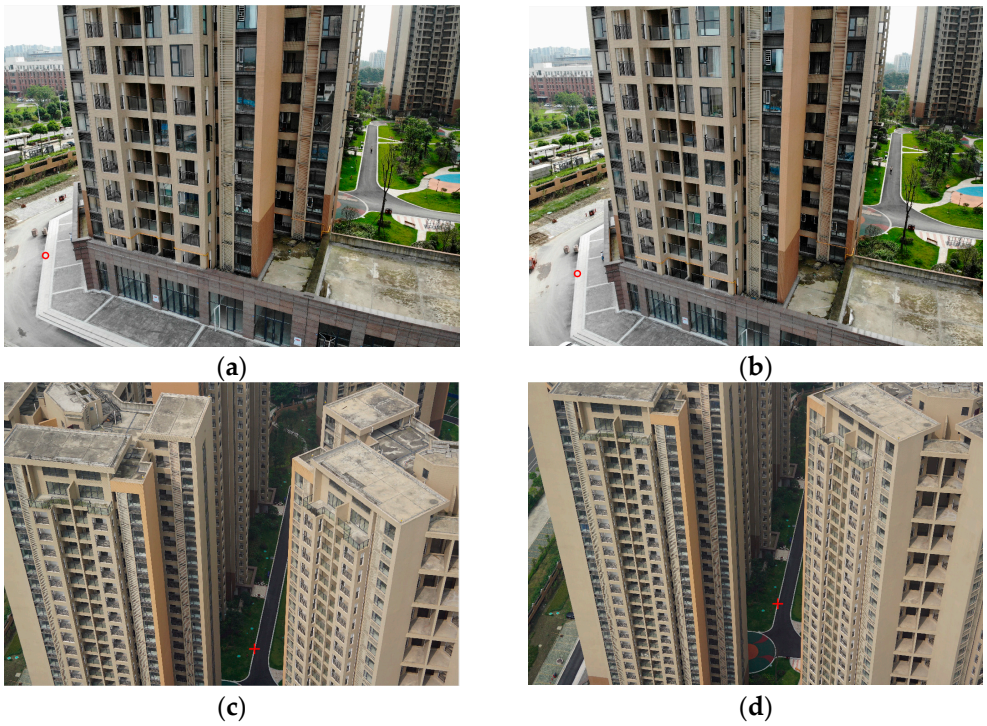
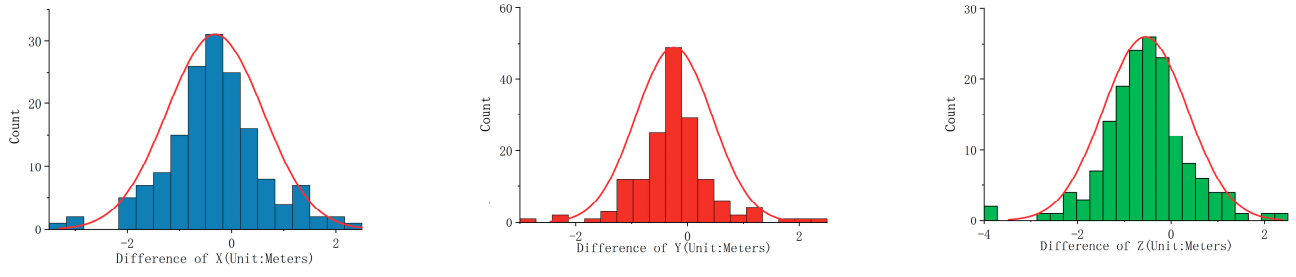


Figure 8. Image observations of an outlier detected in 3D correspondences of the SWJTU-RES dataset: (**a**) DJI_0160.JPG; (**b**) DJI_0161.JPG; (**c**) W0401.jpg; (**d**) W0402.jpg.

Figure 9 illustrates distributions of positional differences of the SWJTU-RES dataset after the removal of the outliers. The red curves show the normal distributions fitted using the inlier samples. It can be seen that the data fit the model well along each axis, which justifies the assumptions of the proposed outlier detection method.



(**a**) (**b**) (**c**)

Figure 9. Histograms and fitted normal distributions of positional differences of SWJTU-RES dataset: (**a**) X; (**b**) Y; (**c**) Z.

After the removal of the outliers from the correspondences, the terrestrial and aerial sparse models were merged and optimized for each dataset. The optimized sparse models were exported to Metashape for further dense reconstruction and texture mapping to generate textured models. Figure 10 shows the reconstructed sparse models and textured models. The reconstructed sparse models are shown in the left column. The textured models reconstructed based on the integrated sparse models are shown in the middle column. The right column shows textured models reconstructed using only aerial images. It can be seen that the terrestrial and aerial sparse models were merged well for all the datasets. The sparse models are visually correct, and no observable distortion is found in the reconstructed scenes.

(**a**)

(**b**)

Figure 10. *Cont.*

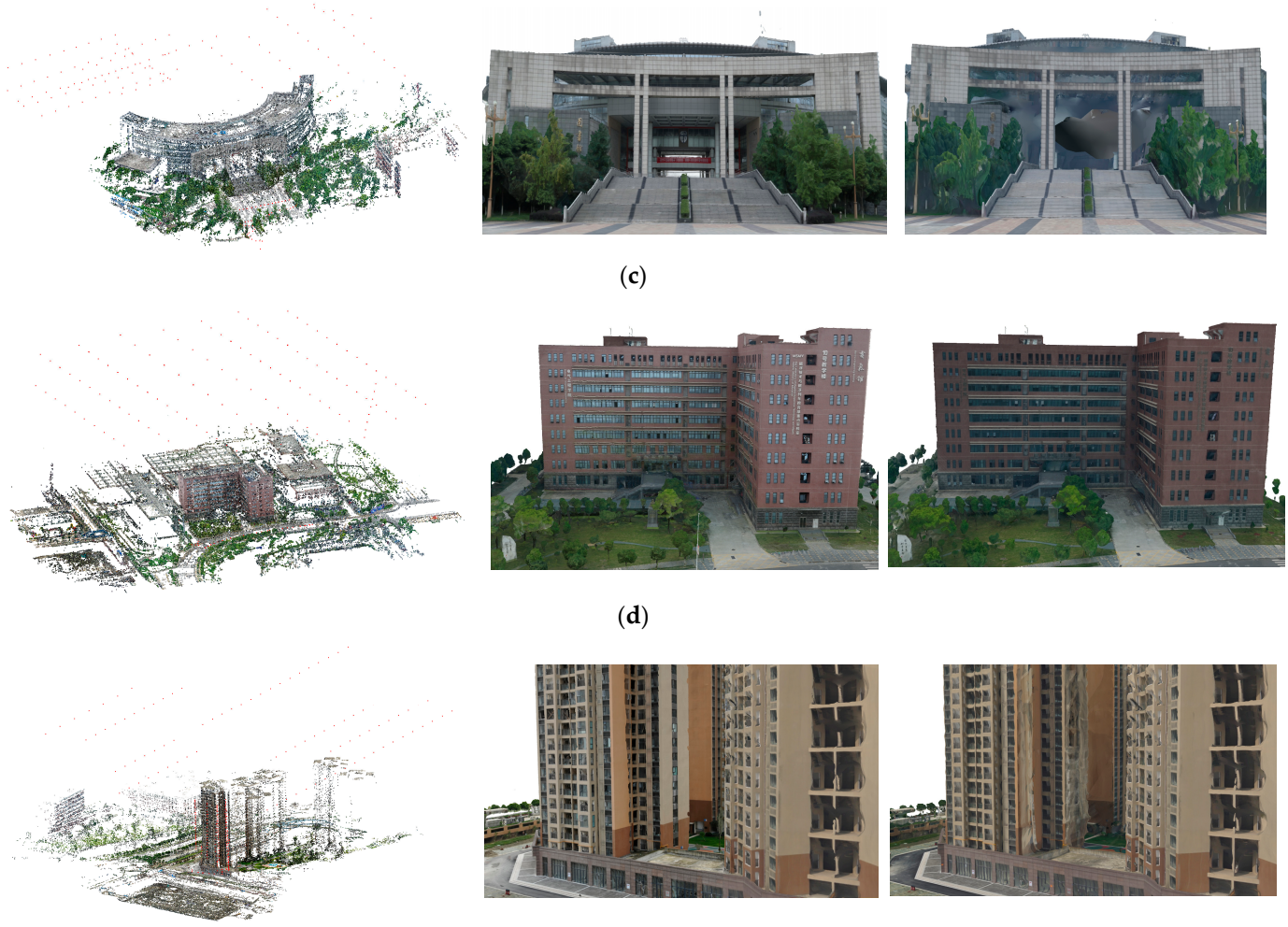


Figure 10. Reconstructed sparse models and textured models: (**a**) Center; (**b**) Zeche; (**c**) SWJTU-LIB; (**d**) SWJTU-BLD; (**e**) SWJTU-RES.

The textured models reconstructed using both terrestrial and aerial images show more details of buildings compared to the textured models reconstructed using only aerial images. The structures of the building facade are more complete and accurate, as terrestrial images provide observations of the structures that are heavily occluded in aerial images. The experimental results demonstrate the effectiveness of the proposed methodology.

4.5. Comparison of Pipelines

The proposed pipeline was compared with a state-of-the-art methodology [9] and software packages, including Metashape, COLMAP [47] and OpenMVG [44]. The configurations of the software packages are listed in Table 7.

- Metashape is a commercial software package widely used by the research community and the industry for the photogrammetric processing of aerial images. In this study, match pairs are selected based on both position and visual similarity of images. The accuracy of the positioning observations of the images of the Center, Zeche, SWJTU-LIB, SWJTU-BLD, and SWJTU-RES datasets was set to 10 m, 10 m, 0.1 m, 0.1 m, and 0.1 m. Metashape leverages a hierarchical SfM for sparse reconstruction. The highest accuracy and the adaptive camera model fitting were set for the SfM reconstruction. The other parameters were set to default values.

- COLMAP is an open-source software package widely used by the research community for image-based 3D reconstruction. In this study, match pairs were selected using a vocabulary tree, and the vocabulary tree file with 32K visual words was downloaded from the official website of COLMAP. SIFT is used for feature point extraction, and all SIFT feature points extracted from raw images were used for pairwise image matching. An incremental SfM strategy was used for sparse reconstruction. The other parameters were set to default values.
- OpenMVG is an open-source software package widely used by the research community for image-based sparse reconstruction. In this study, an exhaustive strategy was used for match pair selection. RootSIFT is used for feature point extraction, and all extracted feature points were used for pairwise image matching. An incremental SfM reconstruction was used for sparse reconstruction. The other parameters were set to default values.

Table 7. Specifications and parameter settings of software packages.

Software Package	Match Pair Selection	Image Matching	SfM	Version	Source
Metashape	Position and visual similarity	Highest accuracy, maximum features: 40,000, maximum tie points: 4000	Hierarchical	2.0.0	https://www.agisoft.com/, accessed on 20 June 2023
COLMAP	Voctree	Full resolution, all feature points	Incremental	3.8	https://github.com/colmap/colmap, accessed on 20 June 2023
OpenMVG	Exhaustive	Full resolution, all feature points	Incremental	2.0	https://github.com/openMVG/openMVG, accessed on 20 June 2023

The terrestrial–aerial integrated reconstruction results are listed in Table 8. It can be seen that the proposed pipeline registered all images for all datasets. Zhu et al.'s method failed to register some aerial images of the Zeche and SWJTU-RES datasets [9]. COLMAP achieved a complete reconstruction of the SWJTU-LIB dataset. However, the estimated positions of all terrestrial images are below the ground points reconstructed from the aerial images. Therefore, the terrestrial images were considered unregistered for this dataset. COLMAP also failed to register the terrestrial images of the SWJTU-BLD dataset. OpenMVG was unable to reconstruct a visually correct model for the SWJTU-BLD and SWJTU-RES datasets. Metashape failed to register aerial images of the SWJTU-RES dataset. The experimental results demonstrate the robustness of the proposed pipeline.

The reported accuracy values show that the proposed pipeline consistently achieved the highest accuracy on all the datasets. OpenMVG also obtained high accuracy on the Center, Zeche, and SWJTU-LIB datasets. The accuracy achieved by COLMAP is a little lower than OpenMVG. Although Metashape exhibited high robustness, it achieved relatively low SfM accuracy on the datasets. No reprojection errors of SfM reconstructions were reported in [9].

The proposed pipeline reconstructed more 3D points than COLMAP and Metashape. It reconstructed more than one million 3D points for the ISPRS datasets and hundreds of thousands of 3D points for the SWJTU datasets. OpenMVG reconstructed more 3D points than the proposed pipeline, as the exhaustive strategy was used by OpenMVG for match pair selection.

Table 8. Comparison of pipelines for terrestrial–aerial integrated reconstruction.

Dataset	Pipeline	Registered Aerial Images	Registered Terrestrial Images	3D Points	RMSE (Pix)
Center	OpenMVG	146/146	203/203	1,493,207	0.56
	Metashape	146/146	203/203	237,464	0.93
	COLMAP	146/146	203/203	235,895	0.70
	Zhu et al.'s method [9]	146/146	203/203	-	-
	Proposed	146/146	203/203	1,566,842	0.39
Zeche	OpenMVG	147/147	172/172	2,172,225	0.46
	Metashape	147/147	172/172	114,884	0.54
	COLMAP	147/147	172/172	296,377	0.71
	Zhu et al.'s method [9]	116/147	172/172	-	-
	Proposed	147/147	172/172	1,758,777	0.30
SWJTU-LIB	OpenMVG	123/123	78/78	1,033,323	0.65
	Metashape	123/123	78/78	132,543	1.09
	COLMAP	123/123	0/78	-	-
	Zhu et al.'s method [9]	123/123	78/78	-	-
	Proposed	123/123	78/78	355,783	0.45
SWJTU-BLD	OpenMVG	-	-	-	-
	Metashape	207/207	88/88	243,673	1.03
	COLMAP	207/207	0/88	-	-
	Zhu et al.'s method [9]	207/207	88/88	-	-
	Proposed	207/207	88/88	467,644	0.41
SWJTU-RES	OpenMVG	-	-	-	-
	Metashape	0/92	192/192	-	-
	COLMAP	92/92	192/192	224,844	0.75
	Zhu et al.'s method [9]	88/92	192/192	-	-
	Proposed	92/92	192/192	348,055	0.43

4.6. Ablation Studies

To demonstrate the effectiveness of the proposed image matching and outlier detection methods, the following two experiments were conducted. In the first experiment, the proposed match pair selection was used to generate the match pairs. The RootSIFT algorithm was used for matching all the match pairs. Then, outliers in the initial matches were removed using the epipolar constraint with the RANSAC framework. And a GNSS-aided incremental SfM reconstruction was conducted based on the tie points to build a sparse model for each dataset. In the second experiment, match pair selection and pairwise image matching were conducted using the proposed methods. A sparse model was reconstructed for each dataset using a GNSS-aided incremental SfM. The proposed outlier detection method for filtering 3D correspondences was not used in either of the experiments. The experiments were implemented based on the OpenMVG framework. The experimental results are listed in Table 9.

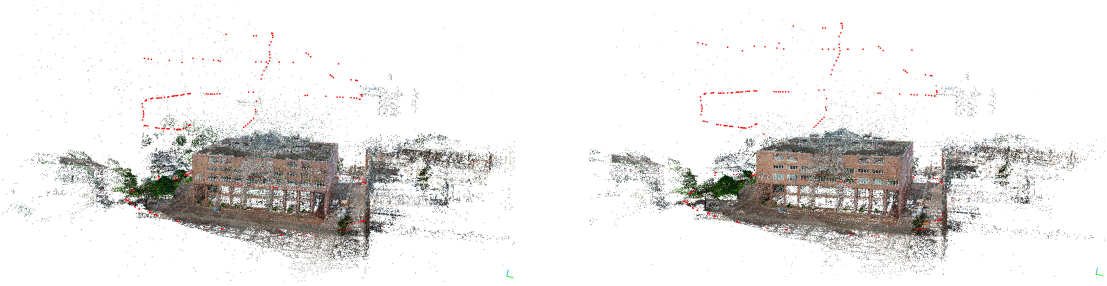
The results of the first experiment on the Center, Zeche, and SWJTU-LIB datasets are almost the same as those by OpenMVG from Table 8. It means that the proposed match pair selection method has little influence on the integrated reconstruction of relatively simple datasets. However, the first experiment achieved better reconstruction on the SWJTU-BLD dataset than OpenMVG, which indicates that precisely selected match pairs can improve the robustness of integrated reconstruction on complex datasets. However, the first experiment still failed to register any terrestrial images of the SWJTU-BLD dataset and could not reconstruct the SWJTU-RES dataset. In comparison, the second experiment achieved complete and accurate reconstruction on the SWJTU-RES dataset, which demonstrates that the proposed image matching method can improve the robustness of integrated reconstruction on complex datasets. The second experiment also achieved comparable results on the ISPRS datasets. However, the reconstruction failed on the SWJTU-LIB dataset in the same way as COLMAP. The estimated positions of all terrestrial images are below

the ground points reconstructed from the aerial images. And the reconstruction also failed on the SWJTU-RES dataset. It demonstrates that outliers in the LoFTR tie points corrupt the incremental SfM reconstruction process. It can be seen from Table 8 that a complete reconstruction of the SWJTU-RES dataset was achieved based on the proposed outlier detection method, which demonstrates that the proposed outlier detection method can further improve the robustness of integrated reconstruction on complex datasets. By comparing the accuracy achieved by the proposed pipeline from Table 8 and the accuracy achieved by the second experiment from Table 9, it can be seen that the proposed pipeline achieved higher reconstruction accuracy on all the datasets. It demonstrates that the proposed outlier detection method can improve the accuracy of integrated reconstruction.

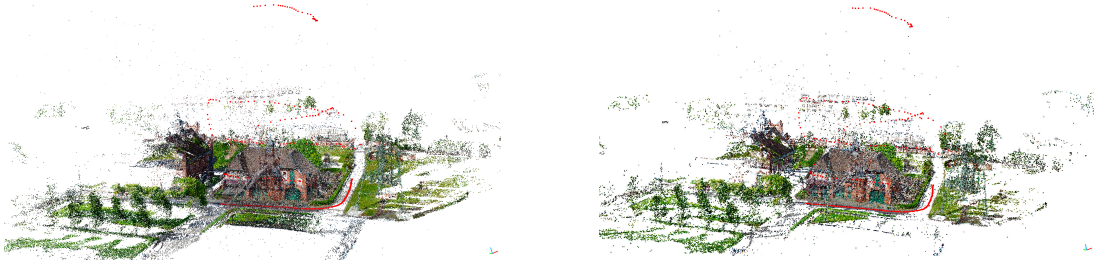
Table 9. Statistics of integrated reconstruction for ablation studies.

Dataset	Experiment	Registered Aerial Images	Registered Terrestrial Images	3D Points	RMSE (Pix)
Center	First	146/146	203/203	1,499,462	0.56
	Second	146/146	203/203	1,005,493	0.61
Zeche	First	147/147	172/172	2,179,790	0.46
	Second	147/147	172/172	1,255,304	0.48
SWJTU-LIB	First	123/123	78/78	1,032,079	0.65
	Second	123/123	0/78	-	-
SWJTU-BLD	First	207/207	0/88	-	-
	Second	207/207	88/88	246,842	0.64
SWJTU-RES	First	-	-	-	-
	Second	-	-	-	-

The experimental results are visualized in Figure 11. The left column shows the sparse models reconstructed in the first experiment, and the sparse models reconstructed in the second experiment are shown on the right. It can be seen from the figures that both experiments achieved geometrically consistent reconstructions on the ISPRS datasets. Figure 11e shows that the sparse models reconstructed from the SWJTU-RES dataset are distorted, and the terrestrial images are misaligned with the aerial images. Figures 10 and 11 together demonstrate the robustness and accuracy of the proposed pipeline for integrated reconstruction.



(**a**)



(**b**)

Figure 11. *Cont.*

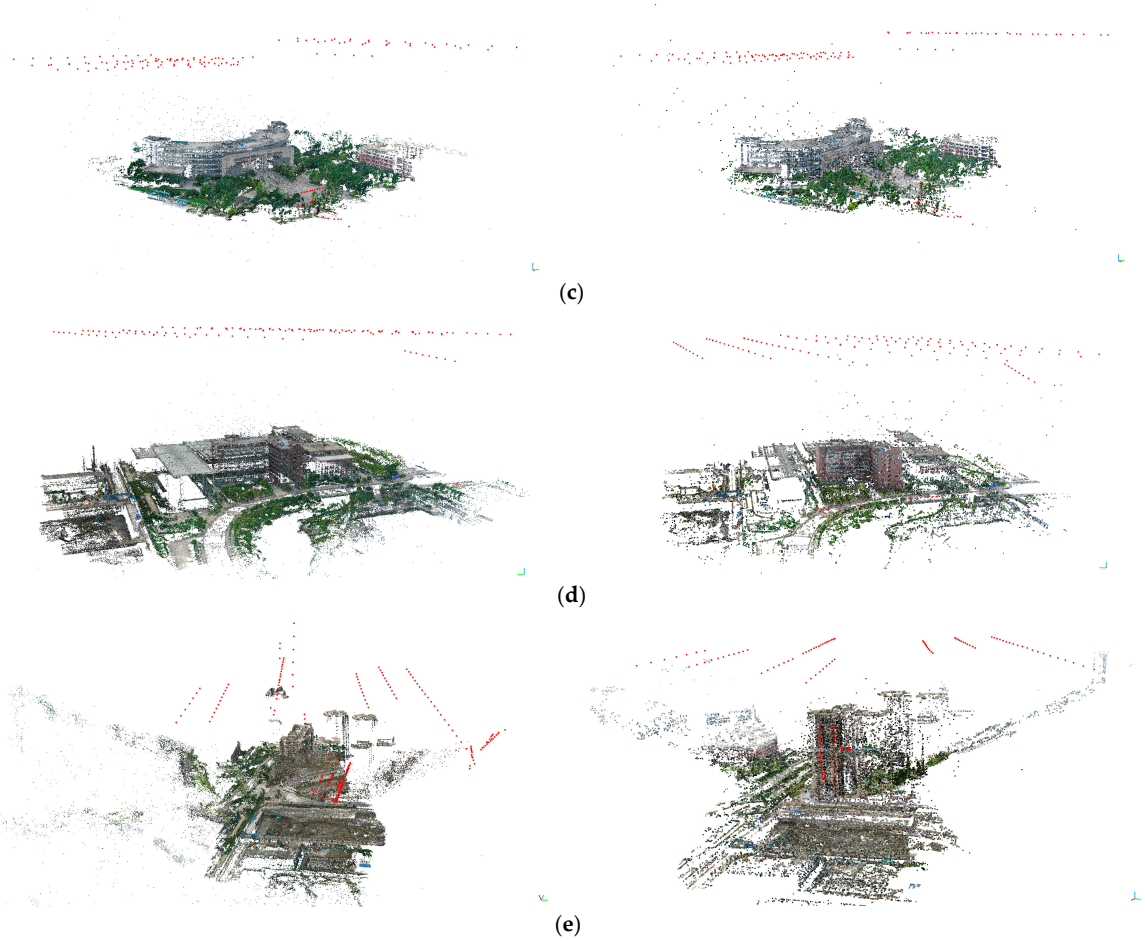


Figure 11. Reconstructed sparse models for ablation studies: (**a**) Center; (**b**) Zeche; (**c**) SWJTU-LIB; (**d**) SWJTU-BLD; (**e**) SWJTU-RES.

5. Discussion

(1) Robustness

There are three factors that affect the robustness of the proposed pipeline. First, the reconstruction of terrestrial and aerial sparse models forms the basis of an integrated reconstruction. The experimental results demonstrate that high-quality terrestrial and aerial sparse models can be obtained using RootSIFT-based image matching and incremental SfM.

Second, the robustness of integrated reconstruction is affected by the precision of match pairs. LoFTR is known to generate tie points even between non-overlapping images. In this case, outliers in the tie points will probably affect the robustness of sparse model fusion and integrated reconstruction. Match pair selection of the proposed methodology is affected by normal vector approximation. The approximation of a normal vector N is based on the assumption that the observation vectors are uniformly distributed in space. Ideally, normal vectors should be estimated using a dense cloud. However, the proposed method avoids using a dense cloud, as dense matching makes the pipeline more time-consuming. Although the approximation may be biased, it still can be used for cross-platform match pair selection by relaxing the angle constraint V_T. In addition, the threshold N_T for the validation of a match pair also influences the match pair selection. When the point density of a terrestrial point cloud is low, this threshold should be lowered to increase the number of match pairs. The values of V_T and N_T are set empirically in this study. The quantitative analysis of the influence of the threshold values on the final results is beyond the scope of the manuscript, and it will be investigated in future work.

Third, the quality of 3D correspondences affects the robustness of integrated reconstruction. As shown by the experimental results, the epipolar constraint cannot remove all mismatches. The remaining outliers will affect the robustness of model fusion. The proposed methodology removes outliers in 3D correspondences, which improves the robustness of integrated reconstruction.

(2) Accuracy

The accuracy of integrated reconstruction is affected by the accuracy of GNSS observations and the quality of tie points and 3D correspondences. First, it is found during the experiments that the elevation accuracy of the terrestrial GNSS observations of the Zeche dataset is low, and therefore, low weights are given to these observations during sparse reconstruction and global optimization. In comparison, the positioning accuracy of the aerial images is generally higher as airborne GNSS observations are not disturbed by ground object occlusion and the multipath effect. Therefore, higher weights are given to aerial GNSS observations during sparse reconstruction and global optimization. Similarly, the proposed methodology merges the terrestrial sparse model with the aerial sparse model rather than the other way around in consideration of the higher accuracy of the aerial GNSS-aided SfM reconstruction. The experimental results demonstrate that the proposed pipeline works as expected with the accuracy of GNSS observations set properly.

Second, the quality of tie points and 3D correspondences also influence the accuracy of the integrated reconstruction. Tie point observations of the proposed methodology are generated by image matching using SIFT and LoFTR. As mentioned above, high-quality tie points can be obtained using SIFT-based image matching during terrestrial and aerial sparse reconstruction. Although mismatches remain in cross-platform tie points, the proposed outlier detection method removes outliers in 3D correspondences to mitigate the influence of the remaining mismatches, which improves the accuracy of the integrated reconstruction.

(3) Efficiency

The proposed pipeline is fully automatic. No human interventions or intermediate processing like cross-view rendering are required, which makes it more streamlined for integrated 3D reconstruction using multi-source images. The current bottleneck of the proposed pipeline is cross-platform image matching due to the low efficiency of the LoFTR implementation, which is much lower than the CPU-parallel RootSIFT implementation from OpenMVG.

6. Conclusions

A novel approach based on robust cross-platform image matching and model fusion is proposed for integrated 3D reconstruction of complex urban scenes using multi-source images. The proposed pipeline works in a hierarchical manner. Firstly, terrestrial and aerial sparse models are reconstructed, respectively. Then, match pairs are selected between terrestrial and aerial images. And cross-platform image matching is performed on the selected match pairs. Thirdly, terrestrial–aerial tie points are triangulated, and outliers in 3D correspondences are filtered. Finally, the terrestrial and aerial sparse models are merged and globally optimized to obtain an integrated model of the scene. The proposed methodology was evaluated on five benchmark datasets. The experimental results demonstrate that terrestrial and aerial images are robustly matched using the proposed cross-platform image matching method. Based on the proposed outlier detection method, terrestrial and aerial models are accurately merged. The proposed pipeline was compared with a state-of-the-art methodology and three software packages. The experimental results demonstrate that the proposed pipeline outperforms the other pipelines in terms of robustness and accuracy. Future work will focus on integrated reconstruction of complex urban scenes using multi-platform and multi-modal remote sensing data.

Author Contributions: Conceptualization, Y.L.; methodology, Y.L.; software, Y.L.; validation, Y.L., Y.M. and Y.Y.; formal analysis, Y.L.; investigation, Y.L.; resources, T.C.; data curation, Y.L.; writing—

original draft preparation, Y.L.; writing—review and editing, Y.L.; visualization, Y.L., Y.M. and Y.Y.; project administration, Y.L.; funding acquisition, Y.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was jointly funded by the Special Foundation for National Science and Technology Basic Research Program of China (grant number 2019FY202500) and the Tianjin Research Innovation Project for Postgraduate Students (grant number 2022SKYZ269).

Data Availability Statement: Data are contained within the article.

Acknowledgments: The authors would like to acknowledge the provision of the datasets by ISPRS and EuroSDR, released in conjunction with the ISPRS scientific initiative 2014 and 2015, led by ISPRS ICWG II/Ia. The authors gratefully acknowledge the provision of the datasets by the research team from Southwest Jiaotong University.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Kalfarisi, R.; Wu, Z.Y.; Soh, K. Crack detection and segmentation using deep learning with 3D reality mesh model for quantitative assessment and integrated visualization. *J. Comput. Civ. Eng.* **2020**, *34*, 04020010. [CrossRef]
2. Lehtola, V.V.; Koeva, M.; Elberink, S.O.; Raposo, P.; Virtanen, J.-P.; Vahdatikhaki, F.; Borsci, S. Digital twin of a city: Review of technology serving city needs. *Int. J. Appl. Earth Obs. Geoinf.* **2022**, *114*, 102915. [CrossRef]
3. Li, X.; Yang, B.; Liang, F.; Zhang, H.; Xu, Y.; Dong, Z. Modeling urban canopy air temperature at city-block scale based on urban 3D morphology parameters—A study in Tianjin, North China. *Build. Environ.* **2023**, *230*, 110000. [CrossRef]
4. Colomina, I.; Molina, P. Unmanned aerial systems for photogrammetry and remote sensing: A review. *ISPRS J. Photogramm. Remote Sens.* **2014**, *92*, 79–97. [CrossRef]
5. Jiang, S.; Jiang, W.; Wang, L. Unmanned Aerial Vehicle-Based Photogrammetric 3D Mapping: A survey of techniques, applications, and challenges. *IEEE Geosci. Remote Sens. Mag.* **2022**, *10*, 135–171. [CrossRef]
6. Nex, F.; Remondino, F. UAV for 3D mapping applications: A review. *Appl. Geomat.* **2014**, *6*, 1–15. [CrossRef]
7. Shan, J.; Li, Z.; Lercel, D.; Tissue, K.; Hupy, J.; Carpenter, J. Democratizing photogrammetry: An accuracy perspective. *Geo-Spat. Inf. Sci.* **2023**, *26*, 175–188. [CrossRef]
8. Li, Z.; Wu, B.; Li, Y.; Chen, Z. Fusion of aerial, MMS and backpack images and point clouds for optimized 3D mapping in urban areas. *ISPRS J. Photogramm. Remote Sens.* **2023**, *202*, 463–478. [CrossRef]
9. Zhu, Q.; Wang, Z.; Hu, H.; Xie, L.; Ge, X.; Zhang, Y. Leveraging photogrammetric mesh models for aerial-ground feature point matching toward integrated 3D reconstruction. *ISPRS J. Photogramm. Remote Sens.* **2020**, *166*, 26–40. [CrossRef]
10. Nex, F.; Gerke, M.; Remondino, F.; Przybilla, H.-J.; Bäumker, M.; Zurhorst, A. ISPRS benchmark for multi-platform photogrammetry. *ISPRS Ann. Photogramm. Remote Sens. Spat. Inf. Sci.* **2015**, *2*, 135–142. [CrossRef]
11. Gerke, M.; Nex, F.; Jende, P. Co-registration of terrestrial and UAV-based images–experimental results. *Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci.* **2016**, *40*, 11–18. [CrossRef]
12. Mikolajczyk, K.; Schmid, C. A performance evaluation of local descriptors. *IEEE Trans. Pattern Anal. Mach. Intell.* **2005**, *27*, 1615–1630. [CrossRef] [PubMed]
13. Yi, K.M.; Trulls, E.; Lepetit, V.; Fua, P. Lift: Learned invariant feature transform. In Proceedings of the Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, 11–14 October 2016; pp. 467–483.
14. Dusmanu, M.; Rocco, I.; Pajdla, T.; Pollefeys, M.; Sivic, J.; Torii, A.; Sattler, T. D2-net: A trainable cnn for joint description and detection of local features. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 8092–8101.
15. Gao, X.; Shen, S.; Zhou, Y.; Cui, H.; Zhu, L.; Hu, Z. Ancient Chinese architecture 3D preservation by merging ground and aerial point clouds. *ISPRS J. Photogramm. Remote Sens.* **2018**, *143*, 72–84. [CrossRef]
16. Wu, B.; Xie, L.; Hu, H.; Zhu, Q.; Yau, E. Integration of aerial oblique imagery and terrestrial imagery for optimized 3D modeling in urban areas. *ISPRS J. Photogramm. Remote Sens.* **2018**, *139*, 119–132. [CrossRef]
17. Ling, X.; Qin, R. A graph-matching approach for cross-view registration of over-view and street-view based point clouds. *ISPRS J. Photogramm. Remote Sens.* **2022**, *185*, 2–15. [CrossRef]
18. Yang, B.; Zang, Y.; Dong, Z.; Huang, R. An automated method to register airborne and terrestrial laser scanning point clouds. *ISPRS J. Photogramm. Remote Sens.* **2015**, *109*, 62–76. [CrossRef]
19. Arya, S.; Mount, D.M.; Netanyahu, N.S.; Silverman, R.; Wu, A.Y. An optimal algorithm for approximate nearest neighbor searching fixed dimensions. *J. ACM* **1998**, *45*, 891–923. [CrossRef]
20. Lowe, D.G. Distinctive Image features from scale-invariant keypoints. *Int. J. Comput. Vis.* **2004**, *60*, 91–110. [CrossRef]
21. Fischler, M.A.; Bolles, R.C. Random Sample Consensus: A Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography. *Commun. ACM* **1981**, *24*, 381–395. [CrossRef]
22. Hartley, R.; Zisserman, A. *Multiple View Geometry in Computer Vision*, 2nd ed.; Cambridge University Press: Cambridge, UK, 2004.

23. Agarwal, S.; Furukawa, Y.; Snavely, N.; Simon, I.; Curless, B.; Seitz, S.M.; Szeliski, R. Building rome in a day. *Commun. ACM* **2011**, *54*, 105–112. [CrossRef]
24. Jiang, S.; Jiang, C.; Jiang, W. Efficient structure from motion for large-scale UAV images: A review and a comparison of SfM tools. *ISPRS J. Photogramm. Remote Sens.* **2020**, *167*, 230–251. [CrossRef]
25. Snavely, N.; Seitz, S.M.; Szeliski, R. Photo tourism: Exploring photo collections in 3D. *ACM Trans. Graph.* **2006**, *25*, 835–846. [CrossRef]
26. Snavely, N.; Seitz, S.M.; Szeliski, R. Modeling the world from internet photo collections. *Int. J. Comput. Vis.* **2008**, *80*, 189–210. [CrossRef]
27. Furukawa, Y.; Ponce, J. Accurate, dense, and robust multiview stereopsis. *IEEE Trans. Pattern Anal. Mach. Intell.* **2010**, *32*, 1362–1376. [CrossRef]
28. Hirschmüller, H. Stereo Processing by Semiglobal Matching and Mutual Information. *IEEE Trans. Pattern Anal. Mach. Intell.* **2008**, *30*, 328–341. [CrossRef] [PubMed]
29. Schönberger, J.L.; Zheng, E.; Frahm, J.-M.; Pollefeys, M. Pixelwise view selection for unstructured multi-view stereo. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 8–16 October 2016; pp. 501–518.
30. Kazhdan, M.; Hoppe, H. Screened poisson surface reconstruction. *ACM Trans. Graph.* **2013**, *32*, 1–13. [CrossRef]
31. Morel, J.-M.; Yu, G. ASIFT: A new framework for fully affine invariant image comparison. *SIAM J. Imaging Sci.* **2009**, *2*, 438–469. [CrossRef]
32. Hu, H.; Zhu, Q.; Du, Z.; Zhang, Y.; Ding, Y. Reliable Spatial Relationship Constrained Feature Point Matching of Oblique Aerial Images. *Photogramm. Eng. Remote Sens.* **2015**, *81*, 49–58. [CrossRef]
33. Liu, J.; Yin, H.; Liu, B.; Lu, P. Tie Point Matching between Terrestrial and Aerial Images Based on Patch Variational Refinement. *Remote Sens.* **2023**, *15*, 968. [CrossRef]
34. DeTone, D.; Malisiewicz, T.; Rabinovich, A. Superpoint: Self-supervised interest point detection and description. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Salt Lake City, UT, USA, 18–22 June 2018; pp. 224–236.
35. Sun, J.; Shen, Z.; Wang, Y.; Bao, H.; Zhou, X. LoFTR: Detector-free local feature matching with transformers. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 8922–8931.
36. Jin, Y.; Mishkin, D.; Mishchuk, A.; Matas, J.; Fua, P.; Yi, K.M.; Trulls, E. Image Matching Across Wide Baselines: From Paper to Practice. *Int. J. Comput. Vision.* **2021**, *129*, 517–547. [CrossRef]
37. Li, H.; Liu, A.; Xie, X.; Guo, H.; Xiong, H.; Zheng, X. Learning Dense Consistent Features for Aerial-to-Ground Structure-From-Motion. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2023**, *16*, 5089–5102. [CrossRef]
38. Sarlin, P.-E.; DeTone, D.; Malisiewicz, T.; Rabinovich, A. Superglue: Learning feature matching with graph neural networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 4938–4947.
39. Bellavia, F.; Colombo, C.; Morelli, L.; Remondino, F. Challenges in image matching for cultural heritage: An overview and perspective. In Proceedings of the International Conference on Image Analysis and Processing, Lecce, Italy, 23–27 May 2022; pp. 210–222.
40. Ji, S.; Zeng, C.; Zhang, Y.; Duan, Y. An evaluation of conventional and deep learning-based image-matching methods on diverse datasets. *Photogramm. Rec.* **2023**, *38*, 137–159. [CrossRef]
41. Jiang, S.; Jiang, W.; Guo, B.; Li, L.; Wang, L. Learned Local Features for Structure from Motion of UAV Images: A Comparative Evaluation. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2021**, *14*, 10583–10597. [CrossRef]
42. Schönberger, J.L.; Hardmeier, H.; Sattler, T.; Pollefeys, M. Comparative evaluation of hand-crafted and learned local features. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 14–19 June 2020; pp. 1482–1491.
43. Arandjelović, R.; Zisserman, A. Three things everyone should know to improve object retrieval. In Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, 16–21 June 2012; pp. 2911–2918.
44. Moulon, P.; Monasse, P.; Perrot, R.; Marlet, R. OpenMVG: Open Multiple View Geometry. In Proceedings of the International Workshop on Reproducible Research in Pattern Recognition, Cancún, Mexico, 4 December 2016; pp. 60–74.
45. Riba, E.; Mishkin, D.; Ponsa, D.; Rublee, E.; Bradski, G. Kornia: An open source differentiable computer vision library for pytorch. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Snowmass Village, CO, USA, 1–5 March 2020; pp. 3674–3683.
46. Li, Z.; Snavely, N. Megadepth: Learning single-view depth prediction from internet photos. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 2041–2050.
47. Schönberger, J.L.; Frahm, J.-M. Structure-from-motion revisited. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, VA, USA, 27–30 June 2016; pp. 4104–4113.