ISPRS Journal of Photogrammetry and Remote Sensing 202 (2023) 463–478

Contents lists available at ScienceDirect

ISPRS Journal of Photogrammetry and Remote Sensing

journal homepage: www.elsevier.com/locate/isprsjprs

Fusion of aerial, MMS and backpack images and point clouds for optimized 3D mapping in urban areas

Zhaojin Li [a], Bo Wu [a,*], Yuan Li [b], Zeyu Chen [a]

[a] Department of Land Surveying & Geo-Informatics, The Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong, China
[b] School of Geospatial Engineering and Science, Sun Yat-sen University, Zhuhai, PR China

ARTICLE INFO

Keywords:
Aerial oblique imagery
Mobile mapping system (MMS)
Backpack
3D mapping

ABSTRACT

Photorealistic 3D models are important data sources for digital twin cities and smart city applications. These models are usually generated from data collected by aerial or ground-based platforms (e.g., mobile mapping systems (MMSs) and backpack systems) separately. Aerial and ground-based platforms capture data from overhead and ground surfaces, respectively, offering complementary information for better 3D mapping in urban areas. Particularly, backpack mapping systems have gained popularity for 3D mapping in urban areas in recent years, as they offer more flexibility to reach regions (e.g., narrow alleys and pedestrian routes) inaccessible by vehicle-based MMSs. However, integration of aerial and ground data for 3D mapping suffers from difficulties such as tie-point matching among images from different platforms with large differences in perspective, coverage, and scale. Optimal fusion of the results from different platforms is also challenging. Therefore, this paper presents a novel method for the fusion of aerial, MMS, and backpack images and point clouds for optimized 3D mapping in urban areas. A geometric-aware model for feature matching is developed based on the SuperGlue algorithm to obtain sufficient tie-points between aerial and ground images, which facilitates the integrated bundle adjustment of images to reduce their geometric inconsistencies and the subsequent dense image matching to generate 3D point clouds from different image sources. After that, a graph-based method considering both geometric and texture traits is developed for the optimal fusion of point clouds from different sources to generate 3D mesh models of better quality. Experiments conducted on a challenging dataset in Hong Kong demonstrated that the geometric-aware model could obtain sufficient accurately matched tie-points among the aerial, MMS, and backpack images, which enabled the integrated bundle adjustment of the three image datasets to generate properly aligned point clouds. Compared with the results obtained from state-of-the-art commercial software, the 3D mesh models generated from the proposed point cloud fusion method exhibited better quality in terms of completeness, consistency, and level of detail.

1. Introduction

Photorealistic 3D models, generated from images collected either from aerial or ground platforms, are the main data sources for digital twin cities (Lehtola et al., 2022) and smart city applications (White et al., 2021). Aerial oblique photogrammetry can retrieve large-scale scenes and has therefore been widely utilized for city-scale 3D mapping. However, the information on near-ground objects is often not fully retrieved, particularly in high-rise metropolitan areas (Ye and Wu, 2018). In contrast, terrestrial images obtained mainly by vehicle-borne mobile mapping systems (MMSs) can provide supplementary views of close-range objects on the ground. Therefore, aerial oblique and MMS images have increasingly been united to achieve improved 3D models (Gao et al., 2018; Wu et al., 2018; Zhu et al., 2020).

However, blind zones limited by vehicle accessibility (e.g., alleys between tall buildings and sidewalks) remain unreachable by vehicle-borne MMSs, which results in a substantial loss of 3D information in urban areas. Attempts have been made to leverage the flexible backpack mapping system to retrieve closer-range observations (Fassi and Perfetti, 2019; Li et al., 2020), but this complicates the already difficult problem of matching and connecting aerial and ground images in the structure from motion (SfM) procedure (Agarwal et al., 2009). Despite the nearly 180° viewing angle difference, the coverage of backpack images is extremely small compared with aerial images. The matching issue may

* Corresponding author.
E-mail address: bo.wu@polyu.edu.hk (B. Wu).

https://doi.org/10.1016/j.isprsjprs.2023.07.010
Received 15 November 2022; Received in revised form 7 July 2023; Accepted 7 July 2023
Available online 12 July 2023

be further exacerbated by different illumination conditions introduced by viewing angle and large resolution discrepancies arising from shooting distance. The scale invariant feature transform (SIFT) (Lowe, 2004), the *de facto* standard of feature extraction and matching, has been proven to fail in such challenging scenarios (Mikolajczyk et al., 2005; Zhu et al., 2020). 3D information obtained from the onboard GPS/IMU system is therefore exploited to align images to the same view (Shan et al., 2014; Wu et al., 2018; Zhu et al., 2020) so that SIFT matching can successfully function. This type of algorithm can effectively address the viewing difference problem for some simple datasets, but it remains incapable of matching more challenging datasets. Recent advances in deep learning, e.g., the SuperGlue algorithm (Sarlin et al., 2020), allow to construct more distinct and robust feature descriptors and solve the matching problem through a convolutional neural network. These learning-based algorithms outperform the traditional methods in many challenging tasks; however they still struggle with the feature matching problem for unordered images collected from different platforms with large differences in perspective, coverage and scale.

Another major obstacle in urban 3D mapping is the strategy to integrate 3D data (e.g., point clouds) from different sources in an optimal way so that to generate 3D models of better quality. Abundant research has focused on registering misaligned point clouds from multiple sources (Huang et al., 2023; Li and Harada, 2022; Mei et al., 2023; Monji-Azad et al., 2023), which could benefit the fusion of point clouds to enrich the 3D scene (Cui et al., 2022). However, the registered point clouds may still suffer from inconsistencies in local regions and unnecessary redundancy in overlapping regions. Simply merging point clouds can lead to problems in geometry and texture quality (Gao et al., 2018). Wu et al. (2018) pioneered the optimal fusion idea by choosing the appropriate images according to the viewing conditions to generate 3D mesh models of better quality, which inspired the optimal fusion of point clouds in this study.

Integration of complementary datasets obtained from different platforms is important for optimized 3D mapping to generate 3D models of better quality in terms of accuracy, completeness, consistency, and level of detail. This paper presents a novel method for the fusion of aerial, MMS, and backpack images and point clouds for optimized 3D mapping in urban areas. The main contributions of this paper are as follows:

(1) A geometric-aware learning-based algorithm is proposed to tackle the feature matching problem in SfM to connect the unordered images from different platforms. Rather than directly feeding the network with the original images (Agarwal et al., 2009), the exterior orientation (EO) parameters of images are used to guide the learning-based feature matching to improve the accuracy and reduce computational complexity. Building façades visible on both the aerial and ground images are segmented and used as reference planes to mitigate the large variations between images from different perspectives so that to generate tentative matching tracks from the unordered images. In addition to the feature encoding method based on 2D image information such as the one used in the SuperGlue algorithm (Sarlin et al., 2020), the 3D geometric information (e.g., the 3D coordinates of feature points calculated from the initial image EO parameters) are encoded in the feature descriptor, so that the relationships among the features can be used to better assist feature matching. Possible wrong matches are further filtered out by multiple geometric constraints including the reference plane, multi-view image geometry, and epipolar geometry.
(2) A graph-based method is proposed for the optimal fusion of point clouds obtained from different platforms. Supervoxels representing clusters of points are segmented from the point clouds, and multi-dimensional features associated with each supervoxel are derived and used for the optimal fusion of point clouds based on the graph-cut method (Boykov et al., 2001) operating on a voxel grid. As both the geometric and textural information are exploited in the determination of the multi-dimensional features for each supervoxel, the method allows the optimal selection and fusion of point clouds to generate 3D models with better quality.

The remainder of this paper is organized as follows. Section 2 reviews the previous related work. Sections 3 elaborates the methods including integrated bundle adjustment of aerial, MMS, and backpack images based on effective tie point matching through the geometric-aware learning-based algorithm and the optimal fusion of point clouds from different sources for better 3D modelling. Section 4 presents the experimental results using representative datasets in Hong Kong and evaluates the performance of the methods. Section 5 discusses the results and limitations of the methods. Finally, the concluding remarks are summarized in Section 6.

2. Related work

To generate 3D models from unordered images, the most representative method is SfM (Agarwal et al., 2009; Schonberger and Frahm, 2016). SfM comprises four consecutive steps, namely, tie-point matching, bundle adjustment, dense matching, and point cloud generation. Among them, tie-point matching to find corresponding points across different images is the critical step (Wu et al., 2018; Zhu et al., 2020), which facilitates the subsequent bundle adjustment of images to eliminate possible geometric inconsistencies among them. Tie-point matching usually extracts distinctive feature points (Förstner and Gülch, 1987; Harris and Stephens, 1988; Wu et al., 2012) for matching, as represented by the landmark SIFT algorithm (Lowe, 2004). Despite considering the distinctiveness of grayscale across the multi-resolution image pyramid, SIFT defines the main direction of each feature, which allows it to match most image pairs with scale, rotation, and translation transformations. Morel and Yu (2009) extended the SIFT algorithm to ASIFT by generating simulated images based on a series of camera positions and rotations, and conducting SIFT on all of the simulated images for the matching pair. ASIFT not only handles the affine transformation but also greatly increases the matching numbers of each image pair.

Many previous works exploited the view-dependent idea by using the position and pointing information of images to mitigate large perspective differences between images to facilitate their matching. Wu et al. (2018) utilized the initial exterior orientation parameters to align each visible aerial and ground image according to the visible building façade, followed by conventional SIFT matching. Similarly, Gao et al. (2018) rectified each ground image relative to the target aerial view. Zhu et al. (2020) extended this idea by rendering the mesh model generated from the aerial images according to the perspective of each ground image, and matching the rendered aerial image and target ground image, before propagating the matches to the original images. Although this mesh-render algorithm reduces computation complexity, rendering city-scale scenes still demands huge computational resources and time. Furthermore, deviations in terms of coverage and spatial resolution of aerial and ground images were hardly discussed.

Nonetheless, these algorithms remain heavily dependent on the SIFT algorithm to extract features and some handcrafted descriptors for matching, which limits the fusion of aerial and ground images for 3D modelling in challenging urban areas (Sarlin et al., 2020). The recent development of deep-learning methods allows learnable features (Sun et al., 2021) to automatically extract deep features, which has surpassed traditional SIFT-like features in many tasks. SuperGlue (Sarlin et al., 2020), the milestone of the deep learning matching algorithm, follows the conventional matching pipeline but subverts manually defined feature descriptors construction and comparison criteria with learned ones. Although this end-to-end strategy works for most close-range tasks, the large differences in coverage, resolution, and perspective in the aerial-ground scenario complicate the algorithm and result in unfavorable matching performances.

Since deep learning is inherently a data-driven method, the training dataset is of great importance for the success of the algorithm (Paullada et al., 2021; Sun et al., 2017). For feature matching tasks, the widely used training datasets are MegaDepth (Li and Snavely, 2018), Phototourism (Jin et al., 2021), and ScanNet (Dai et al., 2017). However, majority of the images in these datasets are close-range images taken from the ground view. For feature matching between images collected from different platforms (e.g., aerial and ground) with large differences in perspective, coverage and scale, a versatile dataset comprising images collected from different platforms of similar characteristics is needed to train the deep-learning model.

Besides, after eliminating geometric inconsistencies among different datasets, data can be simply integrated by treating images from different platforms equally and performing multi-view stereo (MVS) for generating 3D point clouds and mesh models, which is a common solution adopted by much commercial software (e.g., Photoscan, ContextCapture). By increasing the number of observations, the defects of each individual point cloud (i.e., noises, ghost data, holes) are mitigated to a certain extent (Kim et al., 2021; Wolff et al., 2016; Yao et al., 2018). However, this simple union process of multi-source point clouds results in unnecessary data redundancy in the final product, and inevitably causes noise and artifacts. Point cloud filtering is thus important for selecting proper points and removing noise in the point clouds. Many studies have been conducted to denoise point clouds (Rusu and Cousins, 2011) based on the statistics of neighboring points (Balta et al., 2018) or signal processing theory (Digne and de Franchis, 2017). Some recent studies in point cloud processing have leveraged the supervoxel-based algorithm to replace point-based ones due to the high homogeneity of neighbor points (Li et al., 2022; Zhu et al., 2017).

In summary, using geometric information to assist image matching has shown potential in mitigating large perspective differences between image pairs, but the inherent drawbacks of hand-crafted features impede its use in more challenging tasks. While the learning-based algorithm simply feeds the network with the images without geometric information, large-scale aerial images still cannot be favorably matched with localized ground images. It is desirable to develop innovative method for effective matching of aerial and ground images. Furthermore, few studies in the past have focused on the optimal fusion of multi-source point clouds to achieve the highest-quality 3D modelling, which requires further systematic research.

3. Methods

3.1. Tie-point matching and bundle adjustment of Aerial, MMS, and backpack images

Fig. 1 shows the overall workflow of tie-point matching and bundle adjustment of aerial, MMS, and backpack images for 3D point cloud

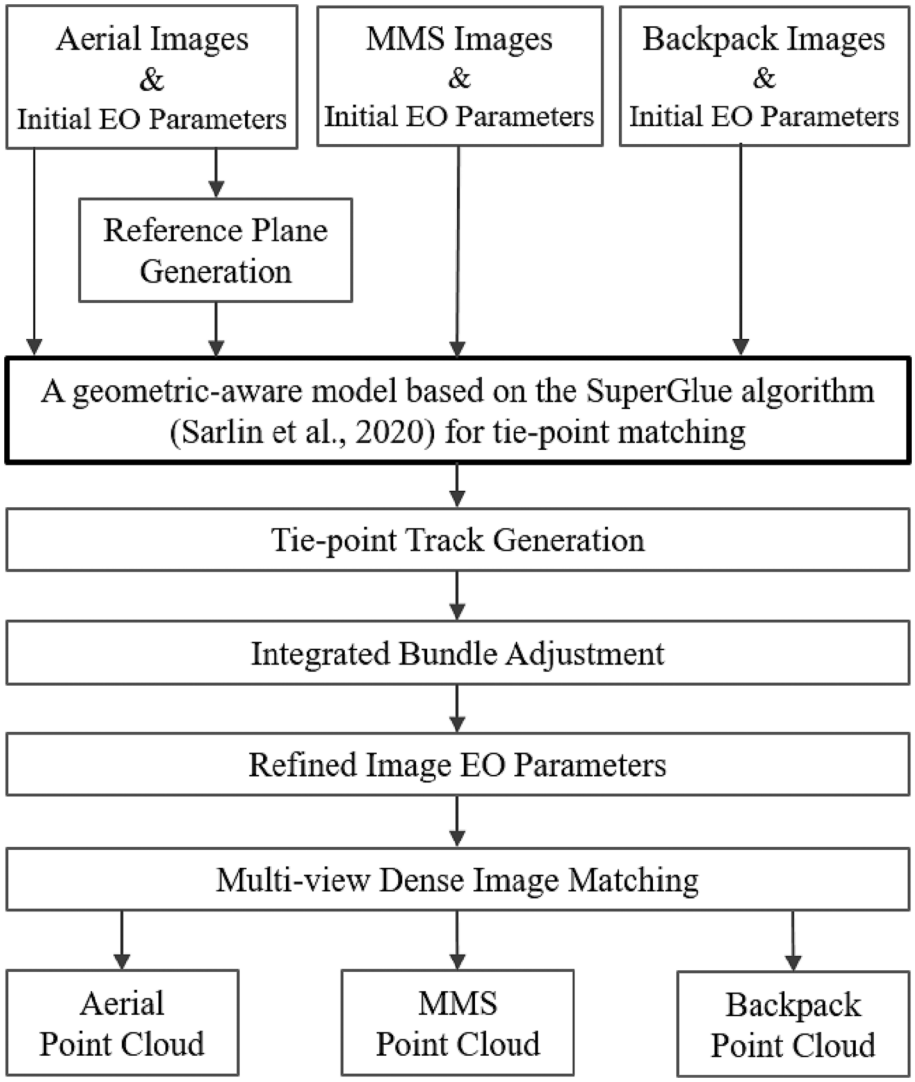


Fig. 1. Workflow of tie-point matching and bundle adjustment of aerial, MMS, and backpack images for point cloud generation.

generation. Aerial images are firstly processed to generate a sparse point cloud, which serves as the input for reference plane (e.g., building façade) retrieval (Fischler and Bolles, 1981). The large unordered image dataset can then be organized according to the visibility of the reference planes. Then, the geometric-aware learning-based algorithm is performed for tie-point matching between aerial and ground images to obtain robust tie points. The pair-wise matches are subsequently integrated to form tie-point tracks (He et al., 2017). Based on the matched tie points and tracks, integrated bundle adjustment is performed to eliminate geometric inconsistencies among the multi-source images by adjusting their exterior orientation (EO) parameters, and between the nominal EOs and the reference datum by using several ground control points (GCPs). Finally, MVS is conducted for aerial, MMS, and backpack images individually, resulting in well aligned 3D point clouds.

3.1.1. A Geometric-aware Learning-based algorithm for tie point matching

Although previous studies have succeeded in matching aerial and ground images with SIFT-like features (Gao et al., 2018; Wu et al., 2018; Zhu et al., 2020), experiments have verified their unfeasibility in densely raised urban areas. There are two reasons for this. First, few SIFT features that present the same point in the real world can be extracted from aerial and ground images simultaneously. Second, hand-crafted descriptors constructed only from image's grayscale information are likely disturbed by the illumination and resolution differences. The deep-learning algorithm SuperGlue (Sarlin et al., 2020; Sun et al., 2021) was hence introduced to distinctively and robustly construct features. Rather than simply using the 256-dimension appearance descriptor extracted from the convolutional neural network (CNN) backbone (DeTone et al., 2018), SuperGlue also incorporates the 2D coordinates *(u,v)* of feature points into the descriptor. Assisted with the attention mechanism (Sarlin et al., 2020) used in the transformer deep-learning model, the self-attention and cross-attention are constructed to describe the relationship between features in one image and across the image pair, respectively, which are embedded in the feature descriptor.

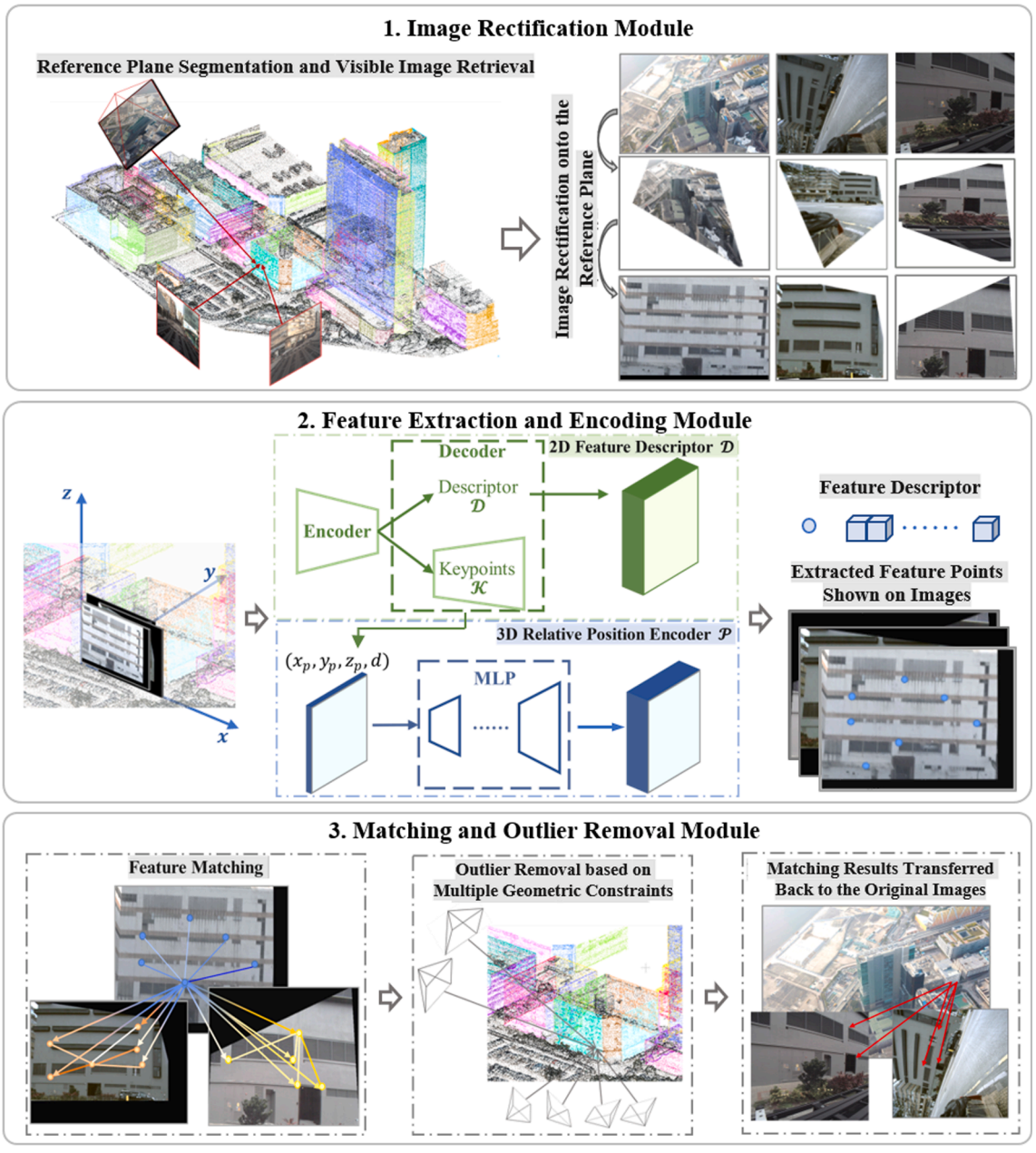


Fig. 2. Illustration of the geometric-aware learning-based algorithm for tie-point matching.

Finally, the one-to-one match selection problem is formulated as an optimal partial assignment based on the Sinkhorn algorithm (Peyre and Cuturi, 2019), and the feature pairs with the highest score are retrieved as the matched tie points.

Leveraging the deep-learning algorithm, feature descriptors can be constructed more comprehensively, making them more resilient to illumination and brightness variations. However, the large variations in coverage, perspective, and scale still present challenges to matching aerial and ground images. Inspired by Wu et al. (2018), a geometric-aware model based on the SuperGlue algorithm is developed, of which geometric information has been exploited to assist the tie-point matching throughout the matching process including:

- The building façades visible on both the aerial and ground view images are firstly segmented from the initial point cloud generated from aerial images using their initial EO parameters, which are then used as reference planes to facilitate the generation of tentative matching tracks from the unordered image dataset. The retrieved images are projected to the reference planes for tie-point matching, so that to mitigate the large variations between the aerial and ground view images.
- In addition to the regular feature encoding method based on 2D image information, the 3D geometric information (e.g., the 3D coordinates of feature points calculated from the initial EO parameters of images) are encoded in the feature descriptor, so that the relationships among the features (self-attention and cross-attention mechanism in the SuperGlue algorithm) can be used to better assist tie-point matching.
- After obtaining the initial matches, multiple geometric constraints including the reference plane, multi-view image geometry, and epipolar geometry are used to filter out possible wrong matches to guarantee the reliability of the final matched tie points.

Fig. 2 illustrates the geometric-aware model for tie-point matching using an aerial image and a ground image covering the same reference plane as an example. First, the normal of each image is calculated from the associated pose information (e.g., from the IMU measurements or EO parameters of images), while the perspective transformation matrix $\boldsymbol{P}$ projecting the image to the reference plane can be derived with the known normal of the plane. Second, the image area outside the plane is excluded to surmount the large coverage variation and the remaining image segments are normalized to the same scale to mitigate the resolution discrepancy, which can be mathematically regarded as a translation $\boldsymbol{T}$ and a scale operation s. After this step, it is explicitly assumed that the 256-dimension descriptors between the two images are aligned as much as possible.

Other than the deep features, the positional encoder (Chu et al., 2021; Zheng et al., 2021) is a crucial part of the original SuperGlue algorithm due to its ability to record structures and distinguish low-texture or repeated pattern regions (Sun et al., 2021). Rather than using image coordinates directly, the 3D relative positions of features are encoded in the positional encoder. 3D relative position is defined as the normalized distance of a point to the center of a plane in the X, Y, and Z directions, which is then extended with the multiple layer perceptron (MLP) structure to a learnable 256-dimension positional encoder. Superior to 2D coordinates, the relative 3D position is resilient to differences in viewing direction, which grants the ability to distinguish the occlusion and repeated region. Unlike simply adding the visual and position encoders (Sarlin et al., 2020), a fully connected layer and a ReLU layer (Nair and Hinton, 2010) are introduced in this architecture to balance the influence between the texture descriptor $\mathscr{D}$ and the position encoder $\mathscr{P}$ through the learnable parameters This process implicitly assumes that the importance of appearance and position should not be identical, and that the assigned weights assist the network to focus on the more important information in different scenarios.

The original SuperGlue algorithm has no outlier removal module, while incorrect matches may occur in the putative matches. Therefore, to guarantee matching quality for the subsequent bundle adjustment, three criteria must be satisfied: epipolar geometry, plane geometry, and intersection constraints, to remove possible mismatches. Epipolar geometry, shown in Equation (1), explicitly assumes that exact consistency has not yet been achieved, and the distance d_{x_1} from point $\boldsymbol{x}_1$ to the epipolar line as calculated with the corresponding point $\boldsymbol{x}_2$ on the other image should be within a reasonable threshold T.

$$d_{x_1} = \boldsymbol{x}_1{}^{*}\frac{F_{21}{}^{*}\boldsymbol{x}_2}{|F_{21}{}^{*}\boldsymbol{x}_2|} < T \tag{1}$$

$$F_{21} = K_1 R\left[R^T t\right]_{\times} K_2^{-1} \tag{2}$$

where F_{21} in Equation (2) denotes the fundamental matrix between two images calculated from the relative rotation matrix R, translation matrix t, and camera matrixes K_1 and K_2. Furthermore, the 3D coordinate of each point is calculated from each matching pair. The incorrect match is rejected, whose 3D coordinate is not near the plane to satisfy the geometry constraint, or the calculated coordinate exceeds 3σ compared with the corresponding point on the other images. Finally, these matched features are reverted to the original image using P, T, and s.

Pair-wise tie-points are extended to long feature tracks to link more images, and short feature tracks involving few images are excluded as features may not be distinct and long feature tracks may be more robust against outliers.

3.1.2. Transfer-Learning for Geometric-Aware SuperGlue

An image dataset including aerial and ground images collected at a typical urban region in Hong Kong (Li et al., 2020) was used to construct the transfer-learning dataset for geometric-aware SuperGlue training. Tie points between the aerial and ground images were manually identified by several operators. Bundle adjustment of the images was carried out to remove tie points with large residuals. A final manual check was conducted for all the images and tie points, and only those tie points across a sufficient number of images (e.g., more than four) were chosen to build the training dataset.

After processing the images into geometric-aware ones, as mentioned in Section 3.1, SuperPoint features (DeTone et al., 2018) are extracted from each image. The features are then matched with all images obtained from the same platform to retrieve the 3D position of each 2D feature point, through multi-view space intersection to facilitate more precise 3D coordinate retrieval. Then, for each aerial-ground image pair, 2D feature points $\boldsymbol{x}_1$ on aerial images are projected to point $\boldsymbol{x}_2^{'}$ on the ground image using the 3D position. If feature point $\boldsymbol{x}_2$ exists in the vicinity of $\boldsymbol{x}_2^{'}$ (e.g., within 1 pixel), $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ are retrieved as tie points.

As tie-points are retrieved strictly from 3D information, some correct tie-points retrieved by the network may be missed by the dataset. The loss function is therefore constructed based on negative log-likelihood loss, considering the supervised $\mathscr{L}_{supervisedmatch}$ and $\mathscr{L}_{supervisedunmatch}$, and the unsupervised part $\mathscr{L}_{unsupervisedmatch}$, as:

$$\begin{aligned}\mathscr{L}_{overall} &= \mathscr{L}_{supervised\,match} + \mathscr{L}_{supervised\,unmatch} + w\mathscr{L}_{unsupervised\,match} \\ &= -\sum_{(i,j)\in\mathscr{M}} \log \mathscr{S}_{ij} - \left(\sum_{i\in\mathscr{M}} \log \mathscr{S}_{i^*} + \sum_{j\in\mathscr{M}} \log \mathscr{S}_{*j}\right) - w\sum_{(i,j)\notin\mathscr{M}} \log\left(1-\mathscr{S}_{ij}\right)\end{aligned} \tag{3}$$

where w is initially assigned as zero and increased gradually to allow the network to first improve the accuracy and then the recall of the matches. $\mathscr{S}_{ij}$ signifies the confidence score if the keypoint i in the first image corresponds to the keypoint j in the second image, $\mathscr{S}_{i^*}$ represents the supervised tie-point when the i^{th} keypoint does not find a match in its image pair, and $\mathscr{M}$ denotes the supervised matched feature pairs.

3.1.3. Integrated bundle adjustment

To eliminate all possible inconsistencies among images from

different platforms, an integrated bundle adjustment of the images is constructed with four types of observation functions, as shown in Equation (4). The first two equations constrain the intersection of the bundles of matches for inner-platform and cross-platform images, respectively. As matching is conducted on a processed image, the operator $f(\hat{\mathrm{A}}\cdot)$ converts the k^{th} match track on the i^{th} image tile to the original image, and $\Pi_i(K_i, X_k)$ projects the unknown 3D coordinate X_k to the original image. Notably, although geometric-aware learned matching retrieves abundant tie-points between the aerial and ground images, the number of tie-points remains incomparable with that of the inner platform. Larger weight w_{cross} is therefore assigned to the cross-platform tie-points to further balance the contribution of different types of tie-points for robust optimization. As the nominal EO parameters of images are roughly correct, the third equation implicitly assumes that the optimized EO parameters should not greatly deviate from the nominals. Optionally, a series of ground control points (GCPs), which are digitized manually on images, can also be used to align the optimized EO parameters to specific spatial references as presented in the forth equation.

$$\begin{cases} \mathscr{L}_{inner} = w_{inner}\sum_{k_{inner}}\sum_{i}\left\|\Pi_i(K_i,X_k) - f\left(\boldsymbol{x}_i^k\right)\right\| \\ \mathscr{L}_{cross} = w_{cross}\sum_{k_{cross}}\sum_{i}\left\|\Pi_i(K_i,X_k) - f\left(\boldsymbol{x}_i^k\right)\right\| \\ \mathscr{L}_{EO} = w_{EO}\sum_{k}\|\boldsymbol{E_k} - \widehat{\boldsymbol{E_k}}\| \\ \mathscr{L}_{GCP} = w_{GCP}\sum_{k_{GCP}}\sum_{g}\left\|\Pi\left(K_g,\boldsymbol{X}_k\right) - p_k\right\| \end{cases} \tag{4}$$

The final optimization function of the integrated bundle adjustment could be formulated as follows:

$$\underset{R,t,K,X}{\text{minimize}}\left(\mathscr{L}_{inner} + \mathscr{L}_{cross} + \mathscr{L}_{GCP} + \mathscr{L}_{EO}\right) \tag{5}$$

where $R$and$t$ denote the rotation and translation of the image, respectively, K refers to the matrix of the camera's interior orientation parameters, and X represents the unknown 3D points. Finally, by minimizing the 2-norm of the summary of the loss, both the optimized EO parameters and the unknown 3D points can be solved simultaneously.

3.1.4. Point cloud generation

With optimized image EO parameters, dense image matching and 3D point cloud generation can be achieved by a patch-based MVS algorithm (Wu, 2021; Furukawa and Ponce, 2010). The algorithm starts from the initial patches generated from corresponding points across multiple images. These initial patches are expanded gradually, considering their distance and depth, and enforce photometric consistency and global visibility constraints. To balance the hue of different data sources, an aerial image is selected as the reference image to adjust the hue of other images through histogram specification. The blue channel is depressed for backpack images to mitigate the severe disturbance from the blue sky due to camera pointing. The patch-based MVS algorithm is performed on aerial, MMS, and backpack images separately to obtain three sets of 3D point clouds with textural information. It implicitly assumes that the best textural information is preserved within the same data sources.

3.2. Fusion of multi-source point clouds for optimized 3D mapping

The point clouds, either from the above photogrammetric processing of different-source images or LiDAR measurements from various

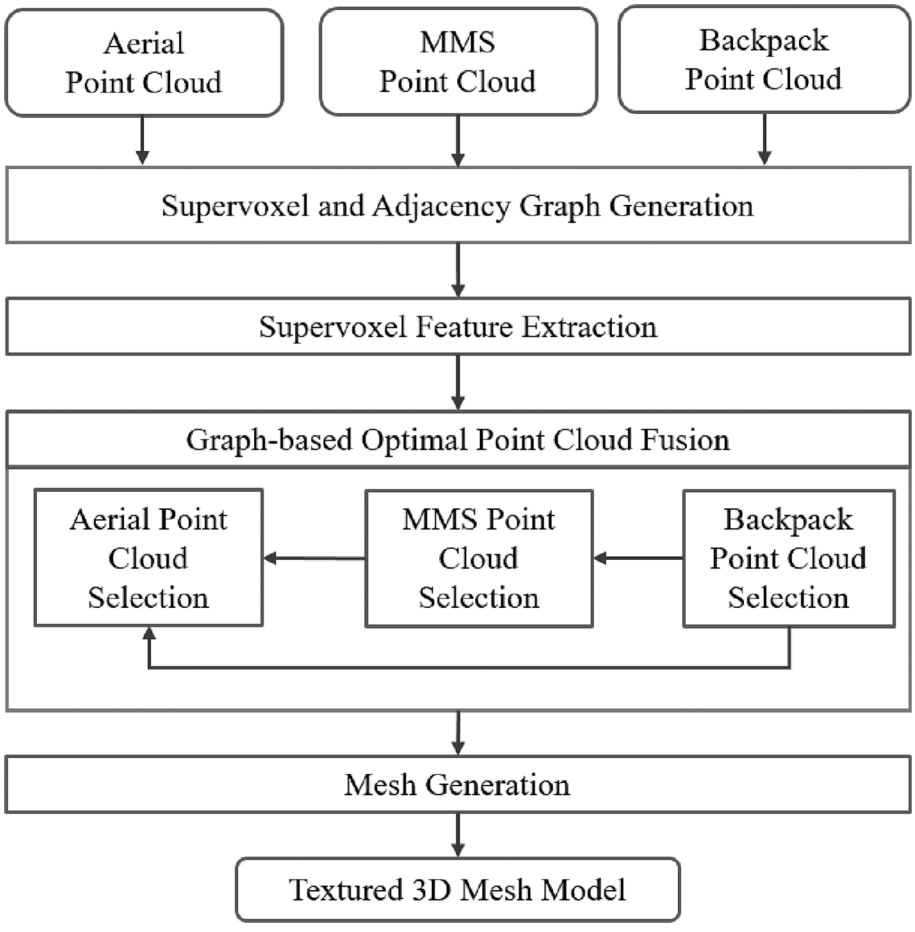


Fig. 3. Workflow of the optimal point cloud fusion method.

platforms, can be integrated for optimized 3D mapping (Li et al., 2019; Zhu et al., 2017). The union of point clouds provides maximal coverage, such that possible holes and merged objects (e.g., due to occlusions) in the original point clouds can be mitigated by combining point clouds from complementary views or sensors. However, directly merging multi-source point clouds may exacerbate the geometric ambiguity among the multiple platforms. Thus, a graph-based multi-source point cloud fusion method was designed to select appropriate points in the aspects of both geometry and texture. As illustrated in Fig. 3, the textured point clouds are first supervoxelized to merge homogenous neighbor points, thus minimizing the influence of noise, and the adjacency graphs of the supervoxels are generated simultaneously (Papon et al., 2013). Then, multi-dimension features (e.g., elevation, density, normal) are extracted for each supervoxel, and fed into the MRF framework to select the favorable supervoxels according to the energy minimization law (Boykov et al., 2001). Finally, points within the selected supervoxels are extracted and triangulated to generate the textured mesh model.

3.2.1. Supervoxel and adjacency graph generation

The object space defined by the point cloud is first segmented and clustered into supervoxels to enhance selection performance and boost efficiency. According to the VCCS algorithm (Li et al., 2019; Papon et al., 2013), three main types of characteristics should be considered in evaluating whether points should be clustered: spatial coordinate distance h_{coor}, color distance h_{color}, and local geometric feature difference h_{geo}. As specified by Zhu et al. (2017), the normal of a point can be used to present the local geometric features. Therefore, the total homogeneity h of a supervoxel can be formulated as follows:

$$h = h_{coor} + h_{color} + h_{normal} \tag{6}$$

where h_{normal} denotes the geometric difference. This criterion is then used to judge whether a point should be accepted by the supervoxel during the octree-based region growing strategy (Vo et al., 2015). The adjacency graph can be obtained directly through the octree structure. Furthermore, to better describe the relationship among various data sources, multi-adjacency graphs are generated to consider spatial distances. Specifically, for supervoxel S_A in point cloud A, if S_B is the nearest supervoxel in point cloud B, S_B and its connected neighbors should also be adjacent to S_A in the multi-adjacency graph.

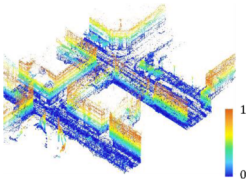
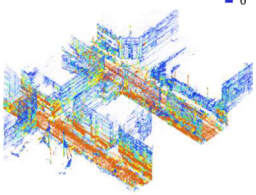
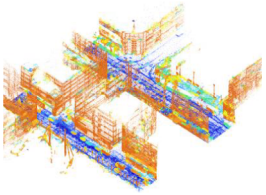
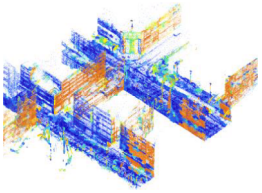
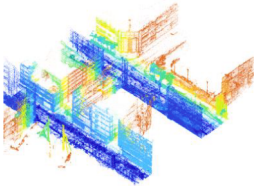
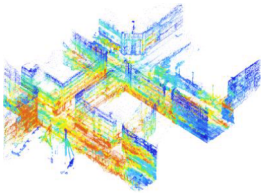
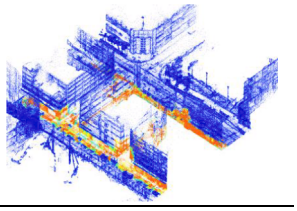
3.2.2. Supervoxel feature extraction

Regarding each supervoxel as a node and the adjacency relation as the edge in a graph, the selection procedure of each dataset can be formulated as an energy minimization problem in the MRF using graph-cut (Boykov et al., 2001), as shown in Equation (7). The overall energy $E_{overall}$ is usually composed of two main components: E_{data} and E_{smooth}. E_{data} evaluates supervoxel quality as constructed from the concerned features. E_{smooth} judges consistency among adjacent supervoxels based on the assumption that the neighbors of a selected supervoxel should also be chosen as neighbor areas that are prone to homogeneity.

$$\begin{gathered} E_{overall} = E_{data} + E_{smooth} \\ E_{data} = E\left(\mathscr{F}_{elevation}, \mathscr{F}_{density}, \mathscr{F}_{normal}, \mathscr{F}_{viewA}, \mathscr{F}_{viewD}, \mathscr{F}_{occlusion}, \mathscr{F}_{conflict}\right) \end{gathered} \tag{7}$$

As the involved features in E_{data} play a decisive role in the optimal fusion performance, the features extracted from supervoxels are presented in Table 1, which are all normalized to [0,1].

Table 1
Description of supervoxel features using part of an MMS point cloud as an example.

Feature	Description	Illustration
$\mathscr{F}_{elevation}$	The elevation of the center of the supervoxel based on the spatial reference. This feature is fundamental to distinguish the aerial- and ground-view point clouds.	
$\mathscr{F}_{density}$	The average number of points per unit to measure the geometric granularity. Dense points extracted from the MVS provide more vertexes for the subsequent mesh generation to recover the 3D building and imply the high quality of the observed images.	
$\mathscr{F}_{normal}$	The normal describes the pointing direction of the supervoxel, constructed by principal component analysis (PCA) (Dunteman, 1989). It can be further used to measure the angle between the horizontal planes $\mathscr{F}_{n2h}$ and the vertical planes $\mathscr{F}_{n2v}$.	
$\mathscr{F}_{viewA}$	The best view angle of the supervoxel, calculated with the pointing direction of the observed images and the supervoxel's normal. The images viewed from the opposite direction (i.e., $\lvert\mathscr{F}_v - 90^\circ\rvert = 90^\circ$) can retrieve the best texture.	
$\mathscr{F}_{viewD}$	The minimum view distance from the visible images to the supervoxel. The supervoxel viewed from a long distance may possess blurry textures.	
$\mathscr{F}_{occlusion}$	Occlusion measures the visibility quality of each supervoxel. The line between the supervoxel and the projected images is generated, and the occurrence of a supervoxel on the line but closer to the image is examined.	
$\mathscr{F}_{conflict}$	The conflict ratio with the already selected supervoxels. The selected points are regarded as the inliers and should not overlap.	

3.2.3. Graph-Based optimal fusion of point clouds for mesh model generation

As three data sources each have their unique merits, the selection tends to utilize these advantages in a sequential manner. Backpack images typically feature the most flexible route and nearest view distance, which yield more favorable point clouds than MMS point clouds. The MMS platform can capture clear textural images of building façades and other near-ground information ignored by the backpack platform. Lastly, the large-scale aerial point cloud can supplement all existing holes among ground-view point clouds.

Therefore, beginning with the voxelized backpack point clouds, the energy of the function can be constructed as Equation (7), which considers all features other than the conflict ratio. Through alpha-expansion (Boykov et al., 2001), labels are assigned to the proper points to achieve the minimum energy. The neighbors' vote strategy is also conducted to guarantee object completeness and minimize the effect of noise and irregular geometry. Selected points are then injected into the MMS

Table 2
Details of the aerial, MMS, and backpack images used for experimental analysis.

Dataset	Sensor	Focal Length (mm)	GSD (cm)	Number of Cameras	CCD Size (pixels)	Number of Images	Collection Date	Height (m)
Aerial	Canon EOS 5DS	49/35	6	12	8,688 × 5,792	121	07/11/2016 to 02/12/2016	~500
MMS	Leica MMS	8	1	6	2,048 × 2,048	1,895	05/07/2019	~2
Backpack	Leica Pegasus Backpack	6	1	5	2,046 × 2,046	4,202	05/07/2019	~2

Table 3
Viewing angle and intersection angle of each experiment.

Experiment Name	**Viewing angle (°)** (the camera pointing angle with respect to the observation plane)			**Intersection angle (°)**	
	Aerial	MMS	Backpack	Aerial-MMS	Aerial-Backpack
Common	132.36	90.82	132.40	104.36	90.79
Renovated	96.11	124.80	80.68	89.48	128.53
Complexed	131.82	88.97	167.97	76.48	49.87
Glassy	107.84	119.87	/	44.66	/
Repeated	133.46	/	129.97	/	55.95

selection module to update the conflict ratio of each MMS supervoxel, followed by a similar selection procedure to that used for the backpack point cloud. Finally, with the known selected backpack and MMS points, aerial points can be analyzed to remove useless points and supplement areas uncovered by the other data sources.

With the selected textured point clouds, the textured mesh model can be generated according to the triangulation algorithm cast as a Poisson problem (Kazhdan and Hoppe, 2013) in the interactable OBJ format and the texture information in the MTL format.

4. Experimental results

4.1. Dataset Description

In this paper, a challenging dataset composed of aerial, MMS, and backpack images was used to evaluate the performance of the proposed workflow. It included various types of buildings and road furniture covering over 40,000 m^2 of the high-rise area in Kowloon Bay, Hong Kong. Table 2 lists the detailed information of three image groups. The 121 aerial images were acquired by a helicopter using an AMC PanOblique, which consisted of 12 Canon EOS 5DS cameras to achieve 360° panoramic oblique images. Limited by a 500-m flying height, the ground sampling distance (GSD) was 6 cm and the clearness of images was affected by sunlight and the atmosphere. The collection of ground images utilized a MMS with six cameras to capture the panoramic images, and the Leica Pegasus backpack system with five cameras. As these images were obtained in 2019, 3 years after the aerial images, many changes have occurred (e.g., reconstruction of buildings, updating of posters, moving of vehicles), which may have further hindered matching between the aerial and ground-view images. The height of the MMS and backpack cameras was approximately 2 m, which is lower than most observed scenes and opposite to the aerial images viewing direction. All images were provided with the initial position and pointing information provided by the GPS and IMU measurements in the Hong Kong 1980 Grid System (EPSG: 2326).

4.2. Evaluation of Tie-point matching between aerial and ground images

To comprehensively evaluate the performance of the proposed matching strategy, five presentative matching results are highlighted: the details of the viewing angle and the intersect angle are listed in Table 3, and the retrieved numbers of tie-points for each experiment are summarized in Table 4. Notably, all of the matching algorithms (SIFT, ASIFT, Harris, SuperGlue, and geometric-aware SuperGlue) were performed on the images, but the former three matching algorithms failed in all experiments. Even the ASIFT algorithm, which simulates many perspective deformation scenarios, could not manage the variation in clearness and the harsh view conditions.

The first experiment focused on the most common and simple buildings in the urban area, and the corresponding region is marked with a red box in Fig. 4(a). Fig. 4(d) illustrates the successful matching between the aerial and MMS images with more than 100° intersect angles. Although this MMS image offers a large view of the building façade (red box), the perpendicular view leads to inevitable occlusion and distortion of the far side. However, nine evenly distributed matches were still achieved, as shown in the zoomed view in Fig. 4(f). The algorithm performed better on the backpack image looking up to the building, with 130 matched tie-points across different planes. This result indicates that through the union of geometric-aware and the learning-based algorithm, the tie-points are no longer limited within a plane as traditional methods (e.g., Wu et al., 2018), and the more focused matching task could manage similar plane confusion. However, no tie-points were obtained for the grey building adjacent to the red box due to repainting and reconstruction.

An in-depth study of this grey building was conducted in experiment "Renovated". As presented in Fig. 5, this building was repainted and reconstructed, particularly at the windows on the right of the façade, such that recognizing the corresponding points is even difficult for a human. By leveraging the proposed method, ~30 tie-points were matched on both MMS and backpack images, mainly at the bottom of the

Table 4
Number of matched tie points between the MMS/BP and aerial images in each experiment using different algorithms (BP is the abbreviation of backpack).

Experiment Name	**SIFT** (Lowe, 2004)		**ASIFT** (Morel and Yu, 2009)		**Harris** (Harris and Stephens, 1988)		**SuperGlue** (Sarlin et al., 2020)		**Geometric-aware SuperGlue**	
	MMS	BP	MMS	BP	MMS	BP	MMS	BP	MMS	BP
Common	0	0	0	0	0	0	0	0	9	130
Renovated	0	0	0	0	0	0	2	0	31	25
Complexed	0	0	0	0	0	0	0	3	43	49
Glassy	0	0	0	0	0	0	3	/	27	/
Repeated pattern	0	0	0	0	0	0	/	0	/	29

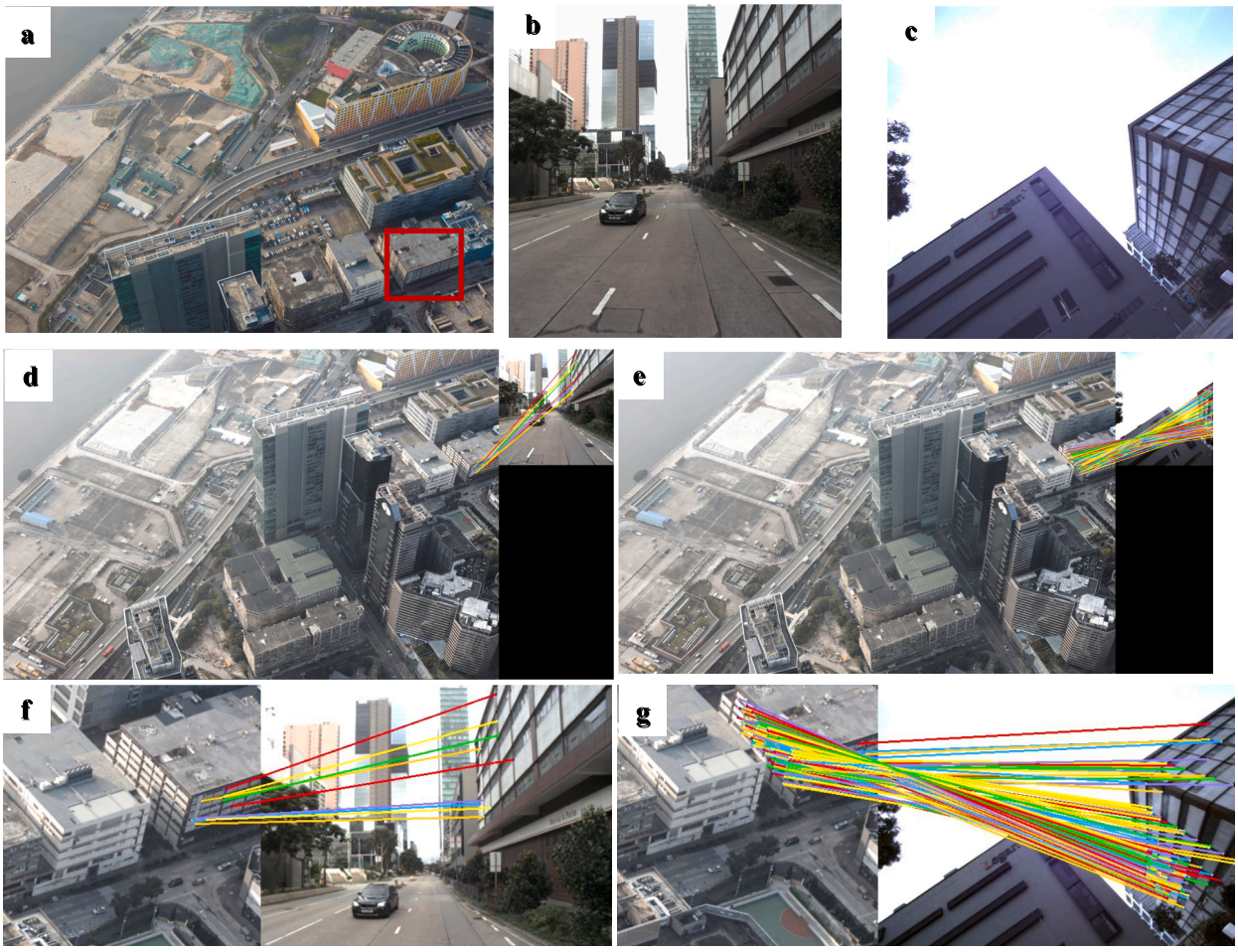


Fig. 4. Tie-pointmatching results of a common building. (a)–(c) The original aerial, MMS, and backpack images; (d)-(e) the matching results of the aerial and MMS/backpack images; and (f)-(g) the zoomed view of (d) and (e). Line color varies for clear presentation.

building, where the original structure was well-preserved. Human checking verified that these tie points were all correct and could be used for the bundle adjustment. These results show the feasibility of using Internet-based community photo collections (CPCs), which provided abundant and incremental information regarding the urban area reconstruction, despite spanning a long time period and undergoing renovation and other changes.

The third experiment, shown in Fig. 6, was conducted on a building with a complicated texture caused by air conditioners, and the observing condition was the worst among the five experiments as the aerial image was captured from a nearly vertical direction. However, the algorithm still performed robustly, as the complex texture possesses sufficient interest points to offer abundant contextual information. Although the building only occupies a small space in the backpack image due to the reflective light condition, the algorithm matched 49 tie points.

As glassy buildings have become common in urban areas, the fourth experiment focused on the skyscraper (Fig. 7). Other than the glassy characteristics, this façade also featured repeated patterns; but with relatively favorable viewing conditions, the aerial and MMS images were matched with 27 tie-points distributed across the entire co-visible region. However, matching between the aerial and backpack images failed due to the reflection of the sky.

The MMS image failed to connect with the aerial image due to the occlusion issue seen in Fig. 8, while the proposed algorithm obtained 29 matched tie points for the backpack image, despite disturbance from the extremely repeated pattern.

Overall, the proposed algorithm could manage the most challenging scenarios. Despite constraints due to visibility and viewing conditions, the MMS and backpack images offered complementary views to increase the number of tie-points with the aerial image.

4.3. Evaluation of integrated bundle adjustment

Using the proposed methods, 13,136 aerial-MMS/backpack tie points are retrieved, involving all the aerial images, 22.32 % MMS and 16.11 % backpack images, as shown in Table 5. As no ground truth data were available to evaluate the matching performance comprehensively, bundle adjustment results were instead utilized for quantitative evaluation. Analysis of the tie-points showed that the re-projection error was improved from 2.63 pixels to 0.97 pixels (Table 5), which suggests that consistency was well-achieved among different data sources. Other than more aerial-ground tie points being used in the bundle adjustment, the improvement of the result was mainly attributed to the sub-pixel precision of the matching algorithm (Sarlin et al., 2020). And this was hard to be guaranteed by manually digitized tie points, especially when large resolution variation was involved. The positioning uncertainty, derived from the covariance matrix of the calculated 3D positions of the tie-points through the bundle adjustment (Morris et al., 2000; Rodríguez-Arévalo et al., 2018), is also analyzed. Table 6 compares the position uncertainty using aerial images only, aerial and MMS images, and the integration of aerial, MMS, and backpack images. Clearly, uniting the MMS ground view with the aerial images caused uncertainty to greatly drop to within 1 cm, which was further reduced ~ 50 % through integrated bundle adjustment with more consecutive and closer-range backpack images. Compared with our previous results (Li et al., 2020) based on manually digitized feature tracks, the overall trend was similar,

Fig. 5. Tie-point matching results of a repainted and reconstructed building. (a)–(c) The original aerial, MMS, and backpack images; (d)-(e) the matching results of the aerial and MMS/ backpack images; and (f)-(g) the zoomed view of (d) and (e).

but the current results are more precise, which indicates that more effective tie-points were retrieved and served as the constraints in the bundle adjustment.

4.4. Evaluation of multi-source point cloud fusion

Fig. 9 shows the overall results of the graph-based optimal point cloud selection and the generated textured mesh model. Intuitively, point cloud selection is mainly distributed according to elevation but is also influenced by other features. For instance, although the elevations of the vehicle road and sidewalk were nearly identical, MMS data could provide a denser and closer viewing distance that better expressed the vehicle roads, while the sidewalks were reconstructed with backpack points due to their direct view without occlusion. Building façades perpendicular to the route of MMS were almost entirely reconstructed from MMS points, whereas only the lower parts of pass-by façades were from MMS images and the higher parts were from aerial images.

The effectiveness of the fusion of multi-source point clouds is illustrated in Fig. 10. The first column visualizes the 3D reconstruction results of aerial images in which the traffic sign and the other details beneath the bridge are missed and the texture of the alley is blurry due to the ~ 500-m flying height. The integrated bundle adjustment of aerial and MMS images generated mesh models, as shown in the second column. Although the texture of the alley was not greatly improved, the information under the bridge (the further part in the middle of the region) unfolded clearly. Through the fusion of multi-source point clouds, the entire 3D scene was reconstructed densely with clear texture and the road furniture (e.g., the fence, road lamp, and motorcycles) was well-retrieved.

To further analyze the advantages of the proposed algorithm, two challenging alleys in the scene were selected and compared with the results from the ContextCapture software. Fig. 11 shows a narrow alley between two buildings. Similar to the results in Fig. 10, the mesh model became more detailed as more data sources were incorporated. As the hue of the images greatly variated in this region, the texture became more mottled when backpack images were integrated. A detailed comparison of a zoomed view is provided in Fig. 12. Regarding each data source equally, the result from ContextCapture in Fig. 12(a) suffers from the mottled texture of the building facades, which can be seen across multiple data sources. Furthermore, the truck on the road is only partially reconstructed due to the ambiguity among the data sources. These texture and fragment issues were avoided by the proposed algorithm as shown in Fig. 12(b). As suggested by Fig. 12(c), the building façades were entirely expressed by MMS points, while the traffic sign and the parked vehicles were retrieved completely. Other than the prior embedded knowledge in $\boldsymbol{E}_{data}$, which guarantees the geometric and textural quality of the selected supervoxels, the $\boldsymbol{E}_{smooth}$ term enforces neighborhood consistency, thus preventing mottled texture and fragmentation. These issues are further mitigated by the neighbor vote strategy, which increases the selection region to eliminate possible fragmentation.

Another alley featuring complicated furniture and vegetation was

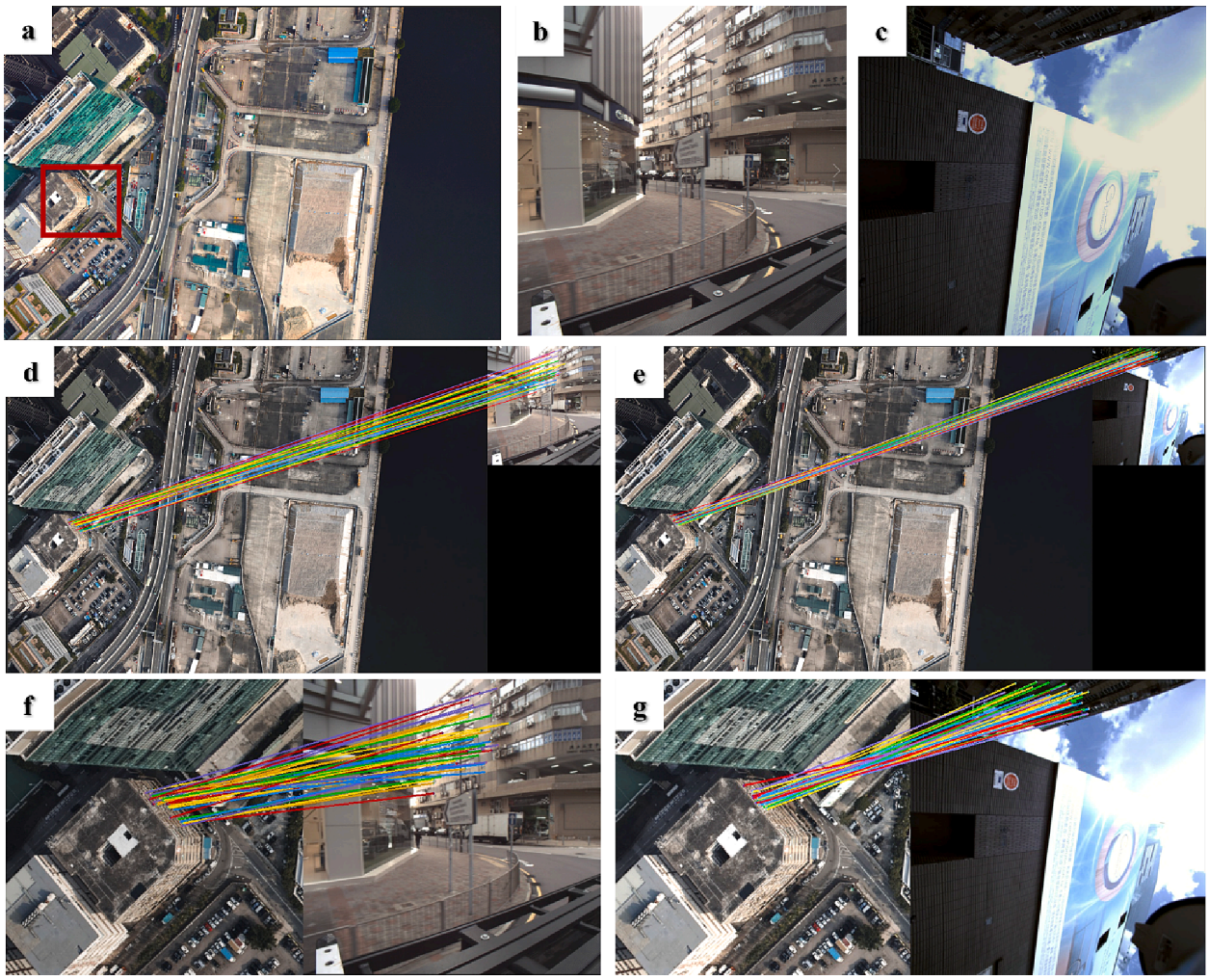


Fig. 6. Tie-pointmatching results of a building with complicated texture. (a)–(c) The original aerial, MMS, and backpack images; (d)-(e) the matching results of the aerial and MMS/ backpack images; and (f)-(g) the zoomed view of (d) and (e).

also selected for comparison (Fig. 13). Similarly, the ContextCapture software result possessed apparent defects on the building façade. The comparison of the zoomed view with the point cloud selection result is provided in Fig. 14. While the ContextCapture software mixes different data sources to present the vehicle road (the furthest upper region), our algorithm chooses the MMS points for dense reconstruction and consistent texture. The alley region was mostly covered with backpack images as the only data source that could capture the inner details. However, the rooftops of the low-rise buildings were automatically reconstructed with aerial points because they could not be seen in the backpack or MMS images. This example demonstrates the strength of the aerial points, which are required for not only the roofs of buildings but also accurate geometry retrieval for some low-rise buildings.

5. Discussion

Despite its success, the proposed solution has some limitations. First, the geometric-aware learning-based algorithm is relatively complex for end-users, as it requires the segmentation of aerial point clouds and the computation of planes for image rectification. The current segmentation algorithm is not robust enough to extract all the building facades and requires manual checking and interactive improvements. Second, the tie points are not distributed evenly due to different image quality and inherent characteristics of the imaging area, especially in a large-scale scenario. The regions with abundant textures may be matched with a large number of tie points, much more than other regions with poor textural conditions. This may lead to a local optimal convergence of the bundle adjustment. In the current solution, the number of tie points in different regions has to be balanced manually, thus a more automatic strategy should be considered to guarantee the performance of bundle adjustment. Third, the point cloud fusion algorithm is inherently an optimal selection strategy that is not able to compensate for possible defects within the original point clouds. However, compared with conventional point cloud registration methods (Huang et al., 2023; Li and Harada, 2022; Monji-Azad et al., 2023) that merge different point clouds together directly after applying some transformations, the proposed point cloud fusion algorithm selects supervoxels of points with optimal geometric and textural properties for 3D modelling, which can improve the problems of geometric and textural inconsistencies in local regions and data redundancy in overlapping regions that are common in conventional methods.

It should be noted that the geometric-aware learning-based algorithm requires the initial EO parameters of images as a-priori knowledge. If no such information provided, a regular SfM processing for each type of images will be necessary to obtain their initial EO parameters, which can be achieved by using any off-the-shelf commercial solution (e.g., ContextCapture). It should also be noted that the radiometric properties of different-source images may vary as they were collected by different sensors onboard different platforms under different illumination conditions. In the current solution, we performed histogram matching to reduce the radiometric inconsistencies (e.g., brightness and hue) among the different-source images. However, more effective

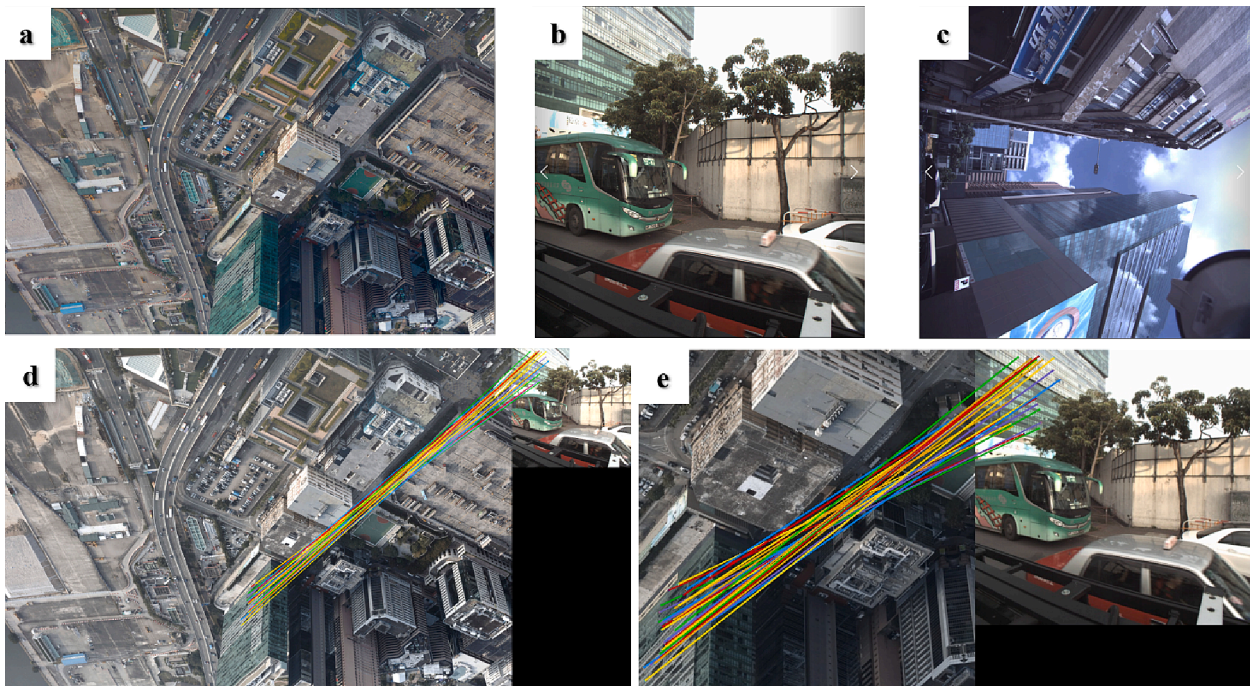


Fig. 7. Tie-point matching results of a glassy building. (a)–(c) The original aerial, MMS, and backpack images; (d) the matching results of aerial and MMS images; and (e) the zoomed view of (d).

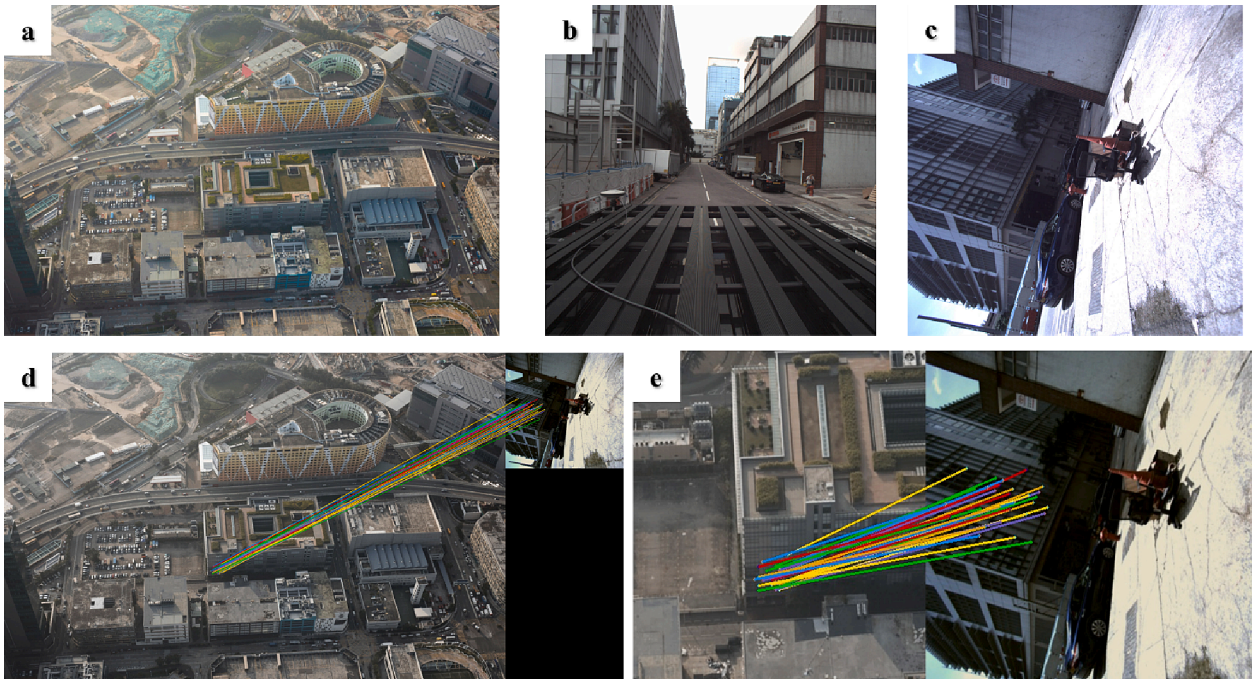


Fig. 8. Tie-point matching results of a repeated pattern building. (a)–(c) The original aerial, MMS, and backpack images; (d) the matching results of aerial and backpack images; and (e) the zoomed view of (d).

Table 5
Comparison of image residuals of tie-points.

Method for Obtaining Tie Points	Aerial-MMS/Backpack Tie-point Number	Matched Image Rate (%)			Mean Residuals (95 %) (Pixels)	RMSE of Residuals (95 %) (Pixels)
		Aerial	MMS	Backpack		
Manual (Li et al., 2020)	10,139	100	20.63	5.50	2.93 (2.63)	3.61 (3.08)
Geometric-aware SuperGlue	13,136	100	22.32	16.11	1.13 (0.97)	1.28 (1.02)

Table 6
Comparison of the positioning uncertainty.

Method for Obtaining Tie Points	Aerial images only (cm)			Aerial and MMS images (cm)			Aerial, MMS, and Backpack images (cm)		
	X	Y	Z	X	Y	Z	X	Y	Z
Manual	9.13	8.55	5.35	0.95	0.83	0.68	0.37	0.34	0.27
(Li et al., 2020) Geometric-aware SuperGlue				0.29	0.29	0.33	0.16	0.15	0.16

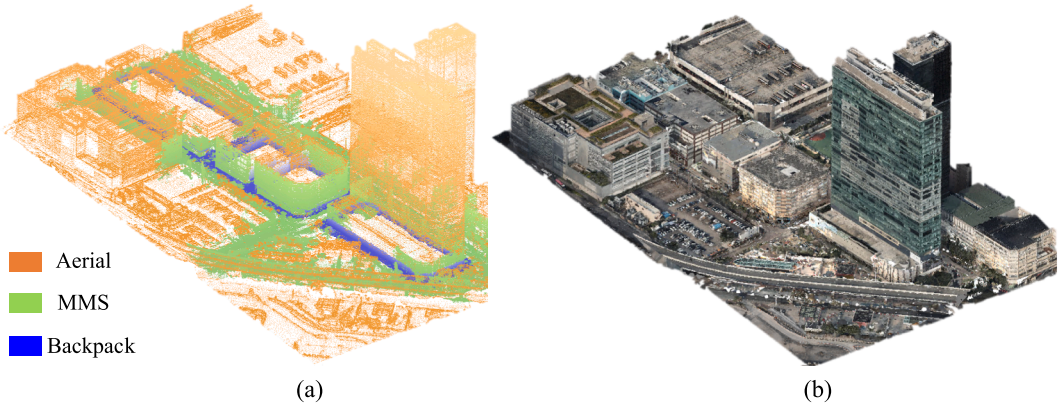


Fig. 9. Overall results of the graph-based optimal point clouds fusion algorithm. (a) The selection of the point clouds from different sources and (b) the textured mesh model generated from the selected points.

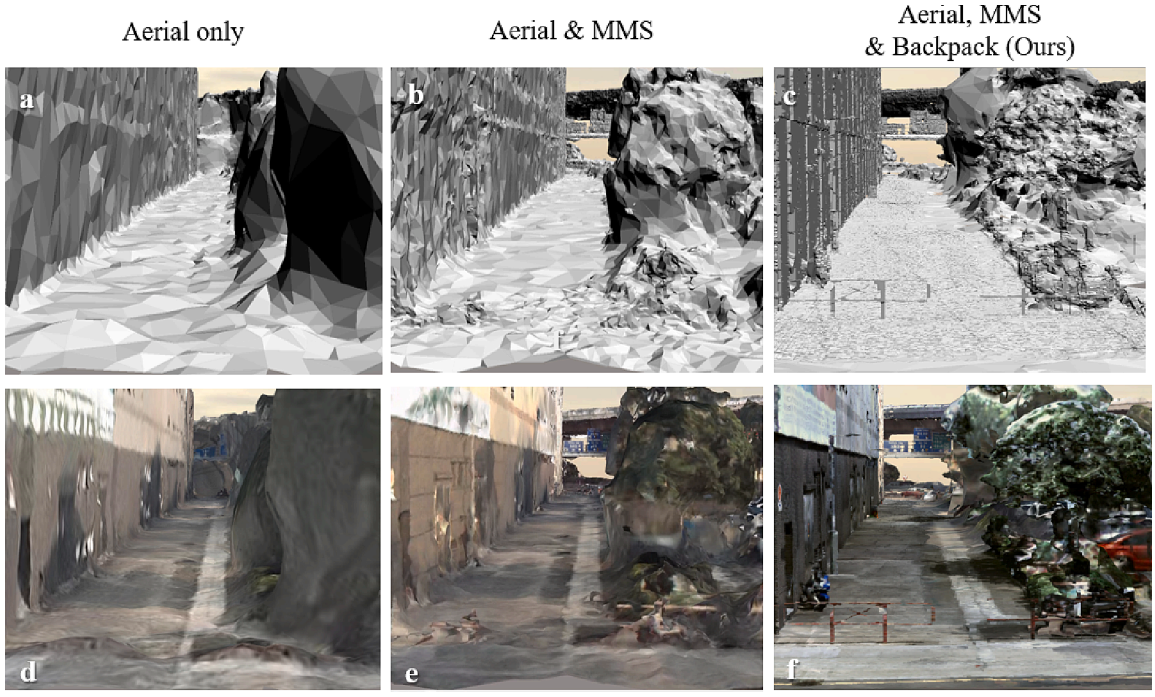


Fig. 10. Effectiveness of the fusion of multi-source point clouds. (a)-(c) The 3D mesh models generated from different image sources; and (d)-(f) the corresponding textured 3D mesh models.

methods are desirable to reduce the radiometric differences of images to generate high-quality textured 3D mesh models.

6. Conclusions

To integrate aerial, MMS, and backpack images and point clouds for optimized 3D mapping in urban areas, this paper presented a geometric-aware model based on the SuperGlue algorithm to solve the tie-point matching problem, followed by the graph-based method for multi-source point cloud fusion. Rather than directly feeding the network with the original images, the nominal EO parameters of images are used to guide learning-based tie-point matching to improve accuracy and reduce computational complexity. Based on the abundant and precise tie points, bundle adjustment and dense image matching could be

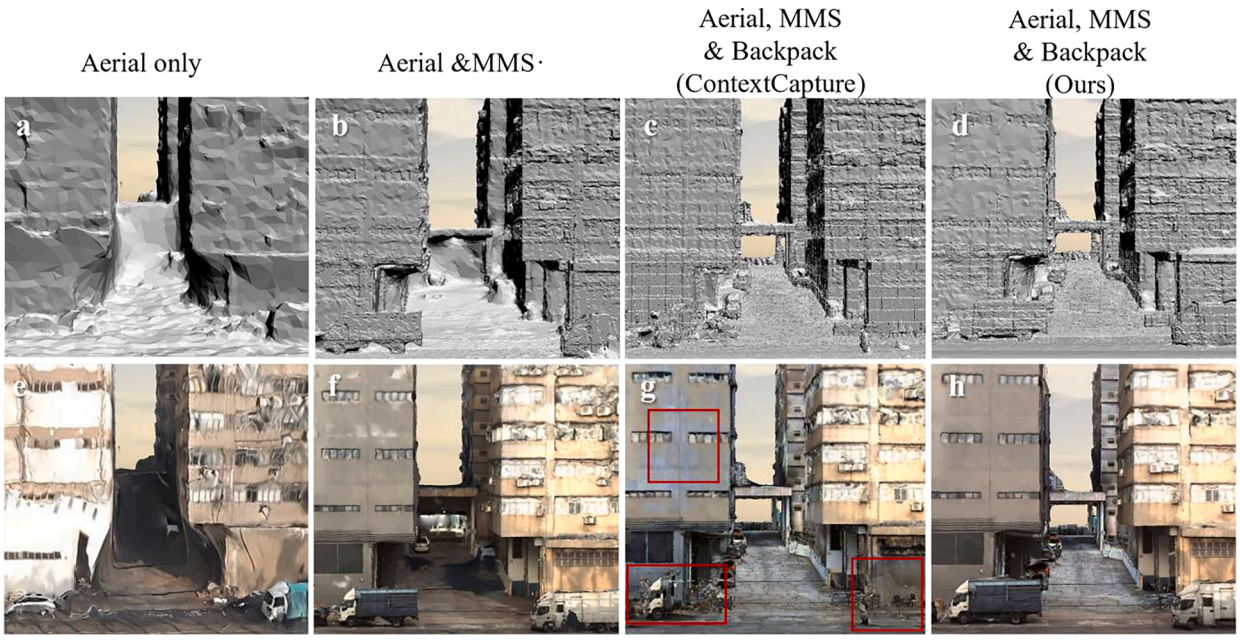


Fig. 11. Effectiveness of the fusion of multi-source point clouds in an alley region. (a)-(d) The 3D mesh models generated from different image sources using different methods; and (e)-(h) the corresponding textured 3D mesh models. The red boxes in the third column indicate defects in the ContextCapture results that are improved in the results of the proposed method. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

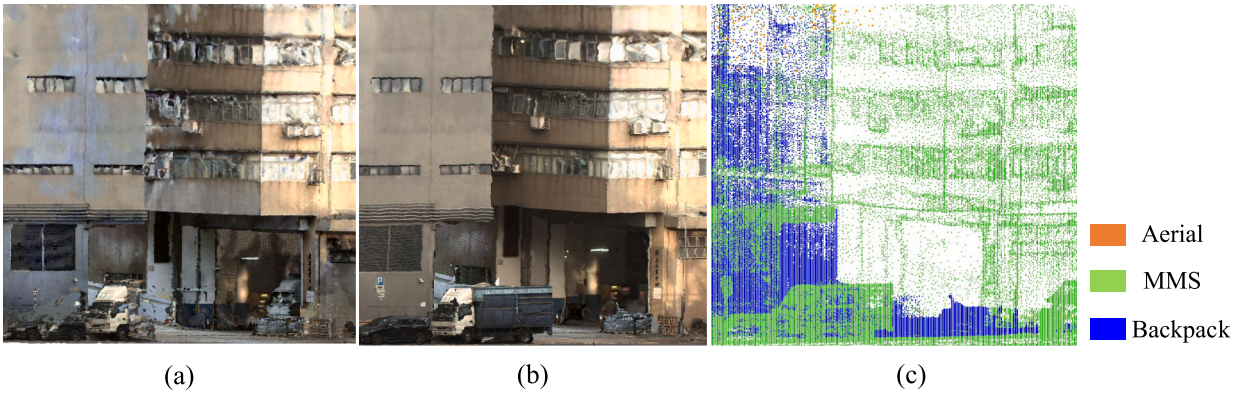


Fig. 12. Enlarged views of the comparison between the results from ContextCapture and our method. (a) The 3D textured mesh model from ContextCapture; (b) the 3D textured mesh model generated from the proposed multi-source fusion algorithm; and (c) the optimal point cloud selection results.

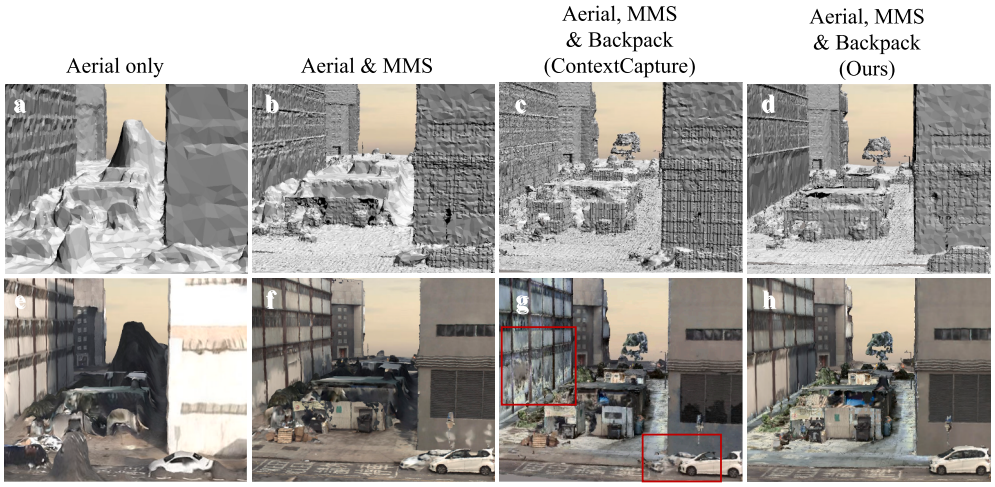


Fig. 13. Effectiveness of the fusion of multi-source point clouds in an alley region featuring complicated furniture. (a)-(d) The 3D mesh models generated from different image sources using different methods; and (e)-(h) the corresponding textured 3D mesh models. The red boxes in the third column indicate defects in the ContextCapture results that are improved in the results of the proposed method. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

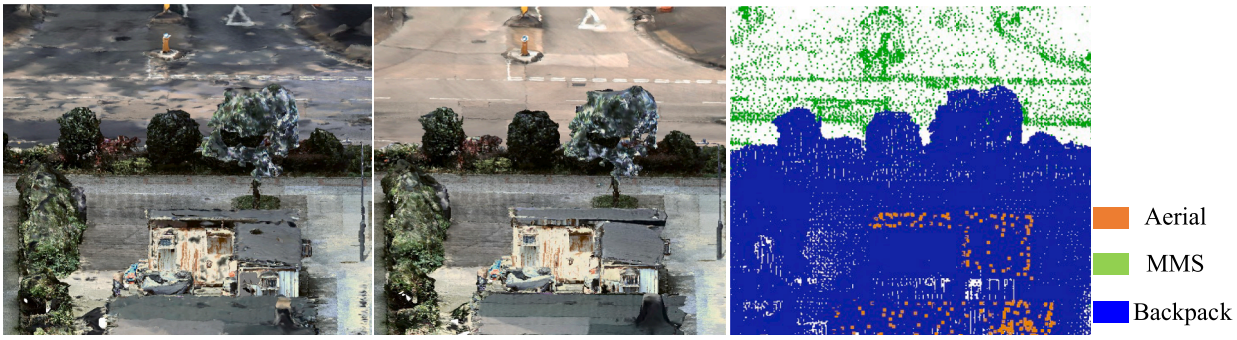


Fig. 14. Enlarged views of the comparison between the results from ContextCapture and our method for the region featuring complicated furniture. (a) The 3D textured mesh model from ContextCapture; (b) the 3D textured mesh model generated from the proposed multi-source fusion algorithm; and (c) the optimal point cloud selection results.

performed to generate point clouds from images collected by different platforms. Finally, these point clouds could be integrated through a graph-based algorithm in the MRF framework by taking advantage of various features of supervoxels and adjacency constraints.

Experiments were performed on a typical urban dataset in Hong Kong consisting of 121 aerial images and more than 6000 ground (MMS and backpack) images. Five representative matching groups between the aerial and ground images were selected and compared with off-the-shelf solutions. The results revealed that the proposed geometric-aware SuperGlue model can obtain dozens of tie points between the aerial and ground images with perspective differences up to 128° (see Tables 3 and 4); while the traditional feature matching methods (such as SIFT, ASIFT, Harris) failed to obtain any tie points (Table 4). Compared to the abundant tie points obtained by the geometric-aware SuperGlue, the original SuperGlue algorithm only obtained several tie points for certain matching groups (Table 4), which showed the effectiveness of the geometric-aware model. The high-quality tie points obtained from the geometric-aware SuperGlue facilitated the integrated bundle adjustment of aerial and ground images, and the re-projection errors of the tie points after bundle adjustment dropped to about one pixel (Table 5), which in turn resulted in positioning accuracy of less than one centimeter (Table 6). For multi-source data fusion, evaluations were performed on three narrow alleys in the datasets, which illustrated the effectiveness of integrating aerial, MMS, and backpack images for producing more completed, consistent, and detailed 3D models (see Figs. 11-14) as compared with the off-the-shelf commercial solution.

The developed methods presented in this paper allow the fusion of images and point clouds collected from different platforms (aerial and ground platforms in particular) for optimized 3D mapping and modelling in urban areas. They can facilitate the generation of 3D city models of the best quality in terms of accuracy, completeness, consistency, and level of detail, which can be used in various applications such as digital twin cities (Lehtola et al., 2022) and smart cities (White et al., 2021).

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

This work was supported by grants from the Hong Kong Polytechnic University (Project ID P0046112), the Research Grants Council of Hong Kong (Project No: PolyU 15210520, Project No: PolyU 15219821, Project No: 15215822), and the National Natural Science Foundation of China (Project No. 42201476). The authors would also like to thank the Survey and Mapping Office of the Lands Department of the HKSAR government for providing the experimental datasets.

References

Agarwal, S., Snavely, N., Simon, I., Seitz, S.M., Szeliski, R., 2009. Building Rome in a Day. IEEE I. C. Comp. Vis. 72–79.
Balta, H., Velagic, J., Bosschaerts, W., De Cubber, G., Siciliano, B., 2018. Fast Statistical Outlier Removal Based Method for Large 3D Point Clouds of Outdoor Environments. IFAC-PapersOnLine 51, 348–353.
Boykov, Y., Veksler, O., Zabih, R., 2001. Fast approximate energy minimization via graph cuts. IEEE Trans. Pattern Anal. Mach. Intell. 23, 1222–1239.
Chu, X., Zhang, B., Tian, Z., Wei, X., Xia, H., 2021. Do We Really Need Explicit Position Encodings for Vision Transformers? ArXiv abs/2102.10882.
Cui, Y., Chen, R., Chu, W., Chen, L., Tian, D., Li, Y., Cao, D., 2022. Deep Learning for Image and Point Cloud Fusion in Autonomous Driving: A Review. IEEE Trans. Intell. Transp. Syst. 23, 722–739.
Dai, A., Chang, A.X., Savva, M., Halber, M., Funkhouser, T., Nießner, M., 2017. Scannet: Richly-annotated 3d reconstructions of indoor scenes. CVPR 2017, 5828–5839.
DeTone, D., Malisiewicz, T., Rabinovich, A., 2018. SuperPoint: Self-Supervised Interest Point Detection and Description, in: Proceedings 2018 IEEE/Cvf Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 337-349.
Digne, J., de Franchis, C., 2017. The Bilateral Filter for Point Clouds. Image Processing on Line 7, 278-287.
Dunteman, G.H., 1989. Principal components analysis. Sage.
Fassi, F., Perfetti, L., 2019. Backpack Mobile Mapping Solution For Dtm Extraction Of Large Inaccessible Spaces. ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences.
Fischler, M.A., Bolles, R.C., 1981. Random Sample Consensus - a Paradigm for Model-Fitting with Applications to Image-Analysis and Automated Cartography. Commun. ACM 24, 381–395.
Förstner, W., Gülch, E., 1987. A fast operator for detection and precise location of distinct points, corners and centres of circular features, ISPRS intercommission conference on fast processing of photogrammetric data, pp. 281-305.
Furukawa, Y., Ponce, J., 2010. Accurate, Dense, and Robust Multiview Stereopsis. IEEE Trans. Pattern Anal. Mach. Intell. 32, 1362–1376.
Gao, X., Shen, S.H., Zhou, Y., Cui, H.N., Zhu, L.J., Hu, Z.Y., 2018. Ancient Chinese architecture 3D preservation by merging ground and aerial point clouds. ISPRS J. Photogramm. Remote Sens. 143, 72–84.
Harris, C., Stephens, M., 1988. A combined corner and edge detector. Alvey vision conference. Citeseer 10–5244.
He, L.F., Ren, X.W., Gao, Q.H., Zhao, X., Yao, B., Chao, Y.Y., 2017. The connected-component labeling problem: A review of state-of-the-art algorithms. Pattern Recogn. 70, 25–43.
Huang, J.H., Birdal, T., Gojcic, Z., Guibas, L.J., Hu, S.M., 2023. Multiway Non-Rigid Point Cloud Registration via Learned Functional Map Synchronization. IEEE T. Pattern. Anal. 45, 2038–2053.
Jin, Y., Mishkin, D., Mishchuk, A., Matas, J., Fua, P., Yi, K.M., Trulls, E., 2021. Image matching across wide baselines: From paper to practice. Int. J. Comput. Vis. 129, 517–547.
Kazhdan, M., Hoppe, H., 2013. Screened Poisson Surface Reconstruction. ACM Trans. Graph. (ToG) 32.
Kim, T., Choi, J., Choi, S., Jung, D., Kim, C., 2021. Just a Few Points are All You Need for Multi-view Stereo: A Novel Semi-supervised Learning Method for Multi-view Stereo. 2021 IEEE/Cvf International Conference on Computer Vision (ICCV 2021), 6158-6166.
Lehtola, V., Koeva, M., Elberink, S., Raposo, P., Virtanen, J.-P., Vahdatikhaki, F., Borsci, S., 2022. Digital twin of a city: Review of technology serving city needs. Int. J. Appl. Earth Obs. Geoinf. 114, 102915.

Li, Z., Wu, B., Li, Y., 2020. Integration of Aerial, MMs, and Backpack Images for Seamless 3D Mapping in Urban Areas. The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences XLIII-B2-2020, 443-449.
Li, J., Li, M., Li, Z., Peng, S., 2022. Super-Voxel Graph Guided 3D Point Cloud Denoising, 2022 14th International Conference on Computer Research and Development, pp. 276-280.
Li, Y., Harada, T., 2022. Lepard: Learning partial point cloud matching in rigid and deformable scenes. CVPR 2022, 5544–5554.
Li, Z., Snavely, N., 2018. Megadepth: Learning single-view depth prediction from internet photos, in: IEEE conference on computer vision and pattern recognition, pp. 2041-2050.
Li, Y., Wu, B., Ge, X.M., 2019. Structural segmentation and classification of mobile laser scanning point clouds with large variations in point density. ISPRS J. Photogramm. Remote Sens. 153, 151–165.
Lowe, D.G., 2004. Distinctive image features from scale-invariant keypoints. Int. J. Comput. Vis. 60, 91–110.
Mei, G., Poiesi, F., Saltori, C., Zhang, J., Ricci, E., Sebe, N., 2023. Overlap-guided Gaussian Mixture Models for Point Cloud Registration, Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 4511-4520.
Mikolajczyk, K., Tuytelaars, T., Schmid, C., Zisserman, A., Matas, J., Schaffalitzky, F., Kadir, T., Gool, L.V., 2005. A Comparison of Affine Region Detectors. Int. J. Comput. Vis. 65, 43–72.
Monji-Azad, S., Hesser, J., Low, N., 2023. A review of non-rigid transformations and learning-based 3D point cloud registration methods. ISPRS J. Photogramm. Remote Sens. 196, 58–72.
Morel, J.M., Yu, G.S., 2009. ASIFT: A New Framework for Fully Affine Invariant Image Comparison. SIAM J. Imag. Sci. 2, 438–469.
Morris, D.D., Kanatani, K., Kanade, T., 2000. Uncertainty modeling for optimal structure from motion, Vision Algorithms: Theory and Practice: International Workshop on Vision Algorithms Corfu, Greece, September 21–22, 1999 Proceedings. Springer, pp. 200-217.
Nair, V., Hinton, G.E., 2010. Rectified linear units improve restricted boltzmann machines, ICML.
Papon, J., Abramov, A., Schoeler, M., Worgotter, F., 2013. Voxel Cloud Connectivity Segmentation - Supervoxels for Point Clouds. Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 2027-2034.
Paullada, A., Raji, I.D., Bender, E.M., Denton, E., Hanna, A., 2021. Data and its (dis) contents: A survey of dataset development and use in machine learning research. Patterns 2, 100336.
Peyre, G., Cuturi, M., 2019. Computational Optimal Transport. Found. Trends Mach. Le. 11, 355–607.
Rodríguez-Arévalo, M.L., Neira, J., Castellanos, J.A., 2018. On the importance of uncertainty representation in active SLAM. IEEE Trans. Rob. 34, 829–834.
Rusu, R.B., Cousins, S., 2011. 3D is here: Point Cloud Library (PCL). 2011 IEEE International Conference on Robotics and Automation (ICRA).
Sarlin, P.E., DeTone, D., Malisiewicz, T., Rabinovich, A., 2020. SuperGlue: Learning Feature Matching with Graph Neural Networks. CVPR 2020, 4937–4946.
Schonberger, J.L., Frahm, J.M., 2016. Structure-from-Motion Revisited. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 4104-4113.
Shan, Q., Wu, C., Curless, B., Furukawa, Y., Hernandez, C., Seitz, S.M., 2014. Accurate Geo-Registration by Ground-to-Aerial Image Matching, 2014 2nd International Conference on 3D Vision, pp. 525-532.
Sun, J.M., Shen, Z.H., Wang, Y.A., Bao, H.J., Zhou, X.W., 2021. LoFTR: Detector-Free Local Feature Matching with Transformers. CVPR 2021, 8918–8927.
Sun, C., Shrivastava, A., Singh, S., Gupta, A., 2017. Revisiting unreasonable effectiveness of data in deep learning era, IEEE international conference on computer vision, pp. 843-852.
Vo, A.V., Linh, T.H., Laefer, D.F., Bertolotto, M., 2015. Octree-based region growing for point cloud segmentation. ISPRS J. Photogramm. Remote Sens. 104, 88–100.
White, A.G., Zink, A., Codecá, L., Clarke, S., 2021.A. digital twin smart city for citizen feedback. Cities 110, 103064.
Wolff, K., Kim, C., Zimmer, H., Schroers, C., Botsch, M., Sorkine-Hornung, O., Sorkine-Hornung, A., 2016. Point Cloud Noise and Outlier Removal for Image-Based 3D Reconstruction. Proceedings of 2016 Fourth International Conference on 3d Vision (3DV), 118-127.
Wu, B., 2021. Photogrammetry for 3D Mapping in Urban Areas. Springer Singapore, pp. 401-413.
Wu, B., Zhang, Y.S., Zhu, Q., 2012. Integrated point and edge matching on poor textural images constrained by self-adaptive triangulations. ISPRS J. Photogramm. Remote Sens. 68, 40–55.
Wu, B., Xie, L., Hu, H., Zhu, Q., Yau, E., 2018. Integration of aerial oblique imagery and terrestrial imagery for optimized 3D modeling in urban areas. ISPRS J. Photogramm. Remote Sens. 139, 119–132.
Yao, Y., Luo, Z.X., Li, S.W., Fang, T., Quan, L., 2018. MVSNet: Depth Inference for Unstructured Multi-view Stereo. Computer Vision - ECCV 2018. Pt Viii 11212, 785–801.
Ye, L., Wu, B., 2018. Integrated Image Matching and Segmentation for 3D Surface Reconstruction in Urban Areas. Photogramm. Eng. Remote Sens. 84, 135–148.
Zheng, J., Ramasinghe, S., Lucey, S., 2021. Rethinking positional encoding. arXiv preprint arXiv:2107.02561.
Zhu, Q., Li, Y., Hu, H., Wu, B., 2017. Robust point cloud classification based on multi-level semantic relationships for urban scenes. ISPRS J. Photogramm. Remote Sens. 129, 86–102.
Zhu, Q., Wang, Z., Hu, H., Xie, L., Ge, X., Zhang, Y., 2020. Leveraging photogrammetric mesh models for aerial-ground feature point matching toward integrated 3D reconstruction. ISPRS J. Photogramm. Remote Sens. 166, 26–40.