Learning Dense Consistent Features for Aerial-to-Ground Structure-From-Motion

Hongjie Li, Aonan Liu, Xiao Xie, Han Guo, Hanjiang Xiong, and Xianwei Zheng, *Member, IEEE*

***Abstract*—The integration of aerial and ground images is known to be effective for enhancing the quality of 3-D reconstruction in complex urban scenarios. However, directly applying the structure-from-motion (SfM) technique for unified 3-D reconstruction with aerial and ground images is particularly difficult, due to the large differences in viewpoint, scale, and appearance between those two types of images. Previous studies mainly rely on viewpoint rectification or view rendering/synthesis to improve the feature matching quality for aligning the aerial and ground models. Nevertheless, these approaches still fail to address the inherent information differences between aerial and ground images. In this article, we propose a learning-based matching framework for direct SfM with ground and aerial images. The key idea of our method is to learn the pixel-wise consistent features between aerial and ground images to handle the large heterogeneity of these two types of images. Specifically, we deploy a learning-based matching framework to robustly correspond the aerial and ground images. With the high-quality feature matching, learned feature maps are used for refining keypoint locations and fusing featuremetric error into bundle adjustment with the consideration of geometric error, both of which can further improve the accuracy and completeness of the recovered 3-D scene. Extensive experiments conducted on six datasets demonstrate that the proposed method can reconstruct high-fidelity 3-D models with direct aerial-to-ground SfM, which cannot be achieved by existing methods. In addition, our method also shows outstanding performance in subtasks of feature matching and point cloud recovery.**

***Index Terms*—Aerial-ground integration, dense consistent features, feature map-based bundle adjustment (BA), keypoint location refinement (LR), structure-from-motion (SfM).**

Manuscript received 8 March 2023; revised 30 April 2023; accepted 9 May 2023. Date of publication 23 May 2023; date of current version 12 June 2023. This work was supported in part by the Open fund of Key Laboratory of Urban Land Resources Monitoring and Simulation, Ministry of Natural Resources under Grant KF202106084, in part by the National Natural Science Foundation of China Project under Grant 42071370, and in part by the Fundamental Research Funds for the Central Universities of China under Grant 2042022dx0001. *(Corresponding authors: Han Guo; Hanjiang Xiong.)*

Hongjie Li and Xianwei Zheng are with the Key Laboratory of Urban Land Resources Monitoring and Simulation, Ministry of Natural Resources, Shenzhen and the State Key Laboratory LIESMARS, Wuhan University, Wuhan 430000, China (e-mail: lihongjie@whu.edu.cn; zhengxw@whu.edu.cn).

Aonan Liu and Hanjiang Xiong are with the State Key Laboratory LIESMARS, Wuhan University, Wuhan 430000, China (e-mail: liuaonan@whu.edu.cn; xionghanjiang@whu.edu.cn).

Han Guo is with the Key Laboratory of Urban Land Resources Monitoring and Simulation, Ministry of Natural Resources, Shenzhen 518000, China (e-mail: guohan@whu.edu.cn).

Xiao Xie is with the Key Laboratory for Environmental Computation and Sustainability of Liaoning Province, Institute of Applied Ecology, Chinese Academy of Sciences, Shenyang 110016, China (e-mail: xiexiao@iae.ac.cn).

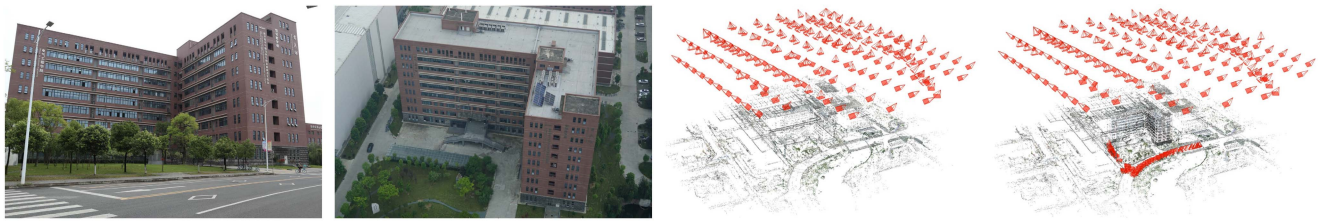
Digital Object Identifier 10.1109/JSTARS.2023.3279199

I. Introduction

With the increasing availability of aerial oblique images, the fast reconstruction of urban scenes has become feasible. However, challenges remain in achieving high-quality urban 3-D models due to occlusion and low resolution caused by the limitations in height and perspective. These issues often result in geometric holes and blurred textures in reconstructed models, particularly on building facades. To address these challenges, recent studies have explored the integration of aerial and ground images as a promising approach for improving the quality of urban 3-D reconstructions [1], [2]. As depicted in Fig. 1, ground images provide close and arbitrary views of the scene, which can complement aerial images to capture details and ensure completeness in the reconstructed model.

The key to joint reconstruction of aerial and ground images lies in accurately registering the 3-D data (e.g., point cloud and mesh) produced by each image source to the same coordinate system and ensuring geometric consistency constraints. Feature matching is a primary strategy that can precisely establish connection relationships between different images by extracting and matching features. However, commonly used handcrafted 2-D features (such as scale-invariant feature transform (SIFT) [3] and affine-SIFT (ASIFT) [4]) cannot tolerate well the heterogeneity of aerial and ground images in terms of viewpoint, lighting, and appearance, making it difficult to find enough matching points to support the effective operation of various components in the structure-from-motion (SfM) [5], such as triangulation and bundle adjustment (BA). Apart from 2-D feature matching, 3-D features from separate 3-D models prebuilt with different image sources are also often used for model matching and alignment, such as pin images (SI) [6], fast point feature histogram [7], rotational projection statistics [8], and so on. Nevertheless, the differences in accuracy, density, and noise level between the aerial and ground models of the same scene still make it difficult to yield satisfactory fusion results.

In order to address the problem of 2-D/3-D feature-based registration difficulties between aerial and ground images or models, some researchers have conducted pioneering studies using viewpoint rectification [9], [10] or rendering/synthesis [1], [2], [11] to improve feature matching performance and achieve registration of aerial and ground models. Viewpoint rectification works focus on identifying view-independent planar structures (such as ground and building facades) in the scene to correct the aerial and ground images to a normalized viewpoint, thus reducing the differences in viewpoint between the aerial and ground



(a) (b) (c) (d)

Fig. 1. Example of recovering both point cloud and camera poses simultaneously from aerial and ground images using SfM, based on SWJTU-BLD (a general building of the Southwest Jiaotong University (SWJTU)) data provided by [1]. (a) Ground image. (b) Aerial image. (c) Aerial Sparse point cloud. (d) Aerial-ground sparse point cloud.

images. However, the structure of urban scenes is usually complex, and it is difficult to guarantee that effective planar features can be extracted even from building facades, so the practicality of this method is often limited. On the other hand, viewpoint rendering/synthesis involves using recovered 3-D data (such as depth maps, point clouds, and meshes) to synthesize a new image at a target viewpoint and match it with the target image (often by projecting ground data onto the aerial viewpoint). Unlike the viewpoint rectification, this strategy combines information from multiple-view images to reduce the differences between aerial and ground images. However, successful implementation of this strategy relies on accurate GPS or geotagged labels for the rough registration of the 3-D data. This can be challenging when the data comes from crowd-sourced databases or when GPS information has high levels of error. In general, although the above two strategies have made some progress, the heterogeneous differences between aerial and ground images still hinder high-quality joint reconstruction with aerial-ground information.

In this article, we propose a dense correspondence learning based SfM approach for the integration of aerial and ground images. Inspired by current developments in deep learning for optical flow estimation [12], [13], the proposed method addresses the heterogeneous differences between aerial and ground images by learning pixel-wise consistent features. Based on dense consistent features, our method can accurately establish correspondences between 2-D local features (subpixel level) preextracted in the images. Furthermore, we propose a multi-view feature consistency refinement to adjust the position of 2-D local features in each track, to overcome the low feature localization accuracy caused by differences in aerial and ground scales and obtain an optimized scene graph. Different from the existing studies on the integration of aerial and ground images, this article not only focuses on solving the feature matching problem existing between aerial and ground images, but also further proposes a feature map-based BA method, which further improves the accuracy of 3-D scene recovery with consideration of both featuremetric error (FE) and geometric error.

In summary, our main contribution is the proposal of a reliable method for integrating aerial and ground images, which mines consistent features from aerial and ground images with extreme information differences by a learning manner to achieve complete and refined sparse SfM point clouds.

The rest of this article is organized as follows. Section II reviews related works. Section III describes the proposed method in detail. In Section IV, we present experiments that validate the performance of our method. Finally, Section V concludes this article.

II. Related Work

Here, we review the works related to 3-D reconstruction based on the integration of aerial and ground images. Specifically, the review is organized into the following three parts: 1) feature matching; 2) viewpoint rectification and 3) view rendering/synthesis.

A. Feature Matching

The first step of image-based 3-D reconstruction methods is usually to extract and match 2-D local features between images. Commonly used 2-D features, such as SIFT [3], ASIFT [4], and oriented FAST and rotated BRIEF (ORB) [14], are useful in handling general baseline scenes but ineffective to cope with matching tasks with extremely wide baseline, such as feature matching between aerial and ground images. Therefore, some studies have proposed self-similarity descriptors based on a simple observation that urban building facades often exhibit high self-similarity to achieve matching between aerial and ground images [15], [16], [17]. However, building facades in cities are usually complex, making it difficult to guarantee that reliable self-similarity features are captured. Additionally, these studies lack the establishment of pixel-level correspondences between images, making it impossible to provide effective inputs for the SfM algorithm. Some researchers pay attention to outlier rejection to handle aerial-ground image matching task by introducing matching priors. These outlier rejection approaches improve the robustness of the feature matching algorithm to high outlier rates and can effectively mine the inliers from cluttered matches. For example, Line et al. [18] proposed to separate outliers by learning the bilateral function (BF) from candidate matches, based on the assumption that the correct matches are consistent in density, smoothness, and spatial distribution. In order to eliminate the incorrect matches brought by repetitive structures, Lin et al. [19] further proposed RepMatch to incorporate random sample consensus (RANSAC [20]) into the BF. Zheng et al. [21] considered that small local areas in real scenes cannot be simply considered as planes, and thus designs the local affine validation (LAV), which eliminates outliers by solving smoothly varying affine functions in small local areas. To address the extreme

scale differences between aerial and ground images, Zhou [22] studied a scale-space-based scale-invariant matching algorithm based on the assumption that the scale ratio of correctly matched feature pairs is close to the image scale ratio. The method introduced by [22] first estimates the image scale ratio based on bag-of-features encoding, and then achieves scale-aware image matching with the estimated scale ratio. In contrast, our method works in a learning-based way to extract features that are invariant to heterogeneous differences such as viewpoint, scale, and illumination between aerial and ground images, so as to achieve accurate matching of aerial-ground images.

B. Viewpoint Rectification

Existing viewpoint rectification based methods use the geometric priors of a given scene to correct the aerial and ground images to a normalized view, thus improving feature matching performance. Typically, the first step is to detect planar structures in the scene that are invariant to viewpoint changes, such as building facades or the ground, and assume that these structures are the same in all images. By projecting all images onto these planes, the viewpoint differences in image data can be reduced [23], [24]. However, in many cases, it is not possible to find planar structures that are visible in all images. In such cases, some studies resort to performing viewpoint rectification for pairs of images based on the planar structures detected in each pair. Wu et al. [9] corrected the images by projecting them onto virtual planes generated from dense point cloud data. Zheng et al. [10] first extracted the building façade structures in the aerial and ground images using the local consistency of features, then verifies images based on the transform-invariant low-rank texture [25], and finally achieves aerial-ground image matching by a mutually supervised manner between extracted façade grid structures and matched seeds. However, the viewpoint rectification approach suffers from the following two problems: 1) even if the viewpoint of both images is successfully verified, the information difference between the aerial and ground images is not handled, and the further scale change brought by the viewpoint rectification will exacerbate the difficulty of feature matching and 2) real scenes are often complex, and it is difficult to guarantee the extraction of planar structures, especially when the building facades are nonplanar.

C. View Rendering/Synthesis

The method of viewpoint rendering/synthesis often uses a coarse-to-fine strategy. First, the coarse registration of the aerial and ground models is achieved using the geotagging of the images. Then, viewpoint rendering/synthesis techniques are used to generate a synthesized image in the target viewpoint, which is then matched with the target image using feature matching. The matched 2-D features are then back-projected to establish 3-D correspondences between the aerial and ground models, and finally, the similarity matrix is estimated to achieve the registration of the aerial and ground models. Considering the low resolution of aerial data, view synthesis often takes the aerial view as the target viewpoint. The view synthesis can thus be realized by using depth maps to warp the ground images [11] or project the dense point cloud visible by ground images to the aerial viewpoint [2]. Compared to the former generation method, the latter incorporates information from multiview images. To avoid holes in the synthetic images, [26] explored to generate synthetic images using spatially continuous meshes generated from ground sparse point clouds. In addition, unlike previous studies, they merged point clouds by BA method instead of estimating the transformations between models, which aims to deal with possible scene drift issues. Zhu et al. [1] inferred synthetic images from ground view using a mesh model recovered from aerial data, and also generated depth and normal maps to tackle the problem of inaccurate feature correspondence caused by the low mesh geometric accuracy and texture blending. In a word, the view rendering/synthesis based approach relies on the GPS information or text labels for the initial coarse alignment of aerial and ground models, which limits the applicability of these methods, especially when the images come from a multisource database without localization information or when the GPS information is inaccurate. Moreover, these approaches still do not address the heterogeneous differences between aerial and ground images, even though they tackle the inconsistency in viewpoint and scale to some extent.

III. Method

Given N^g ground images and N^a aerial images, our goal is to geometrically register these images, and recover high-quality 3-D camera information (intrinsic and extrinsic parameters) and scene structure. The main challenges lie in the significant variations between aerial and ground images about viewpoint, lighting, weather condition, and resolution, as well as the potential for significant occlusion and noise. These factors often result in feature matching failures when attempting to construct a complete scene graph from the set of aerial and ground images. Even in some cases where feature matching appears to be successful, SfM may still generate inaccurate 3-D information. In this work, we make it possible to jointly model complete and detailed scenes using SfM directly by learning consistent features between aerial and ground images. Specifically, we first design a dense correspondence network to learn consistent features among ground and aerial images and generate dense correspondences. To enhance generalization to ground and aerial images, we adopt a multiscale and multistage inference strategy to output high-quality correspondences. We then extract sparse keypoints from each image and establish correspondences between keypoints in the image pairs based on the outputted dense correspondences. To ensure the quality of keypoint matches input to the SfM pipeline, we further use the learned feature map to adjust the locations of the keypoints in the multiview images. Finally, based on the learned feature map, we introduce a new feature consistency error into BA, which effectively eliminates cumulative errors and improves the quality of the recovered 3-D information. The overall pipeline is illustrated in Fig. 2. It should be noted that the fundamental structure from the motion algorithm used in our method is provided by [5].

A. Dense Correspondence Network

Typically, SfM requires a matching graph based on sparse 2-D keypoint correspondences for geometric estimation. However, traditional handcrafted sparse features such as SIFT and ASIFT

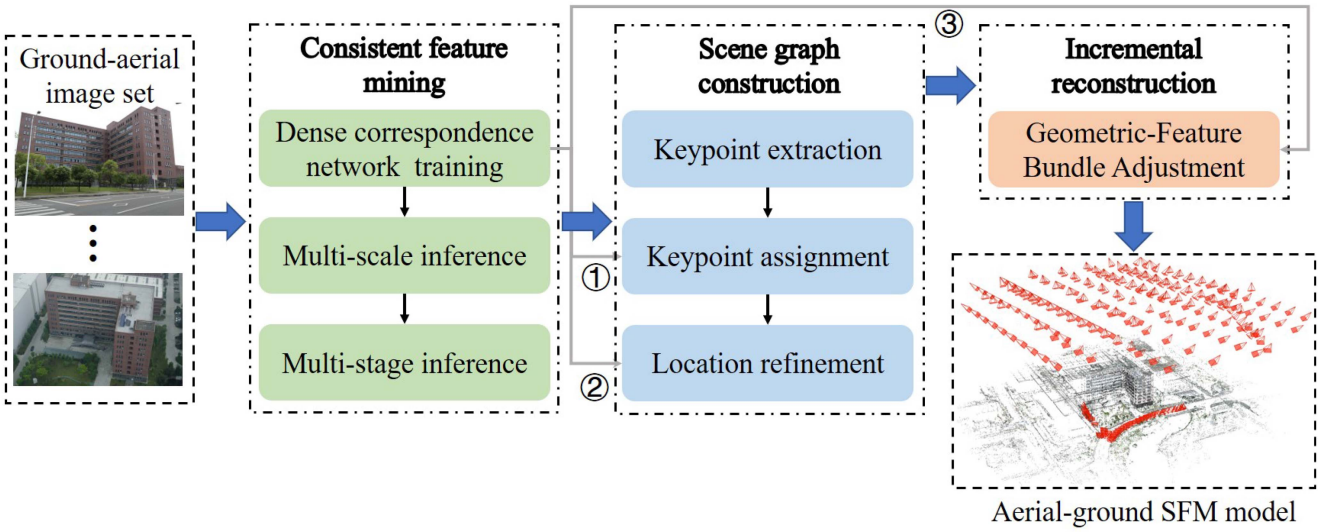


Fig. 2. Proposed aerial-ground scene reconstruction pipeline. Pipeline mainly consists of three parts: consistent feature mining, scene graph construction, and incremental reconstruction. Gray arrows indicate the transfer of intermediate or final outputs from the dense correspondence network for use in subsequent steps. Path ① represents the transmission of dense correspondences, while paths ② and ③ represent the transmission of feature maps.

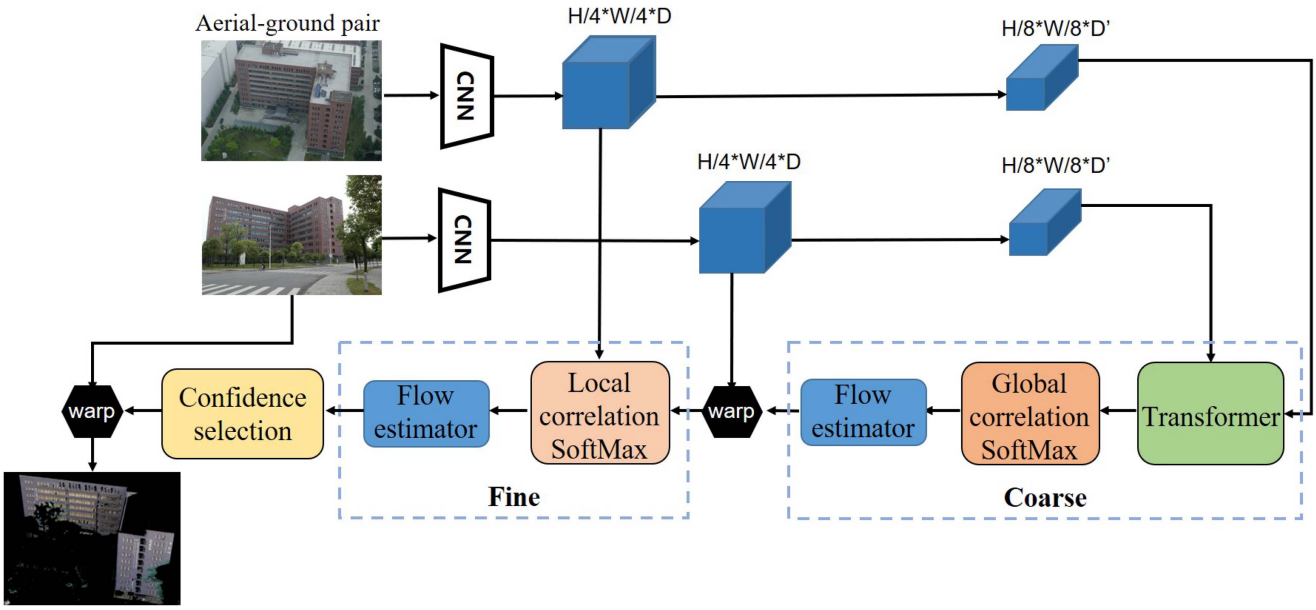


Fig. 3. Structure of the dense correspondence network.

are inadequate for establishing high-quality keypoint correspondences between aerial and ground images. This is mainly due to the sensitivity of the keypoint extractor to changes in image conditions such as viewpoint, scale, and illumination [27], making it difficult to ensure high repeatability of keypoints extracted from aerial-ground image pairs. In addition, the invariance of existing feature descriptors is also insufficient to cope with significant changes in the conditions of aerial and ground images, resulting in the failure of matching strategies based on feature similarity. Although learning-based methods can effectively increase robustness to image condition changes, the large receptive field of convolutional neural networks (CNNs) and down-sampling of feature maps often result in the lower location accuracy of learned keypoints compared to traditionally handcrafted keypoints, which greatly affect the accuracy of geometric estimation [28]. Therefore, instead of directly utilizing the traditional detect-describe-match pipeline to establish keypoint correspondences between aerial and ground images, we first propose a dense correspondence network to learn consistent features between aerial and ground images and establish pixel-level correspondences. Then, in Section III-B 1, dense correspondences are used to perform handcrafted keypoint matching. The idea behind this strategy is that we believe matching pixel by pixel is more accurate than directly using descriptors to match sparse keypoints.

The structure of the proposed dense correspondence network is illustrated in Fig. 3. We first extract two feature maps

$\{F^{1/2} \in \mathbb{R}^{H/2\times W/2\times D}, F^{1/4} \in \mathbb{R}^{H/4\times W/4\times D'}\}$ with $1/2$ and $1/4$ spatial resolution of the original size from the reference image and source image, respectively, where H, W denote the height and width of the input image, respectively, and D, D' denote the channel dimensions of $F^{1/2}$ and $F^{1/4}$. Considering the extreme differences in image conditions between ground and aerial images, we deploy a transformer block [29] to enlarge the receptive field of the CNN and ensure that each pixel can receive full-image contextual information, thus enhancing the discriminability of the features. The design of our dense correspondence network takes a coarse-to-fine strategy, as shown in Fig. 3. In the coarse phase, a global correlation SoftMax layer is used to construct pixel-level similarity between the source and reference feature maps $\{F_s^{1/4}, F_r^{1/4}\}$. The global correlation SoftMax can be defined by the following formulation:

$$C^{1/4}\left(\mathrm{x}_s^i, \mathrm{x}_r^j\right) = SoftMax\left(\frac{F_s^{1/4}\left(\mathrm{x}_s^i\right)\left(F_r^{1/4}\left(\mathrm{x}_r^j\right)\right)^T}{\sqrt{D'}}\right) \tag{1}$$

where x_s^i and x_r^j are the coordinates in the source and reference feature maps, respectively. The resulting correlation volume $C^{1/4} \in \mathbb{R}^{H/4\times W/4\times H/4\times W/4}$ is then fed into a flow estimator to output the coarse flow f_c. In the fine phase, the source feature map $F_s^{1/2}$ is warped to the reference viewpoint based on the coarse flow f_c, and a local correlation softmax layer is used to construct a local correlation volume $C^{1/2}$ between the warped feature map $F_{s(w)}^{1/2}$ and the reference feature map $F_r^{1/2}$. The local correlation softmax can be defined as

$$C^{1/2}\left(\mathrm{x}_r^i, \mathrm{d}\right) = SoftMax\left(\frac{F_{s(w)}^{1/2}(\mathrm{x}_r^i + \mathrm{d})\left(F_r^{1/2}(\mathrm{x}_r^i)\right)^T}{\sqrt{D}}\right) \tag{2}$$

where x_r^i represents the ith coordinate in the reference feature map, while d denotes the offset vector. The local region used to compute the pixel-wise similarity is determined by the radius ($|d| < 4$). Finally, the resulting volume $C^{1/2} \in \mathbb{R}^{H\times W\times(2d+1)^2}$ is fed into another flow estimator to output the residual flow f_r. The final refined flow f_f is obtained by combining the coarse flow f_c with the residual flow f_r. The final flow can accurately generate pixel-to-pixel correspondences (referred to as dense correspondences) between the images. Additionally, a confidence selection module is used to select reliable correspondences with correlation values exceeding a specific threshold as the final output. It should be noted that the confidence selection module is only used in the inference stage. Specifically, after forward pass through the network, we calculate the correlation values of pixel-to-pixel correspondences on the feature maps $F^{1/2}$ and $F^{1/4}$, and set the threshold to 0.8. Further details of the network design and training can be found in Section IV-B 1.

Furthermore, we believe that accurate matching between aerial and ground images is challenging to achieve through a single forward pass of the dense correspondence network. Therefore, a multiscale and multistage inference strategy is used to handle the large viewpoint and scale differences between aerial and ground images. For the multiscale inference, we adopt the idea in [30] to resize the ground images into four different resolutions of 0.5, 0.6, 0.88, and 1, resulting in four image pairs with the aerial images $\{(I_{0.5}^g, I_1^a), (I_{0.6}^g, I_1^a), (I_{0.88}^g, I_1^a), (I_1^g, I_1^a)\}$. Each of these image pairs is fed into the network separately to obtain the corresponding dense correspondence results. The resulting four sets of correspondences are then passed through RANSAC to solve for the homography matrix, and the final correspondences are determined based on the ratio of inliers. For the multistage inference, we follow a coarse-to-fine design similar to the network architecture. In the first forward pass of the network, we can obtain the coarse dense correspondence results and estimate the homography matrix by using RANSAC. We then use this matrix to warp the source image into reference viewpoint and obtain a roughly aligned image pair. The second input of the new image pair is fed into the network to generate a new flow, which can be considered as the residual flow. This flow is added to the flow obtained from the first forward pass. The combined flow results are then converted into correspondences and output as the final result.

B. Scene Graph Construction

The scene graph describes the connectivity between images, where each image is a node and there is an edge between any pair of images with matched keypoints. The set of matched keypoints across multiple views forms a track. As the input to SfM, the quality of the scene graph directly determines the quality of 3-D camera and scene information recovery. To construct a high-quality scene graph, we adopt the following steps to provide the necessary connectivity for recovering the complete model, and sufficient redundancy and accurate initial values for reliable estimation.

1) Keypoint extraction and assignment. Although dense correspondences outputted by the dense correspondence network establish pixel-level matches between images, they are insufficient for accurately estimating 3-D geometry and are limited by viewpoint and resolution, often resulting in many-to-one pixel matches. Therefore, dense correspondences cannot be directly applied to SfM. To address these issues, we first extract sparse keypoints (such as SIFT and SuperPoint [31]) from each image and establish rough matching relationships between keypoints based on dense correspondences, which can be viewed as candidate matches. Although these matches also inherit the many-to-one disadvantage of dense correspondences, they have higher keypoint localization accuracy. Additionally, we believe that the sparsity of keypoints ensures that the probability of multiple keypoints within the same pixel is low, allowing the correspondence problem of multiple keypoints in one matching pixel pair to be ignored. The outliers included in these candidate matches can be handled by RANSAC or other outlier rejection algorithms, such as LAV [21].

2) Refinement of keypoint locations. Usually, sparse keypoints are independently extracted on each image. Therefore, when there are significant variations in image conditions, it is difficult to ensure positional consistency of the matched keypoints across multiple images, resulting in a decrease in the accuracy of both scene structure and camera pose estimation. In order to tackle this problem, we propose a method that refines the keypoint locations based on multiview feature maps. This method adjusts the keypoint locations by minimizing the FEs between matched keypoints in the multiview feature maps. Specifically, the method works as follows:

$$\begin{aligned}\mathcal{P} &= \operatorname{argmin} \sum_{j=1}^{N_t} \sum_{i=1}^{N_j} e_{i,j} \\ &= \operatorname{argmin} \sum_{j=1}^{N_t} \left(\eta \sum_{i=1}^{N_j^g} \|F_i[p_{i(j)}] - F_k[p_{k(j)}]\| \right. \\ &\quad \left. + (1-\eta) \sum_{i=1}^{N_j^a} \|F_i[p_{i(j)}] - F_k[p_{k(j)}]\| \right) \end{aligned} \tag{3}$$

where $\mathcal{P} = \{p_{i(j)}\}_{i=1,\ldots,N_j}^{j=1,\ldots,Nt}$, N_t and N_j ($N_j = (N_j^a + N_j^g) < N = (N^a + N^g)$) refer to the number of tracks and the number of keypoints on jth track, respectively. N_j^a and N_j^g denote the number of aerial images and ground images involved in the jth track, respectively. $p_{i(j)}$ is the ith keypoint on the jth track, F represents the feature map learned from the dense correspondence network (It is noted that we use the feature map $F^{1/2}$ with 1/2 spatial resolution of the original image.). $[\cdot]$ is the sampling operator. We select the keypoint $p_{k(j)}$ with the most matching relationships on the jth track as the anchor point to adjust the locations of other keypoints $\mathcal{P} = \{p_{i(j)}\}_{i=1}^{N_j}, i \neq k$. Taking into account the resolution difference between aerial and ground images, we utilize γ to regulate the influence of keypoint offset on both types of images towards the overall loss. We adopt the Levenberg–Marquardt (LM) algorithm [32] to solve (3) and optimize the location estimation in each iteration as follows:

$$\Delta\mathcal{P}^* = \operatorname{argmin} \|J(\mathcal{P})\Delta\mathcal{P} + E(\mathcal{P})\| + \lambda \|D(\mathcal{P})\Delta(\mathcal{P})\| \tag{4}$$

where $E(\mathcal{P}) = [e_{i,j}]_{i=1,\ldots,N_j}^{j=1,\ldots,N_t}$, $J(\mathcal{P})$ is the Jacobian matrix of $E(\mathcal{P})$, $D(\mathcal{P})$ is a nonnegative diagonal matrix consisting of the square root of the elements on the diagonal of $J(\mathcal{P})^T J(\mathcal{P})$, and $\lambda > 0$ controls the degree of regularization.

C. Geometric-Feature BA

To mitigate the cumulative errors during the incremental reconstruction, it is necessary to perform BA after image registration and triangulation to guarantee the accuracy of 3-D scene estimation. Specifically, given an initial estimation, BA refines the estimation of the scene point and camera pose by minimizing the following reprojection error (RE):

$$\begin{aligned}\mathcal{X} &= \operatorname{argmin} \sum_{j=1}^{N_t} \sum_{i=1}^{N_c} d_{i,j} \\ &= \operatorname{argmin} \sum_{j=1}^{N_t} \sum_{i=1}^{N_c} \left\| \prod (R_i P_j + T_i, C_i) - p_{i(j)} \right\| \end{aligned} \tag{5}$$

where $\mathcal{X} = \left\{ \{R_i, t_i\}_{i=1}^{N_c}, \{P_j\}_{j=1}^{N_t} \right\}$ and N_c is the number of images involved in 3-D reconstruction; $\prod(\cdot)$ is the function that projects the 3-D scene point to 2-D plane; P_j is the jth scene point; C_i and $\{R_i, T_i\}$ are the intrinsic parameter and pose of ith camera, respectively. For better performance of BA, we utilize the image feature maps ($\{F_i^{1/2}\}, i = 1, \ldots, N_c$) from the dense correspondence network and introduce a novel feature-based BA

$$\begin{aligned}\mathcal{X} &= \operatorname{argmin} \sum_{j=1}^{N_t} \sum_{i=1}^{N_c} f_{i,j} \\ &= \operatorname{argmin} \sum_{j=1}^{N_t} \sum_{i=1}^{N_c} \left\| F_i \left[\prod (R_i P_j + t_i, C_i) \right] - F_i \left[p_{i(j)} \right] \right\|. \end{aligned} \tag{6}$$

The final objective function for BA minimizes both geometric and featuremetric consistency error, which is formulated as follows:

$$\begin{aligned}\mathcal{X} = \operatorname{argmin} \sum_{j=1}^{N_t} \left(\eta \sum_{i=1}^{N_c^g} (d_{i,j} + f_{i,j}) \right. \\ \left. + (1-\eta) \sum_{i=1}^{N_c^a} (d_{i,j} + f_{i,j}) \right) \end{aligned} \tag{7}$$

where N_c^a and N_c^g denote the number of aerial images and ground images involved in 3-D reconstruction, respectively. Here, we also use a parameter η to balance the impact of errors on aerial and ground images towards the overall loss.

The solution of (7) is the same as that of (3), which also utilizes the LM algorithm to iteratively update the parameters for optimization

$$\Delta\mathcal{X}^* = \operatorname{argmin} \|J(\mathcal{X})\Delta\mathcal{X} + E(\mathcal{X})\| + \lambda \|D(\mathcal{X})\Delta(\mathcal{X})\| \tag{8}$$

where $E(\mathcal{X}) = [e_{i,j}]_{i=1,\ldots,N_c}^{j=1,\ldots,N_t}, e_{i,j} = d_{i,j} + f_{i,j}$, $J(\mathcal{X})$ is the Jacobian matrix of $E(\mathcal{X})$, $D(\mathcal{X})$ is a nonnegative diagonal matrix consisting of the square root of the elements on the diagonal of $J(\mathcal{X})^T J(\mathcal{X})$, and $\lambda > 0$ controls the degree of regularization.

IV. Experiments

To evaluate the effectiveness of our proposed method in the task of aerial-ground image integration, we conduct a series of experiments on multiple datasets that are both publicly available and collected by ourselves. Firstly, we compare the performance

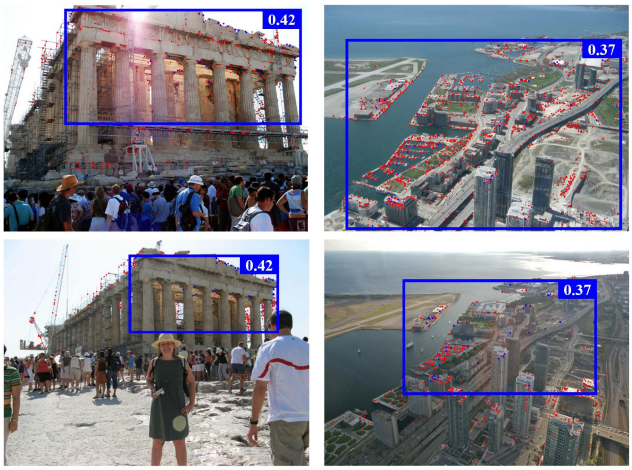


Fig. 4. Two examples of calculating the scale ratio between images in MegaDepth. (a) Image pair 1. (b) Image pair 2.

of our proposed method with state-of-the-art techniques in feature matching. Secondly, we evaluate the quality of the reconstructed 3-D scenes by integrating aerial and ground images. Finally, we demonstrate the impact of our proposed method on the reconstruction of complete and fine-grained surface models. The comparative results with prior arts on datasets are reported in Sections IV-C, IV-D, and IV-E.

A. Dataset

1) Training Data: We use the MegaDepth [33] to train the proposed dense correspondence network. MegaDepth consists of images from photo-tourism with significant variations in appearance and viewpoint, which can simulate the differences between aerial and ground images. The authors use COLMAP to reconstruct 196 different scenes from 1 070 468 internet photos and provide the intrinsic and extrinsic camera parameters and depth maps for 102 681 images among them. Before training, we preprocess the MegaDepth in the following steps.

1) Removing the scenes with low quality depth maps as indicated by [34].
2) (Obtaining image pairs based on whether they have covisible points in the sparse SfM point cloud.
3) Obtaining the 2-D projections of sparse SfM point cloud on images and generating their bounding boxes as their range of distribution on each image. The area ratio of bounding boxes between image pairs is considered as the scale ratio, as shown in Fig. 4.
4) The scale ratios of all image pairs are enumerated in the range of [0.1–0.7], and each scene is divided into multiple subsets according to the scale ratio range in [0.1–0.3], [0.3–0.5], [0.5–0.7].

In order to obtain the ground truth correspondences between image pairs, we project all the points in the source image with depth information into the reference image to obtain coarse correspondences. Subsequently, a depth-check is performed to reject the incorrect correspondences and obtain the final correspondence ground-truth and the mask for the effective loss calculation, as proposed in [34]. It should be noted that the correspondences can be converted into the ground-truth of flow for training dense correspondence network, as done in [35].

2) Test Data: Six datasets are used to evaluate the proposed method, including the ISPRS benchmark dataset collected in the Centre of Dortmund and Zeche of Zurich [36], two datasets (SWJTU-LIB and SWJTU-BLD) collected on the campus of Southwest Jiaotong University (SWJTU) provided by [1], and two datasets (CQ-BAISHA and CQ-StudioCity) collected by ourselves in Baisha Town and a studio city in Chongqing (CQ). The ground sampling distance (GSD) of all images ranged from 0.16 to 1.8 cm. Table I describes the specific details of the six aerial-ground datasets, Fig. 5 shows examples of aerial-ground image pairs, and Fig. 6 shows the scenes reconstructed by SfM from the aerial and ground images separately. In order to obtain ground truth correspondences for quantitative evaluation, we manually select tie points in covisible multiview images to integrate aerial and ground images.

B. Implementation Details

1) Dense Correspondence Network: We implement our network using PyTorch. The backbone of our network is ResNet50 [37], which is pretrained on ImageNet [38]. In the network, we introduce the transformer blocks to increase the global receptive field. However, the computational burden that comes with it cannot be ignored. Therefore, similar to [27], we use the linear transformer to address this issue. We only use one eight-head attention layer. For the flow decoder, we adopt the design provided by [39]. We follow [13] to supervise the training of the network by using the L1 distance between the predicted flow and the ground truth flow. Given the ground truth flow, we are able to calculate the following loss:

$$\mathcal{L} = \gamma_1 \left\| \mathrm{M}_c \left(\mathrm{f}_c - \mathrm{f}_c^{gt} \right) \right\|_1 + \gamma_2 \left\| \mathrm{M}_f \left(\mathrm{f}_f - \mathrm{f}_f^{gt} \right) \right\|_1 \quad (9)$$

where M_c and M_f refer to the ground truth masks at the coarse and fine stages of our network, respectively, while f_c^{gt} and f_f^{gt} refer to the ground truth flow at the coarse and fine stages, respectively. During our experiments, we set the values of γ_1 and γ_2 to 0.7 and 0.9, respectively.

For network training, we use a progressive strategy. Initially, we freeze all the weights of the backbone and train the remaining part of the network on a subset of the scene with a scale ratio of [0.5–0.7]. The learning rate is set to 10^{-4} during this stage. Once this training is completed, we unfreeze the weights of the backbone and fine-tune them on a subset with a scale ratio of [0.3–0.5]. The learning rate is set to 4×10^{-5} during this stage. Finally, to make the network adaptable to challenging scenarios, we train the entire network on the subset with the maximum scale ratio of [0.1–0.3], and set the learning rate to 10^{-6}. The model is trained on image pairs of size 520×520.

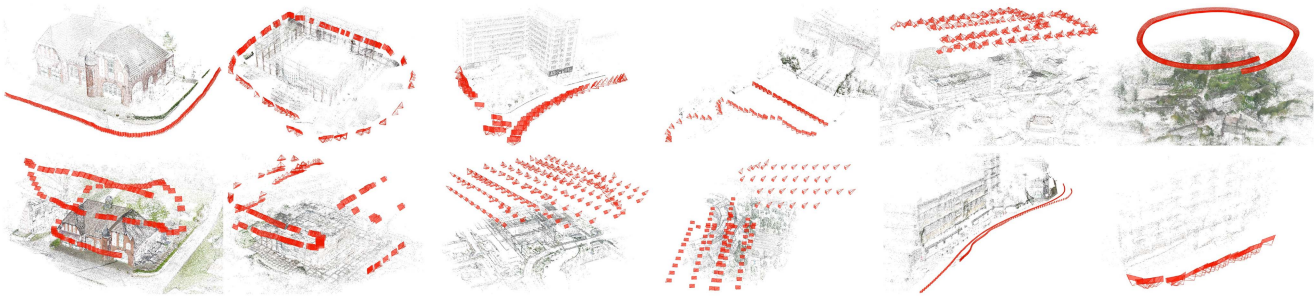
2) Keypoints LR and Optimized BA: In the process of refining keypoint locations and BA, we impose a maximum offset of $\beta = 10$ for each keypoint, and set $\eta = 0.3$ to focus the entire optimization process more on errors in aerial images. It is noteworthy that (4) and (8) reveal the similarity between the solving process of refining keypoint locations and that of BA.

TABLE I
DETAILED DESCRIPTION OF SIX AERIAL-GROUND DATASETS USED FOR EVALUATIONS

Datasets	Sensor		GSD(cm)		Images	
	Aerial	Ground	Aerial	Ground	Aerial	Ground
Zeche	SONY Nex-7	SONY Nex-7	0.56	0.28	172	147
Centre	SONY Nex-7	SONY Nex-7	1.10	0.53	146	204
SWJTU-LIB	SONY ICLE-5100	Cannon EOS M6	1.69	1.06	123	78
SWJTU-BLD	SONY ICLE-5100	Cannon EOS M6	1.93	1.33	207	88
CQ-BAISHA	SONY ICLE-5100	iPhone 11	0.62	0.08	171	505
CQ-StudioCity	SONY ICLE-5100	iPhone 11	1.26	0.19	374	84

(a) (b) (c) (d) (e) (f)

Fig. 5. Selected six image pairs from the six test datasets. The first and second rows show images from ground and aerial sets, respectively. (a) Zeche. (b) Centre. (c) SWJTU-LIB. (d) SWJTU-BLD. (e) CQ-BAISHA. (f) CQ-StudioCity.



(a) (b) (c) (d) (e) (f)

Fig. 6. Six sparse point cloud pairs generated from the six test datasets. The first and second rows show sparse point clouds reconstructed from ground and aerial images, respectively. (a) Zeche. (b) Centre. (c) SWJTU-LIB. (d) SWJTU-BLD. (e) CQ-BAISHA. (f) CQ-StudioCity.

TABLE II
QUANTITATIVE COMPARISON RESULTS OF DIFFERENT METHODS ON SIX PAIRS OF AERIAL-GROUND IMAGES ACCORDING TO THREE METRICS: PRECISION(P), RECALL(R), AND F1 SCORE(F) (UNIT:%)

Methods	Zeche			Centre			SWJTU-LIB			SWJTU-BLD			CQ-BAISHA			CQ-StudioCity		
	R	P	F	R	P	F	R	P	F	R	P	F	R	P	F	R	P	F
Colmap	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
AdaLAM	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
SIFT+SuperGlue	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Superpoint+SuperGlue	0.0	0.0	0.0	31.1	42.3	35.8	10.4	36.8	16.3	16.5	37.3	22.9	9.5	15.9	11.9	12.8	33.3	18.5
SIFT+Ours	68.6	100.0	81.4	82.7	100.0	90.5	0	0	0	25.0	52.2	33.8	69.4	100.0	82.0	56.8	100.0	72.4
Superpoint+Ours	85.6	93.9	89.5	87.9	96.7	92.1	62.7	85.7	72.4	69.5	89.9	78.4	78.4	98.3	87.2	48.7	73.1	58.5

TABLE III
QUANTITATIVE COMPARISON RESULTS OF DIFFERENT METHODS ON SIX DATASETS ACCORDING TO THREE METRICS: N_p, N_m, AND N_o

Methods	Zeche	Centre	SWJTU-LIB	SWJTU-BLD	CQ-BAISHA	CQ-StudioCity
	N_p / N_m / N_o	N_p / N_m / N_o	N_p / N_m / N_o	N_p / N_m / N_o	N_p / N_m / N_o	N_p / N_m / N_o
Colmap	9958 / 52 / 6568	26675 / 24 / 3156	2553 / 38 / 3464	- / - / -	- / - / -	- / - / -
AdaLAM	- / - / -	- / - / -	- / - / -	- / - / -	- / - / -	- / - / -
SIFT+SuperGlue	- / - / -	- / - / -	- / - / -	- / - / -	- / - / -	- / - / -
Superpoint+SuperGlue	6185 / 217 / 1651	3267 / 118 / 871	1813 / 331 / 1637	231 / 80 / 423	2256 / 115 / 1272	428 / 25 / 2738
SIFT+Ours	11248 / 384 / 8194	27342 / 171 / 3640	2736 / 316 / 3671	1557 / 109 / 1021	3467 / 168 / 1675	591 / 65 / 3567
SuperPoint+Ours	11350 / 412 / 8460	27100 / 145 / 3467	2910 / 367 / 3950	2456/ 145 / 1567	3112 / 135 / 1567	475 / 38 / 3342

"-" indicates that the corresponding method cannot successfully match aerial and ground images.

TABLE IV
QUANTITATIVE COMPARISON RESULTS OF DIFFERENT METHODS IN TERMS OF AVERAGE RE AND AVERAGE TL ON SIX DATASETS

Methods	Zeche		Centre		SWJTU-LIB		SWJTU-BLD		CQ-BAISHA		CQ-StudioCity	
	RE	TL	RE	TL	RE	TL	RE	TL	RE	TL	RE	TL
Colmap	1.07	7.21	1.04	4.74	0.58	6.76	-	-	-	-	-	-
RealityCapture	0.50	3.79	-	-	0.86	3.23	-	-	-	-	-	-
Superpoint+SuperGlue	1.52	4.24	1.42	3.67	1.45	4.27	1.41	3.51	1.32	5.03	1.64	8.44
SIFT+Ours ($w/o\,LF$, $w/o\,PE$)	0.63	6.23	0.54	4.37	0.64	5.07	0.85	2.49	0.72	5.26	0.59	4.54
SIFT+Ours ($w\,LF$,$w\,PE$)	0.47	7.36	0.48	4.96	0.53	6.78	0.55	3.18	0.58	5.78	0.42	5.58
SuperPoint+Ours ($w/o\,LF$, $w/o\,PE$)	0.68	6.83	0.62	3.85	0.67	5.62	0.89	2.31	0.85	5.17	0.62	4.20
SuperPoint+Ours ($w\,LF$,$w\,PE$)	0.52	7.63	0.53	4.67	0.54	6.84	0.57	3.28	0.63	5.94	0.47	5.43

"-" indicates that the corresponding method cannot successfully realize the integration of aerial-ground images.
"w" and "w/o" equal to "with" and "without", respectively.
"LF" denotes the operation of keypoint location refinement.
"PE" denotes the operation of introducing the featuremetric error into bundle adjustment.

However, the key difference between the two lies in the fact that refining keypoint locations is performed on a single track, which makes it amenable to acceleration via parallel computing. In contrast, in BA, all camera poses and scene points are optimized simultaneously.

C. Evaluation of Feature Matching Between Aerial and Ground Images

We compare our matching results with the following four advanced methods:

1) The feature matching method embedded in the Colmap system (Colmap) [5];
2) Adaptive locally-affine matching (AdaLAM) [40];
3) SIFT+SuperGlue [41];
4) Superpoint [31]+SuperGlue.

The first two methods use handcrafted feature extractors and outlier filters, the third method uses graph convolutional networks to learn correct matching relationships between features, and the fourth method incorporates a learned feature extractor on top of the graph convolutional network. It is worth noting that our matching results include the results with SIFT and SuperPoint as feature extractors. Additionally, besides using precision (P), Recall (R), and F1-score (F) to evaluate the quality of feature matching results for individual image pairs, we also count the number of aerial-ground image pairs matched, the average number of matches per pair, and the average number of 3-D points observable per image in six aerial-ground datasets to further evaluate the performance of our proposed matching method.

1) Evaluation on Two Views: We select one aerial-ground image pair from each of the six datasets for evaluation. The quantitative comparison of our proposed method and other methods in terms of precision, recall, and F1 score is presented in Table II, while the visual results of feature matching for each method on the six aerial-ground image pairs are shown in Fig. 7. From Table II, it can be observed that the classic SIFT feature and geometric verification method provided by Colmap are insufficient for matching aerial and ground images, resulting in zero scores in all three metrics. AdaLAM, an advanced outlier filter that uses local geometric verification to filter out outliers, surprisingly achieved the same results as COLMAP. This may be due to the following two reasons: 1) the scale, orientation, and other feature frame information of the SIFT features are not reliable for finding matches between aerial and ground images and 2) the small number of true SIFT matches between aerial and ground images is not enough to support local geometric verification. When the SIFT features are input into SuperGlue to learn correct matching relationships, effective matching results cannot be obtained for test image pairs. The results of the above three methods indicate that relying solely on SIFT descriptors is insufficient for aerial-ground image matching tasks under extreme differences.

When SuperPoint, a learned feature, is used instead of SIFT and combined with SuperGlue, there is a significant improvement in the results of aerial-ground image matching. This suggests that learned features perform better than traditionally designed handcrafted features in processing images with significant differences. Nevertheless, the SuperPoint + SuperGlue matching strategy still has many false matches, and even on the two pairs of images in the Zeche and CQ-StudioCity datasets, almost no correct matches are obtained. This may be related to the dataset and training strategy used for network training. On the other hand, using the results output by our dense correspondence network to assist SIFT and SuperPoint matching achieves higher precision, recall, and F1-score. It is worth noting that SIFT still fails to obtain true matches on the SWJTU-LIB image pair. This is because under extreme differences in aerial and ground images, SIFT features are challenging to ensure effective repeatability between images, making it difficult to obtain matches even with high-performance matchers. Additionally, the matches obtained using our method are more evenly distributed in space, as seen in Fig. 7.

2) Evaluation on MultiViews: To comprehensively evaluate the performance of our proposed method, we calculate three metrics on six datasets. The number of matched aerial-ground image pairs (N_p) reflects the robustness of the matching algorithm to differences between aerial and ground images. When the performance of all algorithms is similar on this metric, the average number of matches per aerial-ground image pair (N_m) is used to further evaluate the performance differences between different matching algorithms. The average number of 3-D points observed per image (N_o) is used to evaluate

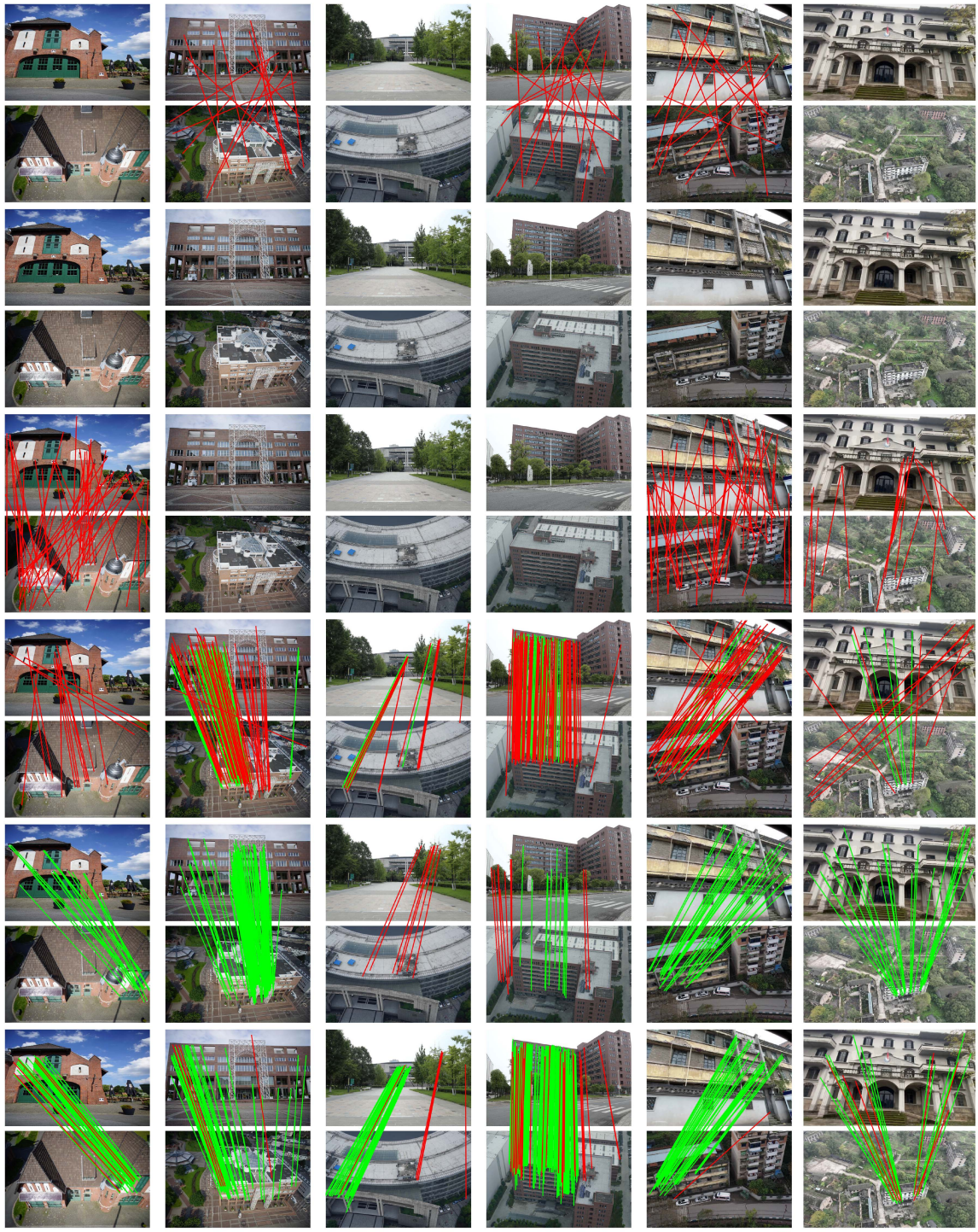


Fig. 7. Qualitative comparison results of our method and four advanced methods. Each row from top to bottom represents the corresponding matching results of Colmap, AdaLAM, SIFT+SuperGlue, SuperPoint + SuperGlue, SIFT+Ours, and SuperPoint + Ours, respectively. (a) Zeche. (b) Centre. (c) SWJTU-LIB. (d) SWTJU-BLD. (e) CQ-BAISHA. (f) CQ-StudioCity.

the matching results from the perspective of recovering 3-D information of the scene. This metric supports the hypothesis that the number and quality of matches do not necessarily have a proportional relationship.

Table III reports a comparison between our method and other methods on the above three metrics. The results in Table III show that our proposed strategy, which utilizes a dense correspondence network to assist SIFT and SuperPoint in matching, can obtain correctly matched aerial-ground images on the six test datasets and exceeds other methods in number. Furthermore, due to the advantage of having more matches on aerial-ground image pair, our method is able to generate more 3-D points. We also find

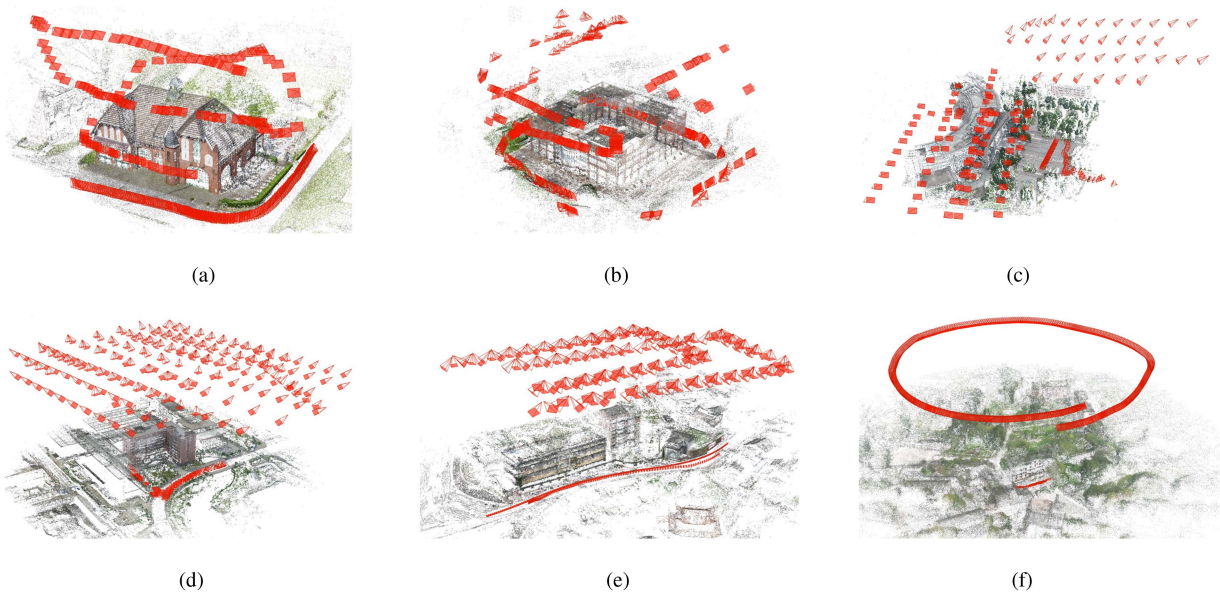


Fig. 8. Qualitative results of our method for generating aerial-ground sparse SfM point cloud on six test datasets. (a) Zeche. (b) Centre. (c) SWJTU-LIB. (d) SWJTU-BLD. (e) CQ-BAISHA. (f) CQ-StudioCity.

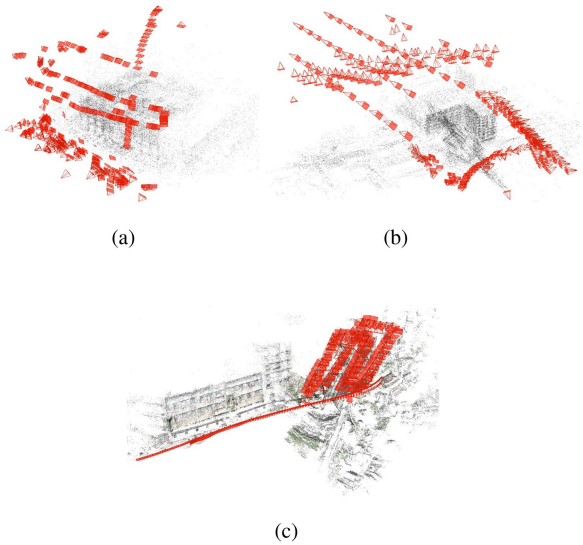


Fig. 9. Examples of failure to reconstruct aerial-ground sparse SfM point cloud. (a) SuperPoint + SuperGlue (Centre). (b) SuperPoint + SuperGlue (SWJTU-BLD). (c) COLMAP (CQ-BAISHA).

that, for Centre, CQ-BAISHA, and CQ-StudioCity, the results based on SuperPoint are slightly lower than those based on SIFT for all three metrics. This could be attributed to the fact that the structures in these three scenes are more suitable for SIFT to extract salient features, such as corners. COLMAP, on the other hand, is able to successfully integrate aerial and ground images in Zeche, Centre, and SWJTU-LIB, and outperformed SuperPoint + SuperGlue in terms of N_p and N_o, but is much lower in N_m. This phenomenon is because the location accuracy of SuperPoint is much lower than that of SIFT, which makes it difficult to effectively recover the 3-D information of the scene even if there are many true matches between the images. This further emphasizes the necessity of refining feature location. Similarly to the evaluation on two views, AdaLAM and SIFT+SuperGlue are unable to integrate aerial and ground images correctly, which is expected.

D. Evaluation on Integrated Sparse Point Clouds

In order to further evaluate the impact of feature matching results on subsequent 3-D scene recovery and the effect of our proposed location refinement (LR) and BA introducing FE on improving the quality of 3-D scenes, we initially feed our method's matching results and SuperPoint + SuperGlue's matching results into SfM for reconstructing 3-D point clouds. We then compare them with the 3-D point clouds generated by software such as COLMAP and RealityCapture. Furthermore, to demonstrate the effectiveness of LR and FE, we generate four different configurations of our method by including or excluding corresponding modules. To evaluate the results, we use two metrics: 1) average RE and average track length (TL), which are presented in Table IV.

From Table IV, it can be seen that both COLMAP and RealityCapture fail to integrate aerial and ground images for SWJTU-BLD, CQ-BAISHA, and CQ-StudioCity scenes, and RealityCapture also fails on Centre. This indicates a high failure rate when using existing software to complete aerial image integration tasks. In comparison, our method, using either SIFT or SuperPoint, not only successfully achieves aerial-ground integration, but also have lower average REs and longer average TLss. It is worth noting that SuperPoint + ours has a significantly higher average RE than SIFT+ours, because feature learning-based methods often have lower location accuracy due to the large receptive field and downsampling operations of CNNs, which can affect the accuracy of 3-D scene reconstruction. Additionally, as shown in the table, adding the keypoint LR and the BA with featuremetric significantly reduce the average RE and increase the average TL. For example, on Zeche, SIFT+ours(wLF, wPE)

Fig. 10. Qualitative comparison results of texture mesh models on six test datasets. First and second rows show the texture mesh models reconstructed solely from aerial images and the ones fused from both aerial and ground images, respectively. (a) Zeche. (b) Centre. (c) SWJTU-LIB. (d) SWJTU-BLD. (e) CQ-BAISHA. (f) CQ-StudioCity.

reduce the average RE by 25% and increase the average TL by 18% compared to SIFT+ours($w/oLF, w/oPE$).

Fig. 8 shows the sparse SfM point cloud obtained by our method using SIFT to integrate aerial and ground images, while Fig. 9 displays some examples of failures in other methods. In Fig. 9(a), all ground cameras are restored to the same facade view. This is because SuperPoint + SuperGlue lacks robustness to repeated structures. In Fig. 9(b), the addition of ground images even destroys the consistency of the original aerial scene, which also shows that the mismatches between aerial-ground images generated by SuperPoint + SuperGlue directly interferes with the operation of the SfM algorithm. In Fig. 9(c), the positional relationship between the aerial-ground cameras is misestimated, resulting in the generated scene points not being accurately registered. In contrast, our method can successfully integrate aerial and ground images by learning more consistent features.

E. Evaluation on Texture Models

Fig. 10 compares the texture mesh models obtained using only aerial images (top row) and the fusion of aerial and ground images (bottom row). In order to highlight the comparison results, we only select part of the facade of the model for display. From Fig. 10, it can be seen that integrating ground images into the aerial model can make the reconstructed model more complete and the texture clearer. This further illuminates that our proposed method can effectively integrate aerial and ground images to generate more accurate and complete 3-D information. Additionally, it is worth noting that in Fig. 10(e), there is ambiguity between the car objects in the model built from aerial images and those in the integrated model. This also indicates that dynamic objects or image differences need to be further considered in the fusion modeling process to obtain more reasonable models.

V. Conclusion

In our article, we propose to improve the SfM algorithm by learning dense consistent features for the integration of aerial-ground images. This approach primarily utilizes the learned features to improve feature matching and BA while introducing a method for adjusting keypoint locations to further refine the accuracy of 3-D scene reconstruction. Extensive experiments have demonstrated that our method not only significantly improves modeling accuracy compared to existing algorithms and software but also achieves effective aerial-ground image integration in challenging scenarios. In the next step, we plan to improve the multiview stereo algorithm to better adapt to aerial-ground images and generate high-quality depth maps and dense point clouds.

REFERENCES

[1] Q. Zhu, Z. Wang, H. Hu, L. Xie, X. Ge, and Y. Zhang, "Leveraging photogrammetric mesh models for aerial-ground feature point matching toward integrated 3D reconstruction," *ISPRS J. Photogrammetry Remote Sens.*, vol. 166, pp. 26–40, 2020.

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

[3] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, pp. 91–110, 2004.

[4] J.-M. Morel and G. Yu, "Asift: A new framework for fully affine invariant image comparison," *SIAM J. Imag. Sci.*, vol. 2, no. 2, pp. 438–469, 2009.

[5] J. L. Schonberger and J.-M. Frahm, "Structure-from-motion revisited," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 4104–4113.

[6] A. E. Johnson and M. Hebert, "Using spin images for efficient object recognition in cluttered 3D scenes," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 21, no. 5, pp. 433–449, May 1999.

[7] R. B. Rusu, N. Blodow, and M. Beetz, "Fast point feature histograms (fpfh) for 3D registration," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2009, pp. 3212–3217.

[8] Y. Guo, F. Sohel, M. Bennamoun, M. Lu, and J. Wan, "Rotational projection statistics for 3D local surface description and object recognition," *Int. J. Comput. Vis.*, vol. 105, pp. 63–86, 2013.

[9] B. Wu, L. Xie, H. Hu, Q. Zhu, and E. Yau, "Integration of aerial oblique imagery and terrestrial imagery for optimized 3D modeling in urban areas," *ISPRS J. Photogrammetry Remote Sens.*, vol. 139, pp. 119–132, 2018.

[10] X. Zheng, H. Li, H. Xiong, and X. Xie, "Lattice-point mutually guided ground-to-aerial feature matching for urban scene images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, no. 3, pp. 4737–4752, Mar. 2021.

[11] Q. Shan, C. Wu, B. Curless, Y. Furukawa, C. Hernandez, and S. M. Seitz, "Accurate geo-registration by ground-to-aerial image matching," in *Proc. IEEE 2nd Int. Conf. 3D Vis.*, 2014, vol. 1, pp. 525–532.

[12] H. Xu, J. Zhang, J. Cai, H. Rezatofighi, and D. Tao, "Gmflow: Learning optical flow via global matching," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 8121–8130.

[13] Z. Teed and J. Deng, "Raft: Recurrent all-pairs field transforms for optical flow," in *Proc. Comput. Vis.–ECCV 16th Eur. Conf.*, Glasgow, U.K., 2020, pp. 402–419.

[14] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, "ORB: An efficient alternative to sift or surf," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2011, pp. 2564–2571.

[15] M. Bansal, H. S. Sawhney, H. Cheng, and K. Daniilidis, "Geo-localization of street views with aerial image databases," in *Proc. 19th ACM Int. Conf. Multimedia*, 2011, pp. 1125–1128.

[16] M. Bansal, K. Daniilidis, and H. Sawhney, "Ultrawide baseline facade matching for geo-localization," in *Large-Scale Visual Geo-Localization*. Cham, Switzerland: Springer, 2016, pp. 77–98.

[17] M. Wolff, R. T. Collins, and Y. Liu, "Regularity-driven facade matching between aerial and street views," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 1591–1600.

[18] W.-Y. D. Lin, M.-M. Cheng, J. Lu, H. Yang, M. N. Do, and P. Torr, "Bilateral functions for global motion modeling," in *Proc. Comput. Vis.–ECCV 13th Eur. Conf.*, Zurich, Switzerland, 2014, pp. 341–356.

[19] W.-Y. Lin, S. Liu, N. Jiang, M. N. Do, P. Tan, and J. Lu, "Repmatch: Robust feature matching and pose for reconstructing modern cities," in *Proc. Comput. Vis.–ECCV 14th Eur. Conf.*, Amsterdam, The Netherlands, 2016, pp. 562–579.

[20] M. A. Fischler and R. C. Bolles, "Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography," *Commun. ACM*, vol. 24, no. 6, pp. 381–395, 1981.

[21] H. Li, X. Zheng, M. Dong, G.-S. Xia, and H. Xiong, "Locally nonlinear affine verification for multisensor image matching," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, no. 12, pp. 1–16, Dec. 2021.

[22] L. Zhou, S. Zhu, T. Shen, J. Wang, T. Fang, and L. Quan, "Progressive large scale-invariant image matching in scale space," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 2362–2371.

[23] H. Hu, Q. Zhu, Z. Du, Y. Zhang, and Y. Ding, "Reliable spatial relationship constrained feature point matching of oblique aerial images," *Photogrammetric Eng. Remote Sens.*, vol. 81, no. 1, pp. 49–58, 2015.

[24] P. Jende, F. Nex, M. Gerke, and G. Vosselman, "A fully automatic approach to register mobile mapping and airborne imagery to support the correction of platform trajectories in gnss-denied urban areas," *ISPRS J. Photogrammetry Remote Sens.*, vol. 141, pp. 86–99, 2018.

[25] Z. Zhang, A. Ganesh, X. Liang, and Y. Ma, "Tilt: Transform invariant low-rank textures," *Int. J. Comput. Vis.*, vol. 99, pp. 1–24, 2012.

[26] X. Gao, S. Shen, Y. Zhou, H. Cui, L. Zhu, and Z. Hu, "Ancient chinese architecture 3D preservation by merging ground and aerial point clouds," *ISPRS J. Photogrammetry Remote Sens.*, vol. 143, pp. 72–84, 2018.

[27] J. Sun, Z. Shen, Y. Wang, H. Bao, and X. Zhou, "Loftr: Detector-free local feature matching with transformers," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 8922–8931.

[28] M. Dusmanu, J. L. Schönberger, and M. Pollefeys, "Multi-view optimization of local feature geometry," in *Proc. Comput. Vis.–ECCV 16th Eur. Conf.*, Glasgow, U.K., 2020, pp. 670–686.

[29] A. Vaswani et al., "Attention is all you need," *Adv. Neural Inf. Process. Syst.*, vol. 30, pp. 6000–6010, 2017.

[30] X. Shen, F. Darmon, A. A. Efros, and M. Aubry, "Ransac-flow: Generic two-stage image alignment," in *Proc. Comput. Vis.–ECCV 16th Eur. Conf.*, 2020, pp. 618–637.

[31] D. DeTone, T. Malisiewicz, and A. Rabinovich, "Superpoint: Self-supervised interest point detection and description," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, 2018, pp. 224–236.

[32] A. Ranganathan, "The levenberg-marquardt algorithm," *Tutoral LM Algorithm*, vol. 11, no. 1, pp. 101–110, 2004.

[33] Z. Li and N. Snavely, "Megadepth: Learning single-view depth prediction from internet photos," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 2041–2050.

[34] M. Dusmanu et al., "D2-Net: A trainable CNN for joint description and detection of local features," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 8092–8101.

[35] P. Truong, M. Danelljan, and R. Timofte, "GLU-Net: Global-local universal network for dense flow and correspondences," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 6258–6268.

[36] F. Nex, M. Gerke, F. Remondino, H. Przybilla, M. Bäumker, and A. Zurhorst, "Isprs benchmark for multi-platform photogrammetry," *ISPRS Ann. Photogrammetry, Remote Sens. Spatial Inf. Sci.*, vol. 2, no. 3, pp. 135–142, 2015.

[37] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.

[38] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Commun. ACM*, vol. 60, no. 6, pp. 84–90, 2017.

[39] D. Sun, X. Yang, M.-Y. Liu, and J. Kautz, "PWC-Net: CNNs for optical flow using pyramid, warping, and cost volume," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 8934–8943.

[40] L. Cavalli, V. Larsson, M. R. Oswald, T. Sattler, and M. Pollefeys, "Adalam: Revisiting handcrafted outlier detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 1–13.

[41] P.-E. Sarlin, D. DeTone, T. Malisiewicz, and A. Rabinovich, "Superglue: Learning feature matching with graph neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 4938–4947.

Hongjie Li received the M.S. degree in geographic information system, in 2020, from the State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing (LIESMARS), Wuhan University, Wuhan, China, where he is currently working toward the Ph.D. degree in geographic information system.

His research interests include image matching, point cloud registration, structure from motion, and 3-D surface modeling.

Aonan Liu received the B.S. degree in surveying and mapping engineering from the School of Geodesy and Geomatics, Wuhan University, Wuhan, China, in 2021. He is currently working toward the M.S. degree in resources and environment with the State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing (LIESMARS), Wuhan University.

His research interests include image matching and 3-D reconstruction.

Xiao Xie received the Ph.D. degree in photogrammetry and remote sensing from Wuhan University, Wuhan, China, in 2016.

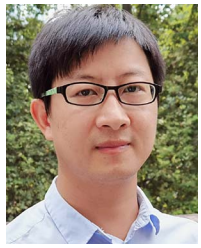
From 2014 to 2016, she was a Research Fellow with the Department of Cartography, Technical University of Munich, Munich, Germany. She is currently a Senior Engineer with the Key Lab of Environmental Computing and Sustainability, Liaoning province, as well as an Assistant Professor with Urban and Environmental Computation, the Institute of Applied Ecology, Chinese Academy of Sciences, Beijing, China. She is also a Postdoctoral Researcher with the School of Geodesy and Geomatics, Wuhan University. Her research interests include 3-D GIS and smart cities.

Han Guo received the B.S. degree in civil and infrastructure engineering from RMIT University, Melbourne, Australia, and the M.S. degree in master of engineering from Melbourne University, Melbourne, Australia, in 2012 and 2014, respectively. He is currently working toward the Ph.D. degree in cartography and geographical information engineering with the School of Resource and Environmental Sciences, Wuhan University, Wuhan, China.

His research interests include 3-D cadaster, and smart city construction.

Hanjiang Xiong received the B.S. degree in photogrammetry and remote sensing from the School of Remote Sensing and Engineering, Wuhan University of Surveying and Mapping, Wuhan, China, and the Ph.D degree in photogrammetry and remote sensing from the State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing (LIESMARS), Wuhan University, Wuhan, China, in 1995 and 2002, respectively.

He has been a Visiting Scholar with Queensland University of Technology for three months in 2011. He is currently a Full Professor in 3-D GIS with Wuhan University. His current research interests include geospatial data management, 3-D visualization, augmented reality, and indoor and outdoor GIS.



Xianwei Zheng (Member, IEEE) received the M.S. and Ph.D degrees in photogrammetry and remote sensing from the State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing (LIESMARS), Wuhan University, Wuhan, China, in 2010 and 2015, respectively.

He is currently an Associate Professor in computer vision and 3-D geographical information science (GIS) with Wuhan University. His research interests include indoor and outdoor scene parsing, 3-D computer vision and reconstruction, and geovisualization.