Aerial–ground collaborative 3D reconstruction for fast pile volume estimation with unexplored surroundings

Sheng Liu, Jingxiang Yu, Zhenghao Ke, Fengji Dai and Yibin Chen

Abstract
Fast and accurate pile volume estimation is a very important basic problem in mining, waste sorting, and waste disposing industry. Nevertheless, for rapid changing or badly conditioned piles like stockpiles or landfills, conventional approaches involving massive measurements may not be applicable. To solve these problems, in this work, by utilizing unmanned aerial vehicles and unmanned ground vehicles equipped with a camera, we propose a collaborative framework to estimate volumes of free-formed piles accurately in short time. Compensating aerial- and ground views enable the reconstruction of piles with steep sides that is hard to be observed by single unmanned aerial vehicle. With the help of red-green-blue image sequences captured by unmanned aerial vehicles, we are able to distinguish piles from the ground in reconstructed point clouds and automatically eliminate concave on the ground while estimating pile volume. In the experiments, we compared our method to state-of-the-art dense reconstruction photogrammetry approaches. The results show that our approach for pile volume estimation has proved its feasibility for industrial use and its availability to free-formed piles on a different scale, providing high-accuracy estimation results in short time.

Keywords
Volume estimation, free-formed pile, collaborative reconstruction, video analysis

Introduction
Piles, usually free-formed, are one of the important material storing and managing forms in the industry. The pile volume is of great significance to scientific management, economic benefit assessment, and storage capacity assessment. With the accelerating development of large-scale storage bases, modern management techniques for stock storage are required. However, current approaches to calculating pile volume are likely to involve massive measurements or high computational complexity, leading to high labor cost or time usage. For situations where stock changes rapidly or piles in bad conditions, like port or landfill, they may not be applicable. To solve the above problem, various studies have been conducted. Current approaches to calculating pile volume are divided mainly into three categories:

Corresponding author:
Sheng Liu, College of Computer Science, Zhejiang University of Technology, Hangzhou 310023, Zhejiang, People’s Republic of China.
Email: edliu@zjut.edu.cn
1. Manually reshape the pile into a regular shape, to calculate the pile volume.
2. Use a total station theodolite (TST) to measure the entire pile and calculate an approximate volume.
3. Densely reconstruct the entire pile with oblique photography, photogrammetry, or light detection and ranging (LiDAR) to get the detailed mesh, and then compute the volume of mesh.1,2

However, these approaches require either intensive human resource or high computational power, causing high time usage in volume calculation. For better illustration, a 11,500m³ pile requires more than 3 h to measure with a TST2; for dense reconstruction approaches, a typical landfill base of approximately 53,628m² can take more than 1 day for a modern computer to reconstruct. Thus, for rapidly changing or badly conditioned piles, those approaches may not be applicable.

At present, reconstruction methods are intensively used in pile volume estimation, reconstruction results from oblique photography can provide high-resolution measurement results with accuracy at centimeter level and is widely used and accepted in geosciences and geomatics industry as a reference of survey and mapping,3–5 but the high-resolution reconstruction also leads to more time-consuming and inefficient methods. In some reconstruction that does not require high resolution, visual simultaneous localization and mapping (SLAM) also have a wide range of applications. SLAM, with allowing reconstructing maps in real time at the expense of accuracy, are some based on conventional visual,6–9 some based on deep learning,10,11 and some based on multisensor fusion.12–14 In a wide range of scenarios, those approach engaging SLAM with a single agent has many defects, including limited sensing range and viewing angle, low data storage capacity, and high onboard computational complexity. By implementing multi-agent, SLAM can simply overcome these difficulties; however, there are challenges in mutual pose correction and information perception among multiple robots, but simply implementing multi-agent SLAM will introduce other problems. For applications like pile volume estimation, reconstruction results from single observe view may be incomplete and does not meet the requirements to accurately estimate the volume.

In this work, we proposed a multi-agent aerial–ground collaborative framework to quickly recover the sparse three-dimensional (3D) information of the pile in unexplored surroundings with marker optimization and lastly calculate the pile volume by the heightmap. Compared with the dense mapping of a single robot, collaborative framework measures the pile volume in a short time with high accuracy. Our main contributions are summarized as follows:

(1) We propose a novel framework of aerial–ground collaborative sparse reconstruction. This framework can easily combine the information for multiple views and generate a global sparse reconstruction of the environment. The keyframe poses of the unmanned aerial vehicle (UAV) and the unmanned ground vehicle (UGV) are optimized by tag-based pose optimization.

(2) We build a new mechanism of tag-based estimation for map scale estimation and interagent local map alignment. Using the tag pose and dimension as prior information, the distance between tag center and the observing camera center can be estimated, and then the duplicated points in the merged map are removed and local maps are aligned and merged into the global map.

(3) The proposed method is applied in pile volume estimation for the first time. In comparison with the single-agent setup, the results show that the overall accuracy of our method is improved significantly.

Related works
This section reviews advances in collaborative SLAM and UAV photogrammetry in a wide range of areas. Early multiple agents are based on the relative positioning and the expensive base-mobile stations positioning system, then gradually to UGVs and UAVs collaborative SLAM. However, lacking the map scale, those methods cannot be used to estimate the real measurement. Our work is based on tag-based estimation for map scale estimation and thus can be used in volume estimation.

Collaborative SLAM
Many research studies have been conducted on collaborative SLAM. For example, Jo et al.15 proposed a method for obtaining the interagent distance based on real-time kinematic (RTK) data of different agents to realizing the relative positioning among agents and reducing the positioning error in the relative positioning process. However, in the case of a weak global position system (GPS) signal, the positioning accuracy of GPS will be worse than that of the conventional sensor like inertial measurement unit (IMU). Some researchers have attempted to use the RTK positioning system by selecting one agent as the RTK base station and others to be RTK mobile stations. While RTK devices are relatively expensive, this attempt is difficult to be widely applied.

Surmann et al.16 proposed a framework that integrates a UAV and a UGV to do SLAM in a relatively complex scenario. In their work, the UGV and UAV equipped with sensors such as LiDAR and vision camera, and the merged map is generated when a similar transformation is done between dense mapping resulting from UAV and a point cloud from UGV. And then Zhang et al.17 proposed an
approach based on environmental perception, which only relied on the cameras of UAV and UGV, and then combined with semantic segmentation to jointly construct the global map.

In some other works, Schmuck and Chli and Van Opdenbosch and Steinbach proposed that multiple agents run SLAM independently, transferring local maps together with keyframe data to a central server, when visual overlaps are detected between the keyframes from different agents, the server merges the local map into a global map. Their fundamental approach does not correct the map scale and thus cannot be used in volume estimation. Karrer et al. were also proposed a collaborative visual-inertial SLAM, which allows agents to share all information with the central server to improve the performance.

**UAV photogrammetry**

Since the high accuracy of UAV photogrammetry or oblique photography, many types of research have been conducted. Mokroš et al. used poles for scale reference and done monocular 3D reconstruction with a single UAV, obtaining a volume error about 10% between values calculated from global navigation satellite system (GNSS) data under an average reconstruction density of 75,300 points per m³. In the work of Nesbit and Hugenholtz, the steep part of structure from motion reconstruction results from UAVs are enhanced by incorporating oblique images. However, their approaches are computationally intensive and can only be performed off-line.

**Aerial–ground collaborative sparse reconstruction**

Framework of aerial–ground collaborative sparse reconstruction We build our UAV and UGV collaborative framework directly on oriented FAST and rotated BRIEF (ORB-SLAM), since its high integrity for the entire SLAM system including visual odometry, back-end optimization, loop detection, and map construction. As shown in Figure 1, (a) the aerial end is comprised of a set of UAVs equipped with a monocular camera and GPS. (b) The ground end consisting of a group of UGVs with a monocular camera and AprilTag printed on its surface. The tracking module in each agent tracks map points and the pose of corresponding unmanned vehicles, sending keyframes to mapping module to create a local map. (c) Central server listens for inbound connections from UAVs and UGVs, receiving transferred data from both ends, merging maps from each agent to create the global map. A tag detection procedure is running on central server polling for keyframes and detects for any presenting AprilTag.

The aerial–ground collaborative sparse reconstruction algorithm has three separate phases, which are as follows:

1. Each agent uses an onboard computing platform to perform the visual SLAM algorithm independently and continuously estimate six-degree-of-freedom (6-DoF) rigid transformation information together with the scale factor.
2. When the central server detects AprilTag in any keyframe, the coordinate of corresponding
observing UAV and observed UGV can be unified. Meanwhile, because of the prior knowledge of the tag’s dimension, the scale factor can be inferred, and the local map is then merged into one consistent global map. Also, the keyframe poses of the UAV and the UGV are optimized by tag-based pose optimization.

3. The optimized pose is then broadcasted to all agents, and a global bundle adjustment is performed to minimize the offset of the global map.

**Tag-based map scale estimation**

After SLAM procedures on agents were initialized and local maps were created on agents with their first keyframe pose as their coordinate origin, whenever a tag is detected, the tag pose can be estimated by solving a perspective-n-point problem. If we know the tag dimension as prior, we can estimate the distance between the tag center and observing camera center, this distance is used as the scale reference in the following tag-based optimizations. The 6-DoF rigid transformation matrix can be written as

\[
T = \begin{bmatrix} R_{3 \times 3} & t \\ 0^T & 1 \end{bmatrix} = P^{-1}
\]  

And transformation matrix \( T_A^G \) that maps homogeneous points in aerial coordinate \( A \) to ground coordinate \( G \) by \( P_G = T_A^G P_A^B \).

\[
P_A^B \text{ is the 6-DoF rigid pose matrix of object } A \text{ in coordinate } B. \text{ For a given keyframe pair } K = \{ A_i, G_j \}, \text{ where } A_i \text{ denotes the } i\text{-th keyframe with tag detected from UAV, } G_j \text{ denotes the } j\text{-th keyframe from UGV, and tag denotes the detected AprilTag. The inter-frame transformation matrix can be written as }
\]

\[
T_{A_i}^{G_{j+1}} = T_{A_i}^{G_0} T_{G_{j+1}}^{G_0}
\]  

We can then compute the scale factor \( s_{AG} \) for the corresponding UAV and UGV when a set of consecutive keyframes containing valid tag captured by UAV. \( t_{G_{j+1}}^{A_i} \) and \( t_{G_j}^{A_i} \) are translation between corresponding keyframes estimated by tracking procedure. Figure 2 shows the keyframe pairs used in this tag-based optimization procedure

\[
t_{G_{j+1}}^{A_i} = s_{AG} (R_{A_i}^{G_{j+1}} R_{A_i}^{G_j} t_{G_j}^{A_i})
\]  

**Tag-based interagent local map alignment**

Besides the scale factor can be recovered from tag prior, the tag is also used to align corresponding maps when a tag pose \( T_A^G \) is estimated, the inter-map transformation matrix can be written as

\[
T^d_{A_0} = T_{A_0}^{G_0} T_{G_j}^{G_0}
\]  

where \( A_0 \) and \( G_0 \) are the coordinate system of the first keyframe of the corresponding UAV and UGV, respectively. The mapping and localization results from a single UGV and a single UAV are shown in Figure 3(a) and (b), respectively. Figure 3(c) shows the result from collaborative reconstruction with a UGV and a UAV. With tag-based optimizations, the duplicated points in the merged map are removed and local maps are aligned and merged into the global map. Three different views are provided for each map.

**Pile volume estimation overview**

In this section, we describe our designed framework for pile volume estimation. Figure 4 shows the block diagram of our designed framework for pile volume estimation. Firstly, a multi-agent reconstruction module continuously analyzes the input video stream, detects, and collects keyframes to generate a sparse point cloud for the captured area. Then the point cloud and keyframe are passed to the ground segmentation module. For each keyframe, we perform the ground segmentation module to calculate the probability that a map point belongs to the ground plane class. After the ground point cloud is collected, random sample consensus (RANSAC)\(^{22} \) is used to fit a ground plane. Once the best fit ground plane is calculated, points will be reprojected to a newly generated coordinate system based on the ground plane and voted into a fine grid to get a heightmap.

In the end, the pile volume can be calculated by summing all elements in the heightmap with positive values and some proper scaling.

**Ground plane estimation**

The correct selection of the ground plane is considered important in volume estimation since the ground plane is used directly in heightmap generation. We use a simple
encoding–decoding network to perform segmentation on every possible keyframe, we resize every frame to the size of $256 \times 256$. The structure of our network is shown in Figure 5, it is a small network with seven layers including five convolutional layers (Conv2D) and two transposed convolutional layers (Deconv2D). We gathered images of piles from the Internet and manually labeled them with binary masks. Since the simplicity of this task, the data set we built is relatively small, with about only 100 images. And in which we randomly selected 20 images for

Figure 3. Constructed maps from the single agent and collaborative reconstruction with three views provided for each map. (a) Local map from single UGV, (b) local map from single UAV, and (c) merged global map with tag-based optimization of collaborative reconstruction with UAV and UGV. UAV: unmanned aerial vehicle; UGV: unmanned ground vehicle.

Figure 4. Overview of our designed framework for pile volume estimation. The entire system is built by five main modules.
After heightmap \( H \) is generated, the pile volume \( V \) can be easily calculated by summing all elements in the heightmap with positive values multiplied by the grid resolution \( r \).

\[
V = \sum_{h > 0, h \in H} h \cdot r^2
\]  

**Experiments and discussion**

In this section, we describe our experiments and evaluations of our approach. We conducted experiments on a wide range of piles, including large-scale landfills and other stockpiles, to evaluate the speed, accuracy, repeatability, and robustness of our proposed approach.

To evaluate the localization accuracy of our proposed approach among multiple agents, we evaluate our aerial–ground collaborative reconstruction approach in a school campus.

We built a data set with several piles with our collaborative multi-agent setup at different flight heights and UGV on the ground. Based on the data set we built, we conducted the following experiments. Figure 8 is a demonstration of the UAV and UGV used in our experiments in collaborative reconstruction, the UGV is marked with a tag. The top view of the landfill and rock pile in our experiments can be found in Figure 9. For comparison, we captured extra image sequences for the photogrammetry method to produce dense reconstruction results.

**Accuracy evaluation on collaborative reconstruction**

In this section, we describe our evaluations on collaborative reconstruction and analyze the accuracy of the mapping and localization procedure. The UAV in this experiment is DJI A3 flight controller-driven drones, equipped with NVIDIA Jetson TX1 onboard computers. And the UGV is driven by Intel NUC. We used a laptop with Intel i7-8750H processor, NVIDIA GeForce 1070 graphics card, and 8 GB RAM as a central server. The overall system is driven by Intel NUC. We used a laptop with Intel i7-8750H processor, NVIDIA GeForce 1070 graphics card, and 8 GB RAM as a central server. The overall system and localization procedure. The UAV in this experiment is DJI A3 flight controller-driven drones, equipped with NVIDIA Jetson TX1 onboard computers. And the UGV is driven by Intel NUC. We used a laptop with Intel i7-8750H processor, NVIDIA GeForce 1070 graphics card, and 8 GB RAM as a central server.

validation, the images from our testing pile are not used in training and validating.

For each map point, we maintain a list of co-visible keyframes. We use a voting mechanism to determine a map point whether belongs to ground or pile, each keyframe votes for the map point, the category with highest votes with more than 60% votes will be selected as the determined category, if no category has more than 60% votes, the map point stays undetermined.

Then the determined ground point cloud is passed into the RANSAC algorithm to fit the parameters of the ground plane \( P : z = Ax + By + C \), the algorithm iteratively finds a good fit with most inlier points. Figure 6(a) to (c) shows an input–output pair of the ground plane estimation procedure. And an example of segmentation results and the heightmap will be easily calculated by summing all elements in the heightmap with positive values multiplied by the grid resolution \( r \).

\[
V = \sum_{h > 0, h \in H} h \cdot r^2
\]  

**Experiments and discussion**

In this section, we describe our experiments and evaluations of our approach. We conducted experiments on a wide range of piles, including large-scale landfills and other stockpiles, to evaluate the speed, accuracy, repeatability, and robustness of our proposed approach.

To evaluate the localization accuracy of our proposed approach among multiple agents, we evaluate our aerial–ground collaborative reconstruction approach in a school campus.

We built a data set with several piles with our collaborative multi-agent setup at different flight heights and UGV on the ground. Based on the data set we built, we conducted the following experiments. Figure 8 is a demonstration of the UAV and UGV used in our experiments in collaborative reconstruction, the UGV is marked with a tag. The top view of the landfill and rock pile in our experiments can be found in Figure 9. For comparison, we captured extra image sequences for the photogrammetry method to produce dense reconstruction results.

**Accuracy evaluation on collaborative reconstruction**

In this section, we describe our evaluations on collaborative reconstruction and analyze the accuracy of the mapping and localization procedure. The UAV in this experiment is DJI A3 flight controller-driven drones, equipped with NVIDIA Jetson TX1 onboard computers. And the UGV is driven by Intel NUC. We used a laptop with Intel i7-8750H processor, NVIDIA GeForce 1070 graphics card, and 8 GB RAM as a central server. The overall system and localization procedure. The UAV in this experiment is DJI A3 flight controller-driven drones, equipped with NVIDIA Jetson TX1 onboard computers. And the UGV is driven by Intel NUC. We used a laptop with Intel i7-8750H processor, NVIDIA GeForce 1070 graphics card, and 8 GB RAM as a central server.
optimization algorithms. This experiment is conducted in a school campus.

Table 1 presents that the average errors of a single UAV and UGV SLAM are 0.134 m and 0.166 m, respectively, in a 50 x 50 m² area. The collaborative setup considerably reduces the average error to 0.05 m. Table 2 presents that the average pose errors of a single UAV and a single UGV are 0.206 and 0.209, respectively. With collaborative approaches, the average error is reduced to 0.195. Additional matching points are extracted in the keyframe sequence due to the implementation of the collaborative setup. In comparison with the single-agent setup, the

Figure 6. Ground segmentation and plane estimation. The ground points are shown in green, pile points are shown in red, whereas the undetermined points are marked in yellow. (a) Segmentation input, (b) segmentation output, (c) overlaid image, (d) side view of segmented point cloud, and (e) point cloud with estimated ground plane.
overall accuracy is improved by 61.37%, the preceding improvements validate the performance of the proposed method on multi-agent collaborative reconstruction.

**Volume estimation performance on small-scale rock pile**

In this experiment, we evaluate the estimation accuracy and speed of our approach on a rock pile as shown in Figure 9(a). We use the calculated volume by a dense method as a reference, the dense mesh is reconstructed from 186 images in 5 different viewing angles at 60 m above ground. Both dense reconstruction and our approach are performed on an Intel i7-8700 K PC with NVIDIA GTX 1070Ti graphic cards.

The estimated values and time usage are listed in Table 3. Sequences 1–3 are three different captured image sequences of the same pile at a flight height of 10, 20, and 30 m, respectively. As shown in the table, flight height has an implicit impact on the estimation accuracy and a direct impact on the estimation time, since higher flying height provides wider viewport coverage, but less texture detail. The optimal flight height is subjected to the pile scale and the pile texture.

As we can see, the estimation time of our approach is shorter than the acquisition time and is several times shorter than the time usage of dense method, which means with the

![Figure 7](image1.png)

**Figure 7.** An example of the heightmap generation. The brighter location is higher. (a) Generated heightmap and (b) heightmap with holes filled.

![Figure 8](image2.png)

**Figure 8.** The photo of (a) UAV and (b) UGV in collaborative setup used in our experiments. UAV: unmanned aerial vehicle; UGV: unmanned ground vehicle.
parallelization of data acquisition and estimation procedure, the estimated values can be read right after acquisition procedure is finished. Compared to the dense method, our approach is fast and accurate, with a minimum volume error of 0.73%. A comparison of the rendered model from generated heightmap and dense reconstruction can be seen in Figure 10.

**Volume estimation performance on large-scale landfill**

In this experiment, we evaluate the accuracy of the pile volume estimation method on a large-scale landfill. This landfill mainly serves for daily waste disposal, with some part of the pile covered with plastic sheets, and the top view of this landfill is shown in Figure 9(b).

The estimated values and time usage are listed in Table 4. Sequences 1–3 are three different captured image sequences of the same pile, in which sequence 1 is captured at a flight height of 60 m, and both sequences 2 and 3 are captured at a flight height of 30 m. Here we also use the calculated volume by the dense method as a reference, the dense mesh is reconstructed from 1677 images in 5 different viewing angles at 60 m above ground.

![Figure 9. Top view of the piles used in experiments. (a) The small-scale rock pile and (b) the large-scale landfill.](image)

**Table 1.** Comparison of the translational error between single- and multi-agent reconstruction.

| Experiment  | Matching point pairs | Mean-square error (m) | Average deviation (m) | Standard deviation (m) | Minimum error (m) | Maximum error (m) |
|-------------|----------------------|-----------------------|-----------------------|------------------------|-------------------|-------------------|
| UAV         | 3256                 | 0.153958              | 0.134467              | 0.074977               | 0.001573          | 0.337134          |
| UGV         | 2948                 | 0.193208              | 0.164689              | 0.098035               | 0.001646          | 0.249191          |
| UAV + UGV   | 4432                 | 0.061033              | 0.051926              | 0.032074               | 0.000356          | 0.192824          |

UAV: unmanned aerial vehicle; UGV: unmanned ground vehicle.

**Table 2.** Comparison of the rotational error between single- and multi-agent reconstruction.

| Experiment  | Matching point pairs | Mean-square error (°) | Average deviation (°) | Standard deviation (°) | Minimum error (°) | Maximum error (°) |
|-------------|----------------------|-----------------------|-----------------------|------------------------|-------------------|-------------------|
| UAV         | 3256                 | 0.250548              | 0.205800              | 0.1402901              | 0.005253          | 1.314837          |
| UGV         | 2948                 | 0.259415              | 0.208513              | 0.154332               | 0.005271          | 1.601783          |
| UAV + UGV   | 4432                 | 0.235729              | 0.194839              | 0.132688               | 0.009711          | 1.096236          |

UAV: unmanned aerial vehicle; UGV: unmanned ground vehicle.

**Table 3.** Evaluation results on small-scale rock pile.

|                | Area (m²) | Volume (m³) | Acquisition time (min) | Estimation time (min) |
|----------------|-----------|-------------|------------------------|-----------------------|
| Dense          | 393.92    | 470.24      | 17.35                  | 46.70                 |
| Sequence 1 (@10 m) | 412.75 (+4.78%) | 457.75 (+2.65%) | 5.98                   | 2.63                  |
| Sequence 2 (@20 m) | 418.06 (+6.27%) | 466.80 (+0.73%) | 4.90                   | 3.39                  |
| Sequence 3 (@30 m) | 420.36 (+6.71%) | 482.51 (+2.61%) | 5.43                   | 2.57                  |
| Average        | 417.06 (+5.87%) | 469.02 (+0.26%) | —                      | —                     |
The conclusion is similar to volume estimation performance on a small-scale rock pile, with collaborative sparse reconstruction method, the estimation time and reconstruction time are shortened. Compared to the dense method, our approach is fast and accurate, with a minimum volume error of 0.12%.

Conclusion

In this study, we propose an innovative and efficient aerial–ground collaborative reconstruction framework to solve the problem of pile volume estimation. In this framework, we present tag-based optimizations to recover map scales and inter-map transformation matrix, so that local maps from different agents can be aligned and merged into a global map. We have conducted a series of experiments on both large-scale and small-scale piles. This method is proved to work reliably on free-formed piles and estimate the volume of the pile with a reasonable accuracy which is capable of industrial usages and is meaningful to stock management and solid waste recycling industry.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by the National Key R&D Program of China (no. 2018YFB1305200), Science Technology Department of Zhejiang Province (no. LGG19F020010), and National Natural Science Foundation of China (no. 61802348).

ORCID iDs

Jingxiang Yu https://orcid.org/0000-0001-9417-5669
Fengji Dai https://orcid.org/0000-0002-1030-2382

References

1. Mokroš M, Tabáček M, Lieskovský M, et al. Unmanned aerial vehicle use for wood chips pile volume estimation. *Int Arch Photogramm Remote Sens Spat Inf Sci* 2016; 41: 953.
2. Arango C and Morales C. Comparison between multicopter UAV and total station for estimating stockpile volumes. *Int Arch Photogramm Remote Sens Spat Inf Sci* 2015; 40(1): 131.
3. Nesbit PR and Hugenholtz CH. Enhancing UAV–SFM 3D model accuracy in high-relief landscapes by incorporating oblique images. *Remote Sens* 2019; 11(3): 239.
4. Westoby MJ, Brasington J, Glasser NF, et al. ‘Structure-from-motion’ photogrammetry: a low-cost, effective tool for geoscience applications. *Geomorphology* 2012; 179: 300–314.
5. Smith M, Carrivick J, and Quincey D. Structure from motion photogrammetry in physical geography. *Prog Phys Geogr* 2016; 40(2): 247–275.
6. Engel J, Schöps T, and Cremers D. LSD-SLAM: large-scale direct monocular SLAM. In: *European conference on computer vision*. Cham: Springer, 2014, pp. 834–849.
7. Mur-Artal R, Montiel JMM, and Tardos JD. ORB-SLAM: a versatile and accurate monocular SLAM system. *IEEE Trans Robot* 2015; 31(5): 1147–1163.
8. Mur-Artal R and Tardós JD ORB-SLAM2: an open-source slam system for monocular, stereo, and RGB-D cameras. *IEEE Trans Robot* 2017; 33(5): 1255–1262.
9. Engel J, Koltun V, and Cremers D. Direct sparse odometry. *IEEE Trans Pattern Anal* 2017; 40(3): 611–625.
10. Bowman SL, Atanasov N, Daniilidis K, et al. Probabilistic data association for semantic slam. In: 2017 IEEE international conference on robotics and automation (ICRA), Singapore, 29 May–3 June 2017, pp. 1722–1729.

11. Zhang J, Gui M, Wang Q, et al. Hierarchical topic model based object association for semantic SLAM. *IEEE Trans Vis Comput Graph* 2019; 25(11): 3052–3062.

12. Qin T, Li P, and Shen S.VINS-Mono: a robust and versatile monocular visual-inertial state estimator. *IEEE Trans Robot* 2018; 34(4): 1004–1020.

13. Liu R, Zhang J, Chen S, et al. Towards SLAM-based outdoor localization using poor GPS and 2.5D building models. In: 2019 IEEE international symposium on mixed and augmented reality, Beijing, China, 14–18 October 2019.

14. Wang Z, Zhang J, Chen S, et al. Robust high accuracy visual-inertial-laser SLAM system. In: 2019 IEEE/RSJ international conference on intelligent robots and systems (IROS), Macau, China, 3–8 November 2019.

15. Jo K, Lee J, and Kim J. Cooperative multi-robot localization using differential position data. In: 2007 IEEE/ASME international conference on advanced intelligent mechatronics, Zurich, Switzerland, 4–7 September 2007, pp. 1–6.

16. Surmann H, Berninger N, and Worst R. 3D mapping for multi-hybrid robot cooperation. In: 2017 IEEE/RSJ international conference on intelligent robots and systems (IROS), Vancouver, Canada, 24–27 September 2017, pp. 626–633. IEEE.

17. Zhang J, Liu R, Yin K, et al. Intelligent collaborative localization among air-ground robots for industrial environment perception. *IEEE Trans Ind Electron* 2018; 66(12): 9673–9681.

18. Schmuck P and Chli M. Multi-UAV collaborative monocular SLAM. In: 2017 IEEE international conference on robotics and automation (ICRA), Singapore, 29 May–3 June 2017, pp. 3863–3870.

19. Van Opdenbosch D and Steinbach E. Collaborative visual SLAM using compressed feature exchange. *IEEE Robot Autom Lett* 2018; 4(1): 57–64.

20. Karrer M, Schmuck P, and Chli M. CVI-SLAM—collaborative visual-inertial SLAM. *IEEE Robot Autom Lett* 2018; 3(4): 2762–2769.

21. Wang J and Olson E. AprilTag 2: efficient and robust fiducial detection. In: 2016 IEEE/RSJ international conference on intelligent robots and systems (IROS), Daejeon, South Korea, 9–14 October 2016, pp. 4193–4198.

22. Fischler MA and Bolles RC. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Commun ACM* 1981; 24(6): 381–395.