Computers and Electronics in Agriculture 227 (2024) 109661

Contents lists available at ScienceDirect

Computers and Electronics in Agriculture

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

Original papers

Benchmarking of monocular camera UAV-based localization and mapping methods in vineyards

Kaiwen Wang [a,b,*], Lammert Kooistra [c], Yaowu Wang [c], Sergio Vélez [d], Wensheng Wang [b], João Valente [e]

[a] Information Technology Group, Wageningen University & Research, 6708 PB, Wageningen, the Netherlands
[b] Agricultural Information Institute, Chinese Academy of Agriculture Sciences, Beijing 10086, China
[c] Laboratory of Geo-Information Science and Remote Sensing, Wageningen University & Research, 6708 PB, Wageningen, the Netherlands
[d] Group Agrivoltaics, Fraunhofer Institute for Solar Energy Systems ISE, 79110, Freiburg, Germany
[e] Centre for Automation and Robotics (CAR), Spanish National Research Council (CSIC), 28500, Madrid, Spain

ARTICLE INFO

Keywords:
SLAM
Structure from Motion
Precision Agriculture
Up-close Sensing

ABSTRACT

UAVs equipped with various sensors offer a promising approach for enhancing orchard management efficiency. Up-close sensing enables precise crop localization and mapping, providing valuable a priori information for informed decision-making. Current research on localization and mapping methods can be broadly classified into SfM, traditional feature-based SLAM, and deep learning-integrated SLAM. While previous studies have evaluated these methods on public datasets, real-world agricultural environments, particularly vineyards, present unique challenges due to their complexity, dynamism, and unstructured nature.

To bridge this gap, we conducted a comprehensive study in vineyards, collecting data under diverse conditions (flight modes, illumination conditions, and shooting angles) using a UAV equipped with high-resolution camera. To assess the performance of different methods, we proposed five evaluation metrics: efficiency, point cloud completeness, localization accuracy, parameter sensitivity, and plant-level spatial accuracy. We compared two SLAM approaches against SfM as a benchmark.

Our findings reveal that deep learning-based SLAM outperforms SfM and feature-based SLAM in terms of position accuracy and point cloud resolution. Deep learning-based SLAM reduced average position error by 87% and increased point cloud resolution by 571%. However, feature-based SLAM demonstrated superior efficiency, making it a more suitable choice for real-time applications. These results offer valuable insights for selecting appropriate methods, considering illumination conditions, and optimizing parameters to balance accuracy and computational efficiency in orchard management activities.

1. Introduction

Viticulture, the cultivation of grapes for winemaking, is a complex and dynamic agricultural practice that necessitates precise monitoring and management (Tardaguila et al., 2021). Precision viticulture integrates an array of remote sensing technologies, advanced sensor applications, variable rate technologies, actuators, and artificial intelligence, fostering an interconnected ecosystem that optimizes grape cultivation practices through monitoring, informed decision-making and targeted interventions (Santesteban, 2019). The structure and geometric characteristics of vineyards can vary significantly depending on the vine training system and pruning methods employed (Keller, 2020). Furthermore, the small size of grapes and the potential for a dense canopy close to the ground can lead to occlusion issues, complicating the detection and management of stems, trunks, and grapes (Moreno et al., 2020). Addressing these issues requires up-close sensing rather than nadir remote sensing, which involves acquiring information at a relatively close distance (Ariza-Sentís et al., 2023).

In recent years, Unmanned Aerial Vehicles (UAVs) equipped with different sensors (e.g., LiDAR, RGB camera, RGBD camera) have emerged as effective up-close sensing tools for precision agriculture, providing a non-intrusive and cost-effective means of gathering high-resolution spatial data (Ariza-Sentís et al., 2024). Although the indoor reconstruction performance of RGBD cameras is very good, their

* Corresponding author.
E-mail address: kaiwen.wang@wur.nl (K. Wang).

https://doi.org/10.1016/j.compag.2024.109661
Received 18 June 2024; Received in revised form 9 November 2024; Accepted 11 November 2024
Available online 15 November 2024

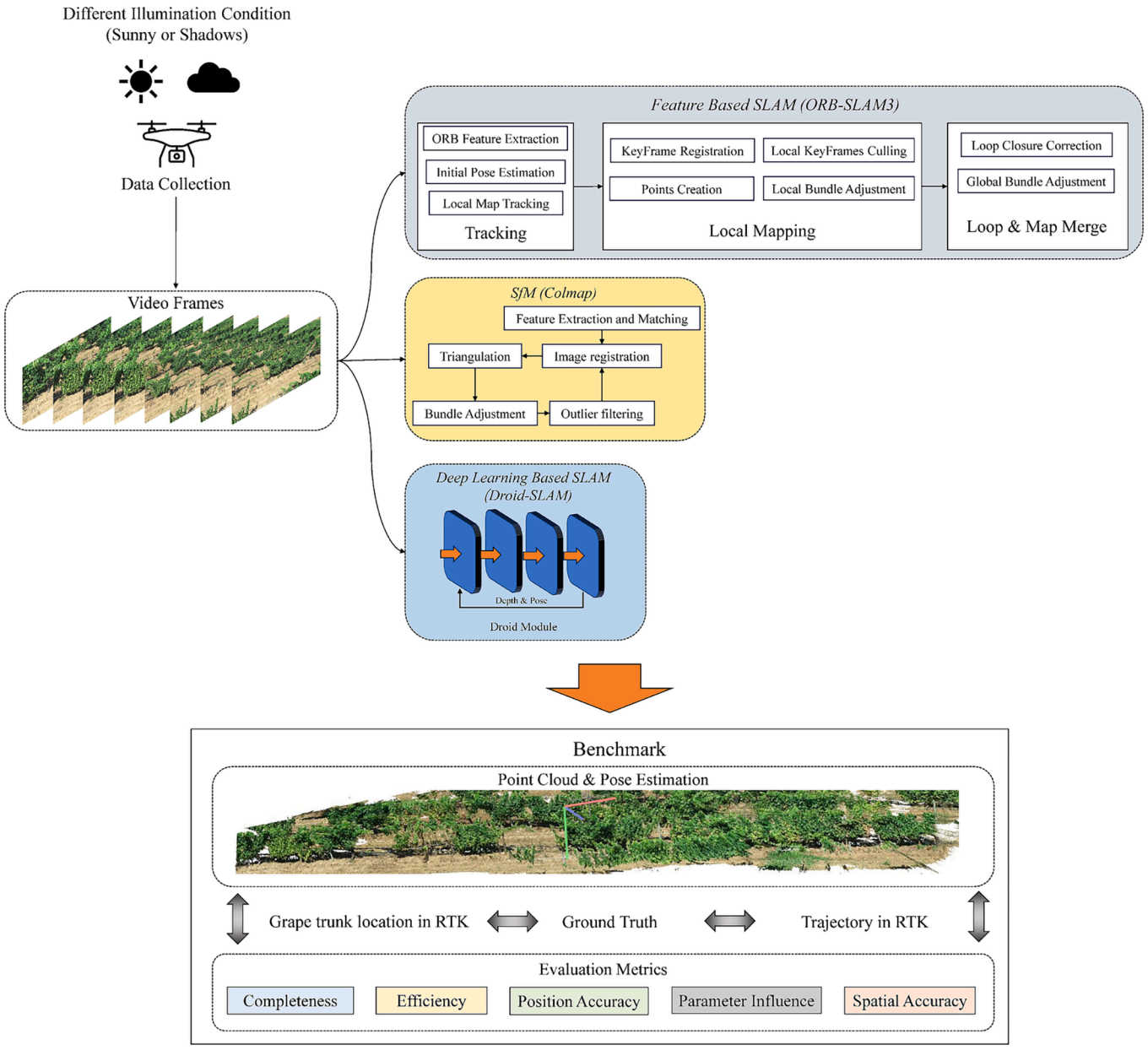


Fig. 1. The overall workflow of the benchmark for SfM (Colmap) and SLAM (ORB-SLAM3 and Droid-SLAM) evaluation in the vineyard under different illumination conditions.

characteristics such as susceptibility to light and small perceptual range cause them to be unsuitable for agricultural applications in field (Li et al., 2022). Monocular cameras, when integrated into the UAV platforms, offer distinctive advantages, such as cost-effectiveness, low power consumption suitable for the constrained battery capacity of UAVs, and a lightweight, compact profile well-suited for the limited payload capacity of UAV platforms, distinguishing them from other sensor modalities (Engel et al., 2014). In orchard management, UAVs typically operate autonomously, following a pre-defined flight path at altitudes exceeding ten meters, with a fixed camera angle to collect RGB images (Ariza-Sentís et al., 2023). After data collection, Structure from Motion (SfM) reconstructs three-dimensional structures of the orchards from a series of photos taken from different and overlapping perspectives. Vegetation modeling results for canopy height using SfM are quite accurate. For instance, Wallace, Lucieer (Wallace et al., 2016) reported a Root Mean Square Error (RMSE) of less than 0.42 m in the horizontal direction and 0.17 m in the vertical direction, similar to the accuracies of LiDAR-based system in the forest. SfM has proven to be an accurate approach for orchard management, providing valuable insights for activities such as pruning, harvesting, yield prediction, and disease detection (Zhang et al., 2021). However, the relatively high flight altitudes result in missing details of the grapevines due to the occlusion by the canopies. Additionally, pre-mapped flight paths may not always be available and a GPS-based navigation system may not work at low altitudes, necessitating the UAV platform to adapt to and account for local variability in the vineyard canopy (Costley and Christensen, 2020). Thus, monitoring crops with UAVs in an inter-row flight mode at low altitudes with robust and accurate localization and mapping is crucial.

Simultaneous Localization and Mapping (SLAM) is a method that enables a system to create a map of an unknown environment by integrating information from sensors while simultaneously determining its position within that environment in real-time (Wang et al., 2024). SLAM operates on the same principle as SfM, which is based on triangulation. However, SfM has certain limitations, including high computational

requirements, long processing times, low real-time performance, and challenges in autonomous navigation within GPS-denied environments (Jiang et al., 2020). On the other hand, SLAM can process more data in real-time and obtain position information simultaneously for robot navigation, making it a promising tool for bridging gaps in SfM for precision agriculture. For instance, a stereo visual SLAM method has been proposed that efficiently works in agricultural scenarios without compromising performance and accuracy compared to other state-of-the-art methods (Islam et al., 2023). This method incorporates an image enhancement technique for ORB point and LSD line features recovery, enabling it to operate in a variety of scenarios and providing extensive spatial information in low-light and hazy agricultural environments. Xiong, Liang (Xiong et al., 2023) proposed a SLAM framework combining a semantic segmentation method with an RGB-D camera based handheld device to generate a semantic map in citrus orchards. Chen, Huang (Chen et al., 2023) developed an autonomous ground robot combining sensor fusion (RGB camera, LiDAR, and IMU), SLAM, and object detection for harvesting pitaya. While many studies have focused on ground robots and sensor fusion for agricultural applications, the challenges differ when applying SLAM to UAVs. Current studies offer some insights into the benchmark of visual SLAM in agriculture. Hroob, Polvara (Hroob et al., 2021) evaluated LIO-SLAM, StaticMapping, ORB-SLAM2, and RTAB-MAP in both LiDAR-based and visual-based methods in a simulated vineyard. However, their focus was limited to trajectory evaluation between these methods using ground robots and did not include evaluations of deep learning-based SLAM methods. Nevertheless, depending on the technique used, challenges in orchard crop modeling can arise due to factors such as leaf gaps, similar plant colors, repetitive structures and random geometric traits of the crops (Moreno and Andújar, 2023). Therefore, it is important to evaluate the effectiveness of each approach.

This paper presents a comprehensive benchmarking study aimed at evaluating the performance of SLAM algorithms using SfM as a benchmark in the specific context of vineyard monitoring with a UAV-mounted monocular camera under various illumination conditions. The main contributions of this paper are (i) collecting vineyard video data between rows using UAV with monocular camera and ground truth using Real-time kinematic positioning (RTK) device in Spain, (ii) proposing evaluation metrics in five aspects (efficiency, completeness, position accuracy, parameter sensitivity, and spatial accuracy at plant level), and (iii) comparing feature-based visual SLAM and deep learning-based visual SLAM using SfM as a benchmark in vineyards under different illumination conditions.

2. Methodology

This section proposed a benchmarking workflow for evaluating UAV-based SfM and visual SLAM in vineyard environments. The workflow included UAV data collection, algorithm implementation, and result analysis with five evaluation metrics. The ground truth contained the trunk location (Vélez et al., 2024) and UAV trajectory in RTK measurement (Fig. 1). This benchmark facilitated evaluation and comparison using the proposed metrics. The next section describes selected SfM and SLAM methods, the evaluation metrics and the experimental setup.

2.1. Study area and data collection

The research was conducted in two different vineyards located in Pontevedra, Spain (41°57′18.3″N 8°47′41.9″W). The grapevines were planted in 1990 with 2.5-meter spacing between plants and 3 m between rows and were aligned in a diagonal direction from northeast to southwest. They were trained to grow vertically using shoot positioning and supported within a vertical trellis system. Each vine was cordon-pruned, leaving between 3 and 6 positions per plant. At the time of data collection, the grapevines were in the late maturity stage, with leaves and grapes in deep green, and some leaves were wilting.

Table 1
Description of flight parameters and operation conditions for the UAV data collection. One collection was in the first vineyard on September 18th, and another collection was in the second vineyard on September 19th.

UAV platform	DJI Phantom 4 RTK, Shenzhen, China
Sensor	RGB FC6310R
Sensor Type	CMOS
Resolution	3840 × 1160
Focal length (mm)	8.8
Flight mode	Manually control within rows
Flight altitude (m)	Around 2
Flight velocity (m/s)	Around 0.5
Video Frame rate (fps)	30
Collection data & start time	Sep 18th, 2023, 03:21 PM to 03: 24 PM, Sep 19th, 2023, 04:53 PM to 05:27 PM (before harvesting)
Wind Speed (m/s)	2.5
Illumination conditions	Sunny with occasional clouds
Temperature (°C)	17

A commercial UAV with a high-resolution monocular camera was used for video collection in the vineyard location (Table 1). The UAV was manually maneuvered to fly between rows of grapevine plants at a distance of approximately 0.5 to 1 m from the plants. Three flight modes were set up to evaluate the performance of the methods with different camera orientations: 1) side view flights, 2) front view flights, and 3) side view flights with a loop closure. The camera settings were set to autofocus to maintain consistent video quality despite illumination changes during the flights. Multiple flights were conducted at similar altitudes, around 2 m above the ground. We also calculated the digitization footprint after the data collection in the areas. The digitization footprint was approximately 2.27 Gb/ha.

To maximize the evaluation of various conditions during inter-row flights in the vineyard, we selected video data under different illumination conditions (shadows or sunlight), shooting directions (front view or side view), and loop closure, respectively (Fig. 2, Table 2). The video data were collected in the same row in the afternoon on September 18th by the same UAV under different illumination conditions.

2.2. Ground truth measurements

In this study, two types of ground truth data were recorded for the vines: 1) position and orientation data of the UAV platform at a frequency of 0.1 s, and 2) the location of each vine trunk. The Real-time Kinematic (RTK) data during each flight were extracted from the UAV's flight record log files and decoded using AirData[1] online. The RTK data considered the ground truth for the UAV's pose and trajectory included latitude, longitude, flight height, speed from x, y and z-axis directions, and the orientation data (compass, pitch, and roll). The location of 43 grapevine trunks was taken using a Trimble R2 Integrated GNSS system with a TSC3 Controller (Trimble Inc., California, USA) that provides centimeter positioning accuracy (Fig. 3).

2.3. Selection of the localization and mapping frameworks

Currently, there are two primary methods for achieving localization and mapping tasks: SfM and SLAM. SfM is an accurate reconstruction method for UAV-based remote sensing (Jiang et al., 2020). Previous studies have shown many approaches stemming from various algorithm architectures, differing in details such as feature extraction, feature matching, and filtering methods (Wang et al., 2024). Our method selection is motivated by the principles underlying these approaches, including SfM, feature-based SLAM, and the more recent deep learning-based SLAM.

[1] https://airdata.com/.

(a) UAV flight position (b) Grapevines under sunlight (c) Grapevines under shadows

Fig. 2. Flight position of UAV in the vineyard during data collection. (a) shows the position in vineyard rows (adapted from (Ariza-Sentís et al., 2024), (b) and (c) shows the vineyard rows under sunlight and cloud shadows, respectively.

Table 2
Characterization and description of the UAV video data evaluated in this study. The elevation and azimuth indicate the sun angle at that moment when collecting data.

Dataset Type	Camera View (within rows)	Illumination Condition	Time/Location/Elevation/Azimuth	Loop Closure	Duration (s)	#Frames
Front_View_Light	Front View	Under sunlight	3:21 PM/41°56′16.20″ −8°47′32.90″/24.31°/248.44°	No	36	1080
Front_View_Dark	Front View	Under shadow	3:22 PM/41° 56′ 15.99″ −8° 47′ 34.32″/24.14°/248.64°	No	18	540
Side_Short_Light	Side View	Under sunlight	3:24 PM/41° 56′ 16.09″ −8° 47′ 34.37″/41° 56′ 16.09/ 23.79°/249.03°	No	30	900
Side_Short_Dark	Side View	Under shadow	3:23 PM/41° 56′ 15.91″ −8° 47′ 35.38″/23.96°/248.83°	No	26	780
Side_Loop	Side View	Under sunlight	4:53 PM/41° 57′ 20.12″ −8° 47′ 39.48″/7.35°/264.81°	Yes	441	13,320
Side_Long	Side View	Under sunlight	5:21 PM/41° 57′ 20.09″ −8° 47′ 40.04″/2.15°/269.51°	No	156	4680

Generally, visual SLAM begins with the detection of distinctive features (feature-based method), or photometric variation (direct methods) followed by the estimation of the camera's pose through tracking the feature or the photometric variation. After detection, a map initialization uses a subset of features and prior knowledge to establish a reference frame. Ongoing mapping involves incremental refinement of the spatial map and optimization of the camera trajectory using local and global optimization methods, such as bundle adjustment, which refines the entire map for consistency. ORB-SLAM3 was selected as the representative method for feature-based SLAM due to its robustness, positional accuracy and efficiency in complex environments (Wang et al., 2024).

Deep learning is an effective means of feature extraction, integrating trained deep models from large datasets into the SLAM framework. For instance, DeTone, Malisiewicz (DeTone et al., 2018) proposed incorporating self-supervised interest points as descriptors for visual SLAM, improving the performance of photometric constancy under challenging illumination conditions. Yang, Stumberg (Yang et al., 2020) introduced a monocular framework involving depth prediction, pose estimation, and uncertainty calculation using a deep network. Teed and Deng (Teed and Deng, 2021) proposed Droid-SLAM, an accurate and robust deep learning-based SLAM system incorporating recurrent iterative camera pose and pixel-wise depth updates through a dense bundle adjustment layer and pose estimation layer. The model is trained on challenging datasets from TartanAir (Wang et al., 2020), which contain various illumination conditions, weather scenarios, and environments. Therefore, Droid-SLAM was selected for comparison in our vineyard datasets.

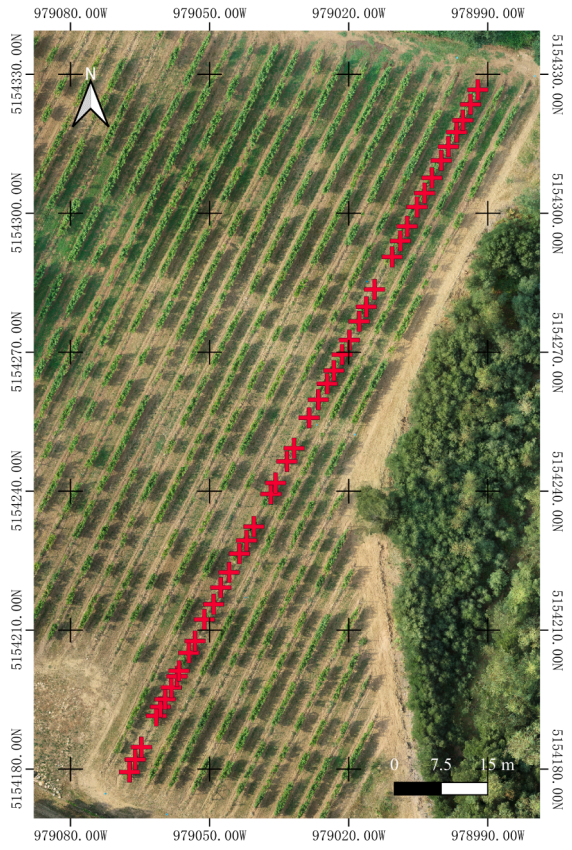
Compared to SLAM methods, SfM is based on photogrammetric approaches from remote sensing and computer vision. Several SfM software implementations exist, with Colmap outperforming in terms of efficiency, point cloud completeness and bundle adjustment accuracy compared to AliceVision, VisualSfM, RealityCapture, Metashape, and Pix4mapper in UAV-based large scenarios (Jiang et al., 2020). Based on their research, Colmap was selected as the representative tool for the SfM method in this paper.

The property of each selected method is listed as follows:

- Colmap 3.8 (Schonberger and Frahm, 2016): Copmap is an open-source SfM solution that provides both a command line and a graphical interface. It supports sparse and dense incremental reconstruction from multiple data sources with different camera models (e.g.; pinhole, radial, fisheye). Colmap integrates multiple feature-matching strategies and two adjustment solvers, SBA and Ceres, to adapt to various kinds of data and meet the diverse requirements of users.
- ORB-SLAM3 (Campos et al., 2021): ORB-SLAM3 is an open-source framework for visual, visual-inertial, and multi-map SLAM. It can perform SLAM with monocular, stereo, and RGB-D cameras, using pinhole and fisheye lens models. As a feature-based SLAM system, ORB-SLAM3 relies fully on Maximum-a-Posteriori (MAP) estimation. It can operate robustly in real-time in both small and large, indoor and outdoor environments, and is 2 to 5 times more accurate in localization than previous approaches (e.g., Mono-SLAM, SVO, LSD-SLAM). The system can reuse all previous information at any processing stage, allowing the inclusion of co-visible keyframes in bundle adjustment. It provides high parallax observations and boosts accuracy, even if they are widely separated in time or come from a previous mapping session.
- Droid-SLAM (Teed and Deng, 2021): Droid-SLAM is a deep learning-based SLAM system capable of performing SLAM with monocular, stereo, and RGB-D cameras using a Dense Bundle Adjustment layer. The system is recurrent and iterative, allowing it to update camera pose and pixel-wise depth. Droid-SLAM is accurate and robust, suffering from substantially fewer failures than previous systems.

2.4. Camera calibration

To implement SLAM, camera intrinsic parameters including focal length, principal point, and lens distortion coefficients from the DJI Phamtom4 RTK, were measured and calibrated using checkerboard patterns (Fig. 4). A one-minute and 19-second video was recorded using the UAV's camera, maintaining a distance of 20 to 30 cm from the checkerboard, with the camera positioned at various angles.



(a) The grapevine trunk location measured by the RTK device

(b) RTK device used in grapevine trunk location

Fig. 3. Grapevine trunk location measurement by RTK device manually for datasets from 19th Sep. (a) shows the position of each grapevine trunk with red cross symbol, and (b) shows the integrated GNSS system used to measure grapevine trunks ((adapted from Vélez et al., 2024).

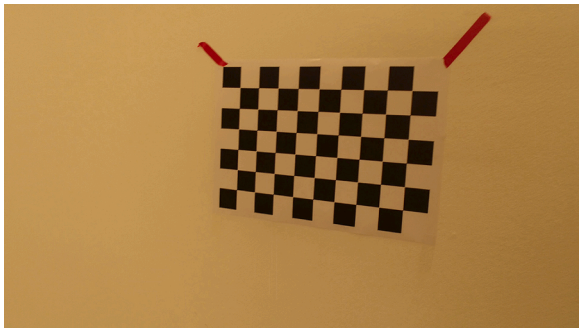


Fig. 4. Checkerboard used for camera setting calibration.

Then, a series of pictures of the checkerboard were taken using the same camera setup, and a camera calibration script in Python with OpenCV was used to derive the intrinsic parameters of the camera. This included the camera matrix with focal lengths and principal point (Equation (1), as well as the distortion coefficients (Equations (2) and (3).

$$K = \begin{bmatrix} f_x & 0 & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{bmatrix} \quad (1)$$

Table 3
Camera calibration parameters in ORB-SLAM3 and Droid-SLAM. Fx, fy, cx, and cy are intrinsic matrices, and k1, k2, p1, p2, and k3 are distortion factors.

Parameters		Value
Camera Type		Pinhole
Calibration and distortion parameters		fx: 2684.79393
		fy: 2686.71103
		cx: 1913.21165
		cy: 1049.37983
		k1: 0.00358061
		k2: −0.12530172
		p1: −0.00161459
		p2: −0.00149304
		k3: 0.20271981
Camera resolution		3840 W 2160H
ORB SLAM Parameter Setup	ORB Feature point number	2000
	ORB scale factor	1.2
	ORB nLevel	1
Droid SLAM Parameter Setup	Beta	0.3
	Filter_thresh	2.4
	Warmup	8
	Keyframe_thresh	4.0
	Frontend_thresh	16.0
	Frontend_window	25
	Frontend_radius	2
	Frontend_nms	1
	Backend_thresh	22.0
	Backend_radius	2
	Backend_nms	1

$$x_{dis} = x\left(1 + k_1 r^2 + k_2 r^4 + k_3 r^6\right) y_{dis} = y\left(1 + k_1 r^2 + k_2 r^4 + k_3 r^6\right) \quad (2)$$

$$x_{dis} = x + \left[2p_1 xy + p_2\left(r^2 + 2x^2\right)\right] y_{dis} = y + \left[p_1\left(r^2 + 2y^2\right) + 2p_2 xy\right] \quad (3)$$

In equations (1), 2 and 3, , x, y are the undistorted pixel location, x_{dis}, y_{dis} are the distorted pixel location, f_x, f_y are focal lengths, c_x, c_y are principal points, and k_1, k_2, k_3, p_1, p_2 are the distortion coefficient.

The measured intrinsic parameters of the camera are given in Table 3.

2.5. Evaluation metrics

In previous research, a comprehensive evaluation of UAV-based SfM solutions was conducted (Jiang et al., 2020). The authors introduced three metrics (efficiency, completeness, and accuracy) to evaluate the performance of six SfM software implementations. To extend these metrics to complex environments, such as vineyard, we retained efficiency and point cloud completeness. Then, the accuracy from the previous study was modified specifically to the position accuracy of the UAV. Moreover, two new metrics were introduced: parameter sensitivity in SLAM and spatial accuracy at the plant level.

Therefore, five metrics were proposed to evaluate UAV-based localization and mapping methods in vineyard environments.

- Efficiency

The efficiency is evaluated based on the process time and the hardware acceleration method used (CPU + GPU, or CPU only). Processing time is measured as the duration required to complete the entire SfM or SLAM process, from data input to point cloud generation. Hardware acceleration methods, such as GPU acceleration, are also considered. This metric assesses the process speed and resource utilization of the SfM and SLAM methods.

- Completeness

The completeness is evaluated based on the number of points in the point cloud generated by different methods and the number of input video frames used during the processing. Point cloud completeness evaluates how well the SLAM or SfM systems capture the entire environment, including all relevant structures and features.

- Position Accuracy

Position accuracy is evaluated based on the camera's pose estimation error compared to the ground truth. Two quantitative metrics, Absolute Pose Error (APE) and Relative Pose Error (RPE), are used to measure the accuracy of the UAV's position. Global consistency is assessed by investigating the absolute distance error between the estimated pose and ground truth (Equation (4)). Local accuracy over a fixed time interval is measured by the RPE (Equation (5)). In this study, we implemented evo,[2] an open-source Python package, to calculate the APE and RPE between estimated poses and ground truth. In addition, monocular SLAM reconstruction lacks absolute scale information, necessitating alignment of the quaternions of the camera poses using the Kabsch-Umeyama algorithm before comparison with ground truth (Lawrence et al., 2019). This algorithm calculates the optimal rotation matrix to minimize the root mean squared deviation (RMSD) between paired sets of points, enabling alignment of translation and rotation matrices between ground truth and estimated trajectories. With this algorithm, the translation and rotation matrix can be aligned between the ground truth trajectories and estimated trajectories.

$$APE = \sqrt{\frac{1}{n}\sum_{i=1}^{n} \left\| trans\left(Q_i^{-1} S P_i\right) \right\|^2} \quad (4)$$

$$\mathrm{RPE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} \left\| trans\left(Q_i^{-1} Q_{i+\Delta}\right)^{-1} \left(P_i^{-1} P_{i+\Delta}\right) \right\|^2} \quad (5)$$

where $Q_i \in SE_3$ are a sequence of ground truth poses, $P_i \in SE_3$ are a sequence of estimated poses, S is the rigid-body transformation, which is transformation of the camera on the UAV here. This metric is crucial for creating precise maps and ensuring the UAV follows the intended flight path.

- Parameter Sensitivity

This metric evaluates the impact of various algorithmic parameters on the performance of visual SLAM systems in a vineyard environment. Two studies investigated the impact of parameters in SLAM implementations, highlighting that sensitivity analysis can help to fine-tune the approach and improve reliability in specific applications (Norzam et al., 2019; Abdelrasoul et al., 2016). Furthermore, a recent study showed that parameter tuning is the key to optimizing the performance of SLAM performance (Sossalla et al., 2022). Moreover, the performance of SLAM under different parameters in an orchard environment is crucial for algorithm implementation and suitable algorithm selection. By systematically adjusting parameters such as feature detection thresholds, matching criteria, and optimization settings, we analyze how these changes affect the overall accuracy, robustness, and efficiency of the SLAM process. Understanding parameter influence helps identify optimal configurations that maximize system performance and adaptability to the unique challenges posed of vineyard terrains, such as varying illumination conditions and dense foliage.

- Spatial Accuracy at plant level

Traditional metrics in SLAM such as APE, RPE might not fully capture the nuances required for specialized agricultural applications such as pruning, harvesting. Thus, this metric was introduced to address a specific aspect of spatial performance that is crucial for orchard management. The metric is evaluated based on the locations of each trunk at the plant level collected with the GNSS high accuracy receiver. Spatial accuracy is measured as the relative distance error between the estimated location of each grapevine trunk and the RTK measurement (Equation (6)). Every point cloud map was processed using CloudCompare to determine the location of each grapevine trunk in the point cloud. The canopies and ground were cropped, retaining only the grapevine trunks, after which the distances between each trunk were measured.

$$RMSE_{crop} = \sqrt{\frac{1}{m}\sum_{i=1}^{m} \left(y_i - \widehat{y}_i\right)^2} \quad (6)$$

where y_i is the ground truth relative distance between each two grapevine trunks by RTK device, and $\widehat{y}_i$ is the estimated relative distance between each two grapevine trunks by SfM and SLAM.

2.6. Experimental setup

For the six selected videos (Table 2), we reduced the frame rate to 10 FPS due to the high frame rate of the original videos, and the extremely high processing time required by the SfM-based method. For the datasets Front_View_Light, Front_View_Dark, Side_Short_Light, Side_Short_Dart and Side_Long, we used the parameter setup as shown in Appendix. We followed the steps of SfM (Schonberger and Frahm, 2016), which include feature extraction, feature matching, sparse reconstruction,

[2] github.com/MichaelGrupp/evo.

Table 4
Efficiency performance statistics in two aspects (time cost during each procedure and hardware acceleration method). The hardware acceleration method includes CPU and GPU when implementing Colmap-SfM, ORB-SLAM3, and Droid-SLAM. The bolded fonts and numbers indicate the best performance in different processing methods, and '#' means number.

Methods		Front_View_Light	Front_View_Dark	Side_Short_Light	Side_Short_Dark	Side_Loop	Side_Long	Hardware Acceleration
Colmap	#Input Image	**360**	**180**	**257**	**300**	**4396**	**1560**	CPU and GPU
	Feature Extraction	1 min 31 s	46 s	1 min 2 s	1 min 4 s	20 mins 30 s	9 mins 11 s	
	Feature Matching	48 s	18 s	36 s	37 s	11 mins 18 s	5 mins 21 s	
	Sparse Reconstruction	33 mins 19 s	6 mins 15 s	13 mins 26 s	11 mins 33 s	132 mins 30 s	454 mins 33 s	
	Total time in Sparse Reconstruction	35 mins 38 s	7 mins 19 s	15 mins 4 s	13 mins 14 s	164 mins 18 s	469 mins 5 s	
	Image Distortion	50 s	23 s	46 s	46 s		19 mins 45 s	
	Stereo	127 mins 28 s	64 mins 45 s	95 mins 44 s	98 mins 6 s		560 mins 4 s	
	Dense Fusion	7 mins 56 s	1 min 49 s	4 mins 12 s	7 mins 12 s		226 mins 18 s	
	Total Time	171 mins 52 s	74 mins 16 s	115 mins 46 s	119 mins 18 s		1275 mins 2 s	
ORB-SLAM3	#Input Image	1080	539	770	900	13,188	4680	**CPU**
	Time Cost	**4 mins 53 s**	**2 mins 15 s**	4 mins 47 s	**4 mins 59 s**	**22 mins 4 s**	**7 mins 45 s**	
Droid-SLAM	#Input Image	1080	539	770	900	13,188	4680	CPU and GPU
	Time Cost	5 mins 39 s	3 mins 2 s	**4 mins 35 s**	5 mins 5 s	79 mins 56 s	28 mins 52 s	

stereo reconstruction, and data fusion.

Regarding the dataset Side_Loop, which is a long video containing more than ten thousand frames, there are challenges in using the same sparse reconstruction method as with the previous video datasets. The processing time increases exponentially when using incremental mapping in sparse reconstruction due to the large data size. Therefore, we used the hierarchical mapping method to address the issue. Hierarchical mapping divides a large dataset into several smaller datasets through feature extraction and feature matching, then performs sparse reconstruction on each segmented dataset, and finally fuses the sparsely reconstructed camera poses and point clouds using the matched features from each separate dataset (Xu et al., 2021).

ORB-SLAM3 and Droid SLAM require a series of parameters, including camera calibration parameters, feature point settings, and scale factor. The same camera calibration parameters were used to calibrate the camera and initialize ORB-SLAM3 and Droid-SLAM (Table 3). We implemented all three methods in the same hardware configuration with Intel Core i9-10940X CPU, NVIDIA Titan RTX GPU and 64 G memory.

3. Results

According to our proposed benchmark workflow, the efficiency performance, point cloud completeness, and UAV positioning accuracy were evaluated using default parameter settings and the same camera intrinsic parameters (Table 3). Different parameter settings for ORB-SLAM3 and Droid-SLAM were also evaluated to demonstrate the influence of these parameters. Additionally, we explored the potential for spatial accuracy at the plant level.

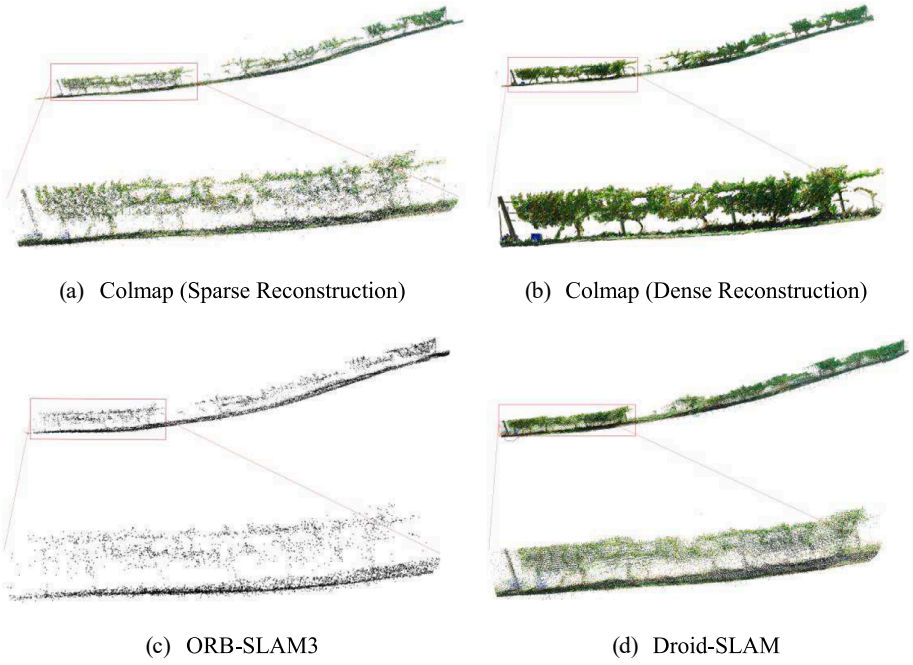
3.1. Efficiency performance

Although the number of input images in the SfM approach is less than in SLAM, SLAM has a shorter processing time. It should be noted that for large datasets with more than 2000 images, the processing efficiency of SfM decreases exponentially as the number of images increases (Table 4). For example, when we processed the dataset Side_Loop using the same incremental processing method (Özyeşil et al., 2017) as we used for datasets Front_View_Light, Front_View_Dark, Side_Short_Light, Side_Short_Dark and Side_Long, the processing time in the sparse reconstruction step exceeded one week and eventually failed due to running out of memory. Therefore, for the Side_Loop dataset, the incremental processing method was replaced with the hierarchical mapper method (Xu et al., 2021) in Colmap. Side_Loop was divided into nine sub-datasets using a hierarchical mapper, and sparse reconstruction was conducted on each.

To reduce the deviation of the results, we conducted the experiment three times under the same configuration. The deviation of the results in Appendix was below 5 %, which is within acceptable limits. However, for the Front_View_Dark dataset, in the implementation of the ORB-SLAM3, two of the three experiments showed tracking failures, which showed less robust than Droid-SLAM. In five additional datasets, ORB-SLAM3 demonstrates a considerable advantage over both sparse and dense reconstruction methods in SfM with respect to computational efficiency. Moreover, it outperforms Droid-SLAM in terms of hardware-accelerated calculations, substantially lowering the requirement for specialized hardware (Table 4). The advantage in processing speed makes ORB-SLAM3 particularly suitable for time-sensitive agricultural activities. Regarding illumination conditions, the processing time of SfM is faster under shadowed conditions compared to sunlight in both side and front views. However, this difference is not significant in the two SLAM approaches.

Table 5
Completeness assessment in six datasets for three methods (Colmap-SfM, ORB-SLAM3, and Droid-SLAM). '#' means number, and the values of sparse point, and dense point indicate the point number in the point cloud. The bolded numbers indicate the best performance in the same dataset using different method implementations.

Datasets		Front_View_Light	Front_View_Dark	Side_Short_Light	Side_Short_Dark	Side_Loop	Side_Long
Colmap-SfM	#Input image	**360**	**180**	**257**	**300**	**4396**	**1560**
	#Sparse Point	86,404	63,628	170,458	172,299	1,662,063	918,266
	#Desnse Point	**13,984,112**	**8,376,848**	**15,457,578**	**17,700,649**		**87,104,785**
ORB-SLAM3	#Input image	1080	539	770	900	13,188	4680
	#Sparse Point	5079	5426	16,621	12,743	200,480	89,423
Droid-SLAM	#Input image	1080	539	770	900	13,188	4680
	#Sparse Point	11,987	37,610	129,055	114,904	1,428,106	630,952



(a) Colmap (Sparse Reconstruction) (b) Colmap (Dense Reconstruction)

(c) ORB-SLAM3 (d) Droid-SLAM

Fig. 5. Visualization for the vineyard point cloud of Side_Short_Light dataset in (a) Colmap-SfM (Sparse), (b) Colmap-SfM (Dense), (c) ORB-SLAM3 and (d) Droid-SLAM. The red bounding boxes are the same zoom areas for the vineyard.

Table 6
The RMSE of absolute pose error (APE) and relative pose error (RPE) of the UAV in six vineyard datasets. The path length shows the total distance of the UAV trajectory during the flight. The bolded numbers indicate the best performance in the same dataset using different method implementations.

Dataset	Path length (m)	APE RMSE (m)			RPE RMSE (m)		
		Colmap-SfM	ORB-SLAM3	Droid-SLAM	Colmap-SfM	ORB-SLAM3	Droid-SLAM
Front_View_Light	44.8	0.33	**0.15**	0.36	0.25	**0.19**	0.22
Front_View_Dark	30.6	0.86	**0.72**	0.78	0.33	**0.27**	**0.27**
Side_Short_Light	38.4	0.66	**0.16**	0.58	**0.31**	0.33	**0.31**
Side_Short_Dark	35.3	0.61	0.30	**0.29**	**0.26**	**0.26**	0.27
Side_Loop	293.8	5.51	**5.16**	7.50	**0.10**	**0.10**	**0.10**
Side_Long	126.9	2.36	**0.48**	0.80	**0.14**	0.15	0.15

3.2. Point cloud completeness

In terms of completeness of the vineyard reconstruction, SfM holds a significant advantage over SLAM approaches. SfM can achieve higher point cloud resolution with fewer images compared to SLAM, for both front and side view datasets. Additionally, SfM can generate accurate depth images for each frame through stereo vision matching, resulting in precise and dense point cloud reconstructions of the vineyard. Interestingly, in front view datasets, SfM outperforms Droid-SLAM by 86.1 % (under sunlight) and 40.9 % (under shadow) respectively in terms of completeness in sparse reconstruction (Table 5). This difference may be attributed to the fact that side view data contain more crop features per frame, while front view data often include large areas of land, sky, and shadows, resulting in fewer features (Fig. 2). Furthermore, for all three methods, the number of point clouds generated for side view datasets is significantly higher than for front view datasets, indicating the presence of more features in the side view.

To provide a visual representation of the completeness (Table 5) of the point cloud generated by the three methods, Fig. 5 illustrates the results for Dataset Side_Short_Light and selects the same areas to show more details. The completeness of dense reconstruction from Colmap is evident, capturing all details in the vineyard, including grape stems, berries, canopies, soils, weeds, and other tools with high accuracy (Fig. 5b). In sparse reconstruction, Colmap performs similarly to Droid-SLAM, capturing the grape trunk and most of the canopy, though smaller details like individual grapes are not observed. ORB-SLAM3 exhibits the poorest performance, with the fewest points and no color information, attributable to its reliance on tracking ORB features for pose estimation and mapping, and its support for grayscale images only in the original implementation of ORB-SLAM3.

3.3. UAV position accuracy

The results in Table 6 show that the position accuracy of the UAV in SfM is superior to SLAM-based methods in estimating RPE regardless of illumination conditions and camera positions. However, for long-distance and large-scale data (Side_Loop and Side_Long), the advantage of SfM in RPE is not as pronounced as with SLAM methods. ORB-SLAM3 demonstrates high accuracy performance in the RMSE of APE across different illumination conditions and camera positions. Nevertheless, in the vineyard dataset, SfM, ORB-SLAM3, and Droid-SLAM all exhibit subpar performance in longer datasets containing loop closures.

In the comparison of UAV trajectories generated by ORB-SLAM3, Droid-SLAM, and Colmap across various illumination conditions, distinct performance can be observed (Fig. 6). When tested in simple flight mode with a straight-line flight under both light and dark conditions, ORB-SLAM3 performs better than other methods compared to ground truth with little drift. In the more complex long and loop trajectory (Fig. 6e and Fig. 6f), it is apparent that the drift error in the trajectory accumulates over time for both SfM and SLAM approaches in

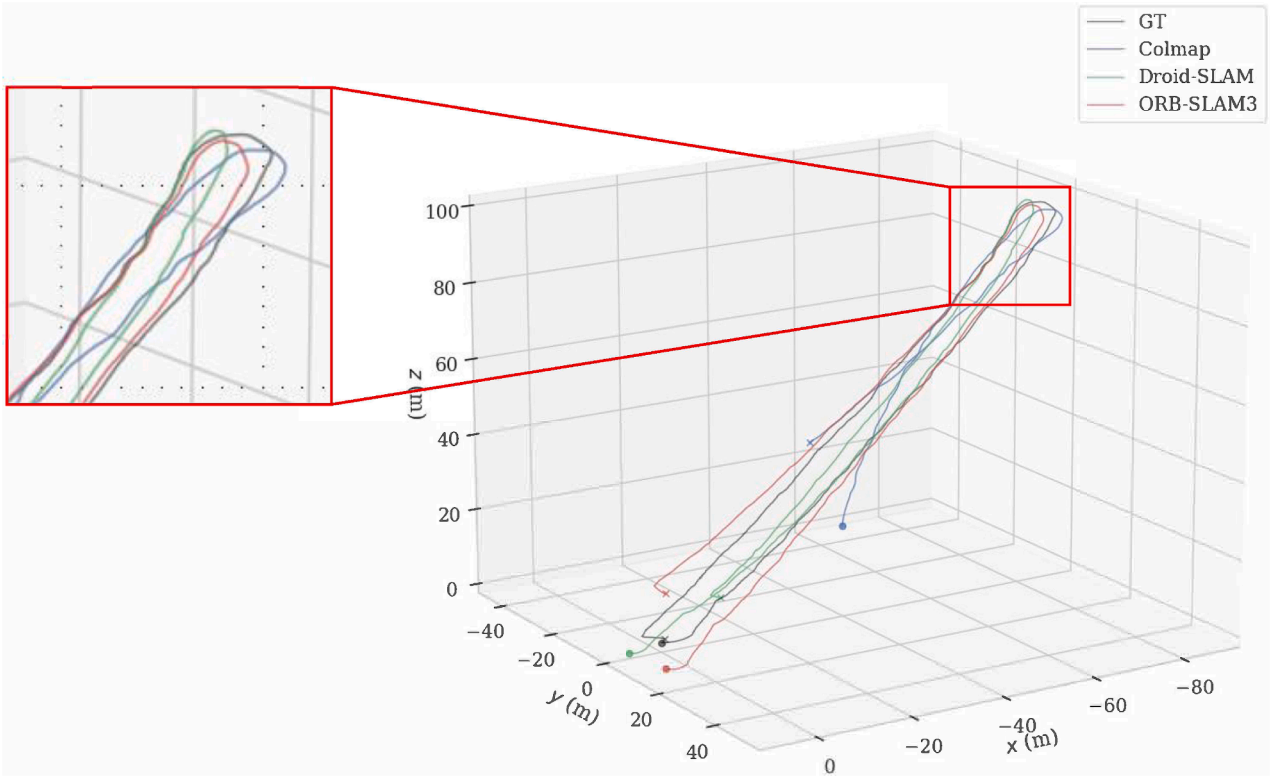


Fig. 6. Comparison of the trajectories with three methods to ground truth. "GT" means the ground truth of the trajectories, the dots indicate the starting point of the UAV flight, and the "x" shapes with red, green, blue and brown indicate the end point of the UAV flight, the red box shows the zoomed-in area of the corner of the turn of the UAV's trajectory.

Table 7
The influence of the number of ORB feature points in ORB-SLAM3 in efficiency, point cloud completeness, and positional accuracy in APE. "Tracking Error" means tracking errors and program failures during ORB-SLAM3 implementation. The completeness shows the point number in the point cloud.

Dataset	#ORB Feature point	Efficiency	Completeness	Positional Accuracy (APE) (meter)
Front_View_Dark	500	Tracking Error	Tracking Error	Tracking Error
	5000	1 min 34 s	5703	0.72
	10,000	Tracking Error	Tracking Error	Tracking Error
	50,000	Tracking Error	Tracking Error	Tracking Error
Front_View_Light	500	Tracking Error	Tracking Error	Tracking Error
	5000	5 mins 9 s	7758	0.12
	10,000	5 mins 14 s	6459	0.12
	50,000	Tracking Error	Tracking Error	Tracking Error
Side_View_Dark	500	Tracking Error	Tracking Error	Tracking Error
	5000	5 mins 13 s	15,482	0.30
	10,000	5 mins 53 s	16,030	0.31
	50,000	Tracking Error	Tracking Error	Tracking Error
Side_View_Light	500	Tracking Error	Tracking Error	Tracking Error
	5000	4 mins 54 s	23,320	0.16
	10,000	5 mins 25 s	20,171	0.16
	50,000	Tracking Error	Tracking Error	Tracking Error

Table 8
The influence of the scale factor in the performance of ORB-SLAM3. "Tracking Error" means tracking errors and program failures during ORB-SLAM3 implementation. The completeness shows the point number in the point cloud.

Dataset	scale factor	Efficiency	Completeness	Position Accuracy (APE) (meter)
Front_View_Dark	1.1	2 mins 19 s	6033	0.7165
	1.5	2 mins 1 s	5379	0.7174
	1.8	1 min 57 s	5783	0.7157
Front_View_Light	1.1	4 mins 52 s	4987	0.1600
	1.5	Tracking Error	Tracking Error	Tracking Error
	1.8	Tracking Error	Tracking Error	Tracking Error
Side_View_Dark	1.1	4 mins 56 s	12,369	0.3056
	1.5	4 mins 25 s	12,381	0.3046
	1.8	4 mins 13 s	12,319	0.3042
Side_View_Light	1.1	4 mins 46 s	17,136	0.1611
	1.5	4 mins 43 s	16,912	0.1607
	1.8	4 mins 4 s	15,925	0.1620

large-scale datasets, resulting in misalignment between the start and end points. This accumulation of drift error can cause significant inaccuracies in point cloud reconstruction. Specifically, in the Side_Loop dataset, where the camera captures a row of grapevines, the drift error in the trajectory estimation leads to an inaccurate camera position, ultimately resulting in the misalignment of the two sides of the grapevine row in the point cloud display. Table 6 shows that ORB-SLAM3 significantly outperforms Droid-SLAM in handling large-scale datasets.

3.4. Parameters sensitivity in SLAM

In the previous sections, we presented the results of SfM and SLAM with their default parameter settings. In this section, we explore the

Table 9
The influence of the number of levels in the scale pyramid in the performance of ORB-SLAM3. "Tracking Error" means tracking errors and program failures during ORB-SLAM3 implementation. The completeness shows the point number in the point cloud.

Dataset	#levels	Efficiency	Completeness	Position Accuracy (APE)
Front_View_Dark	3	2 mins 33 s	5548	0.7273
	5	2 mins 39 s	4748	0.7274
	8	2 mins 53 s	3026	1.1966
Front_View_Light	3	6 mins 19 s	4265	0.1030
	5	7 mins 14 s	3961	0.1047
	8	Tracking Error	Tracking Error	Tracking Error
Side_View_Dark	3	6 mins 37 s	12,796	0.2916
	5	7 mins 55 s	11,538	0.3153
	8	9 mins 19 s	10,656	0.3150
Side_View_Light	3	5 mins 56 s	15,729	0.1697
	5	6 mins 33 s	14,724	0.1797
	8	7 mins 19 s	14,495	0.1743

performance variations of ORB-SLAM3 and Droid-SLAM by adjusting a range of parameters. For ORB-SLAM3, we experimented with the number of ORB feature points (ranging from 500 to 50000), the scale factor (ranging from 1.1 to 1.8), and the number of levels in the scale pyramid (ranging from 3 to 8) (Table 7, Table 8, and Table 9). The number of ORB feature points significantly influences the robustness of ORB-SLAM3. Both excessively low and excessively high values can lead to SLAM tracking failure (Table 7). As the number of ORB feature points increases, the computational efficiency decreases, the point cloud completeness diminishes, and the position accuracy worsens. These observations suggest that while a higher number of feature points can provide more data for tracking and mapping, it also imposes a greater computational burden, which negatively impacts the overall performance of ORB-SLAM3.

In terms of the scale factor, the results indicate that while increasing the scale factor can enhance processing efficiency, it may negatively impact completeness and does not significantly improve position accuracy. The impact on point cloud completeness and position accuracy is within 10 % (Table 8). The optimal scale factor may depend on the specific dataset and the desired balance between processing speed and reconstruction quality.

The analysis of the influence of the number of levels in the scale pyramid on ORB-SLAM3′s performance reveals significant impacts on efficiency, completeness, and position accuracy (Table 9). As the number of pyramid levels increases, the processing time also rises, indicating a direct correlation between the complexity of the scale pyramid and computational load. For instance, in the Front_View_Light dataset, the processing time escalated from 6 min 19 s with three levels to 8 min 52 s with eight levels. This pattern of increased processing time is consistent across all datasets examined. Furthermore, the completeness of the point cloud decreases markedly with higher pyramid levels, as seen in the Front_View_Dark dataset, where completeness declined from 5548 points at three levels to 3026 points at eight levels. This reduction suggests a significant loss in the number of reconstructed points, which is critical for accurate spatial representation. Position accuracy, measured by Absolute Pose Error (APE), also deteriorates with an increasing number of levels, with the Front_View_Dark dataset showing an APE increase from 0.7273 to 1.1966. The negative impact on position accuracy is notably severe in some cases, such as the Front_View_Light dataset, where the algorithm fails at eight levels. These findings underscore the trade-off between computational efficiency and the quality of reconstruction, highlighting the need for careful selection of scale pyramid levels to optimize performance in specific scenarios.

Droid-SLAM has more types of parameters than ORB-SLAM3. According to the original paper (Teed and Deng, 2021), three parameter combinations were established for three public datasets, including eth3d (Schops et al., 2017), euroc (Burri et al., 2016), and tum (Schubert et al., 2018) (Table 10).

The performance of the three parameter setups (Setup1, Setup2, and Setup3) in Droid-SLAM was evaluated across four datasets, namely Front_View_Dark, Front_View_Light, Side_View_Dark, and Side_View_Light (Table 11). Across the Front_View_Dark dataset, Setup3 consistently exhibited the highest completeness, albeit accompanied by the highest APE values. Conversely, Setup2 demonstrated a commendable balance between completeness and APE, outperforming Setup1 in both metrics. Similarly, in the Front_View_Light dataset, Setup3 maintained higher completeness, while Setup1 offered superior position accuracy. Notably, Setup2 consistently displayed competitive performance across datasets, showcasing efficient processing times alongside commendable completeness and APE values. In the Side_View_Dark dataset, all setups performed comparably in terms of position accuracy, with Setup3 achieving the highest completeness. However, Setup2 emerged as the

Table 11
The performance of the parameters setup1, setup2, and setup3 using Droid-SLAM in four datasets. The completeness shows the point number in the point cloud.

Dataset	Parameters	Efficiency	Completeness	Position Accuracy (APE) (meter)
Front_View_Dark	Setup1	2 mins 50 s	26,435	2.10
	Setup2	2 mins 43 s	52,048	0.78
	Setup3	2 mins 35 s	62,751	3.08
Front_View_Light	Setup1	5 mins 20 s	26,359	0.32
	Setup2	5 mins 13 s	17,776	0.76
	Setup3	4 mins 52 s	34,542	1.23
Side_View_Dark	Setup1	4 mins 45 s	135,988	0.30
	Setup2	4 mins 11 s	121,901	0.30
	Setup3	4 mins 43 s	156,018	0.30
Side_View_Light	Setup1	4 mins 14 s	153,695	0.59
	Setup2	3 mins 46 s	139,326	0.58
	Setup3	4 mins 19 s	180,232	0.60

Table 10
Different combinations of parameter setup in Droid-SLAM. The setup1, setup2, and setup3 are followed by eth3d, euroc, and tum datasets in the original paper.

Value	Parameter Setup										
	Beta	Filter_thresh	Warmup	Keyframe_thresh	frontend				backend		
					thresh	window	radius	nms	thresh	radius	nms
Setup1	0.5	2.0	8	3.5	16.0	16	1	0	22	2	3
Setup2	0.3	2.4	15	3.5	17.5	20	2	1	24.0	2	2
Setup3	0.6	1.75	12	2.25	12.0	25	2	1	15.0	2	3

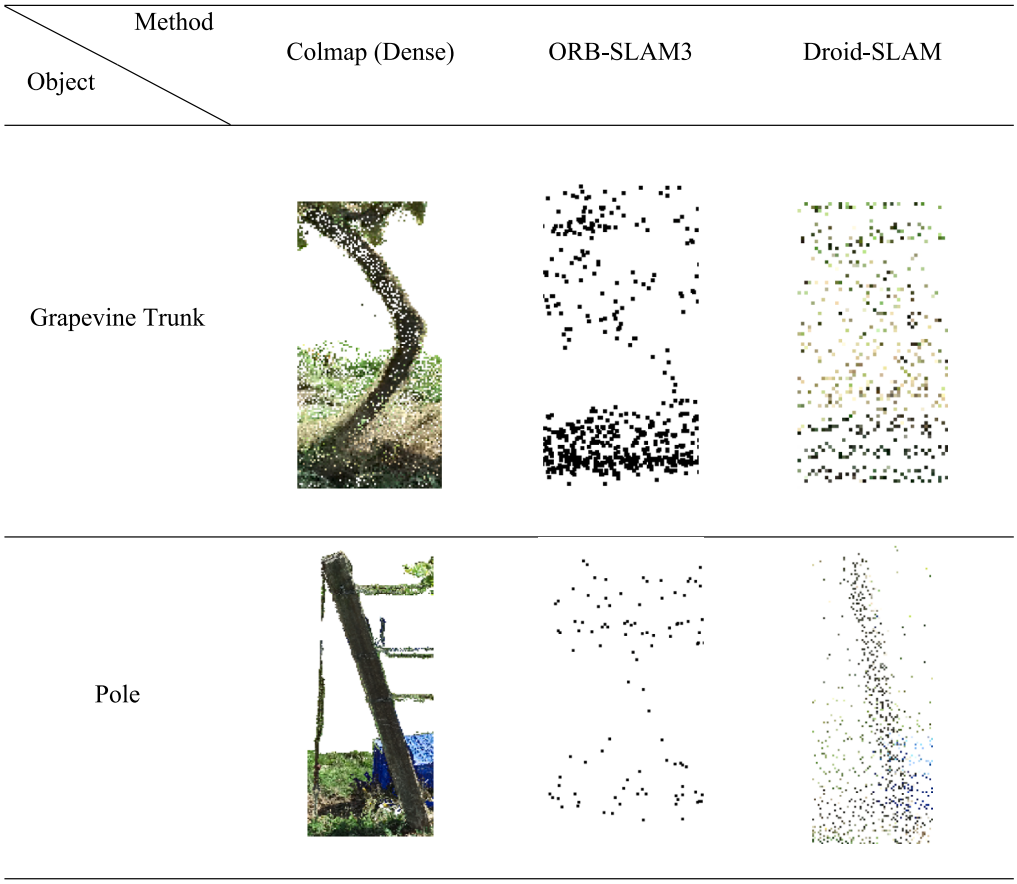


Fig. 7. Point clouds generated using Colmap (Dense), ORB-SLAM3, and Droid-SLAM of the grapevine trunk and pole from the Side_View_Dark dataset.

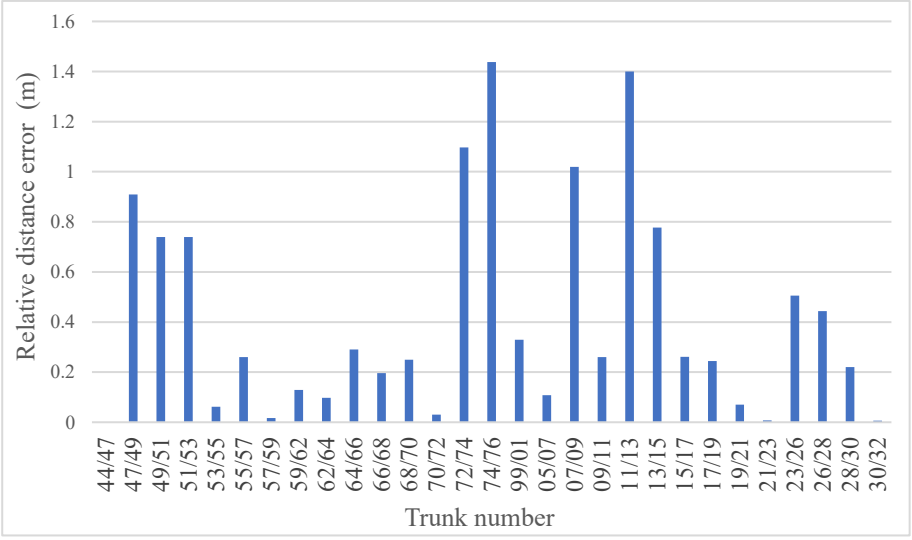


Fig. 8. The relative distance error between each grapevine trunk position using Dense Colmap from Side_View_Long dataset. The horizontal coordinate is the ID of the grapevine trunks.

most efficient option in the Side_View_Light dataset, delivering competitive completeness and APE metrics within a notably shorter processing time.

3.5. Spatial accuracy at plant level

To evaluate spatial accuracy at the plant level, a post-process step was employed to identify the location of grapevine trunks within the vineyard's point cloud using CloudCompare. Initially, the targeted row of the vineyard was segmented from the entire point cloud, following which the ground and canopy regions were cropped out, leaving behind the remaining points, predominantly representing grapevine trunks. Utilizing the dense reconstruction generated by Colmap, the most discernible and unobstructed grapevine trunks were successfully

(a) Grapevine trunk (b) Grape stem (c) Grape canopy

Fig. 9. Point cloud of the grape phenotype of trunk, stem, and canopy in Dense Colmap-SfM from dataset Side_View_Long.

extracted from the point cloud. However, in ORB-SLAM3 and Droid-SLAM, the point density was insufficient for accurate object recognition when compared to dense reconstruction from Colmap by a naked-eye evaluation (Fig. 7). Consequently, only grapevine trunks identified in the dense reconstruction were considered for the assessment of spatial accuracy. To evaluate the spatial accuracy, the relative distance error between each grapevine trunk was calculated (Fig. 8). The total RMSE of the relative distance was 0.5945 m in Dense Colmap.

4. Discussion

This study presents a benchmarking framework to assess three UAV-based monocular localization and mapping methods within vineyard environments across varying illumination conditions. As our previous study has shown, most SLAM research focused on obtaining better performance on public datasets, but there is a lack of targeted evaluation scenarios (Wang et al., n.d.). Thus, our findings indicated variability in the performance of non-deep learning and deep learning methods across the evaluated aspects under diverse conditions in agricultural environment. Therefore, the effect on illumination conditions and how to select the most suitable method for the corresponding application needs to be further discussed.

4.1. The influence of illumination conditions in orchards

RGB cameras are known to be sensitive to changes in illumination conditions, particularly in agricultural settings such as orchards where shadows and sunlight variations are prevalent, potentially leading to exposure phenomena (Jiang et al., 2022). Many studies collected data in orchards with UAVs under sunny conditions, considering crops illuminated by sunlight to be in optimal lighting conditions (Ariza-Sentís et al., 2023). However, our results (Table 5) indicate that the completeness of the point cloud in the front view of the camera under shadow conditions is superior to that under sunlight. Colmap achieved 73.6 % and 59.9 % sparse and dense reconstruction completeness, respectively, in shadow conditions using only half the number of input images compared to sunlight conditions. Furthermore, in terms of the point cloud completeness in SLAM, ORB-SLAM3 and Droid-SLAM performed 6.8 % and 313.8 % better, respectively, in shadow conditions using only half the input images. This finding aligns with the research by Gené-Mola, Llorens (Gené-Mola et al., 2020), which demonstrated that point cloud resolution remains stable in low illumination conditions but decreases significantly in high illumination conditions.

For side view data, the effect of illumination conditions appears less significant. The results of SfM for the side view are as expected, with point cloud resolution decreasing as the number of input images decreases. However, for SLAM results, point cloud resolution increases as the number of input images decreases. This may be attributed to the camera's position, capturing more crop features than in the front view (Raman et al., 2022). Additionally, strong sunlight on the surfaces of grape leaves and vines can cause excessive reflections, generating noise in the reconstruction (Mirhaji et al., 2021). Moreover, variations in illumination conditions do not appear to affect process efficiency and UAV position accuracy.

4.2. Localization and mapping method choices in orchards applications

Various applications in orchard management involve resource efficiency (e.g., water stress estimation), geometric traits (e.g., geometric parameters estimation and phenotype), and field management activities (e.g., ripeness detection, yield estimation, and health status monitoring) when managing orchards (Zhang et al., 2021). According to the results from our vineyard dataset, UAV-based monocular localization and mapping methods show potential for practical applications primarily concerned with geometric traits, and field management activities.

4.2.1. UAV-based geometric traits and phenotyping in orchards

Currently, UAV-based SfM is a common method for measuring geometric traits and phenotyping due to the straightforward camera setups and high efficiency in data collection (Paulus, 2019). For example, Gené-Mola, Sanz-Cortiella (Gené-Mola et al., 2021) proposed SfM for size estimation of apples in apple orchards and compared different SfM methods under different shading conditions. Yuan, Hua (Yuan et al., 2023) proposed using UAV-based orthophotography with SfM to reconstruct apple orchards and estimate blossom density. Marks, Bömer (Marks et al., 2023) introduced a UAV-based point cloud dataset for sugar beet phenotype. Due to the ultra-high resolution of the pictures (11664 px × 8750 px), the processing time for point cloud reconstruction with SfM can be significant. The 3D features of grape phenotypic features can illustrate detailed information, such as trunk diameter, grape bunch size, and volume, more accurately than 2D images (Fig. 7 and Fig. 9). However, rapid phenotyping has been a major limitation in crop breeding (Su et al., 2019). Due to the low efficiency of SfM in the reconstruction (Jiang et al., 2020; Paulus, 2019), visual SLAM presents a potential alternative for phenotyping and measuring plant geometric traits. As indicated in Table 4 and Table 6, monocular SLAM can process the same data in a shorter time with higher positional accuracy. Additionally, deep learning-based SLAM significantly improves point cloud completeness compared to feature-based SLAM (Table 6), especially for side view camera position, achieving a similar order of magnitude in point cloud resolution as the SfM approach in sparse reconstruction.

However, there are still limitations to monocular localization and mapping methods. The lack of scale information can lead to errors in the scale size of the positions and mappings, and in large-scale applications, sensor and position estimation errors can accumulate, resulting in trajectory drift (Table 6 and Fig. 6). These errors can be mitigated in SLAM by incorporating multi-sensor data such as IMU and LiDAR. Recent evaluations of deep learning-based monocular SLAM have shown that deep neural networks can estimate depth information to acquire relative scale. Additionally, position estimation in SLAM can be optimized using neural network modules and bundle adjustment, combining the advantages of both SfM and traditional SLAM in terms of reconstruction efficiency, completeness, and accuracy. Therefore, deep learning has significant potential to enhance the reconstruction efficiency of SfM and

Table 12
Comparison of efficiency and accuracy requirements for five orchard management activities and SLAM implementation setting recommendations. "+" means low requirement, and "+++" means high requirement, "L" means light, "D" means dark, "F" means front view, and "S" means side view. The setup1, 2 and 3 are shown in Table 10.

Orchard Management Activities		Pruning	Fertilization	Irrigation	Weed and pest control	Harvest
Efficiency requirement		++	++	++	+++	+++
Position accuracy requirement		+++	++	++	++	+++
Completeness requirement		+++	++	+	++	+++
ORB-SLAM 3	Camera direction	S	F	F	S	S
	Illumination conditions	L	L	L	L or D	D
	# ORB feature point	++	++	++	++	++
	Scale factor	+	+	++	+	+
	# levels	+	+	++	+	+
Droid-SLAM	Camera direction	S	F	F	S	S
	Illumination conditions	D	L	L	L	D
	Parameter setting	Setup2	Setup2	Setup3	Setup2	Setup2

the reconstruction accuracy and completeness of SLAM for supporting plant phenotyping.

4.2.2. *UAV-based field management activities in orchards*

In orchard management, field management activities such as pruning, fertilization, irrigation, harvest, and weed and pest control have both ecological and economic impacts (Zhang et al., 2021). However, different field management activities may have varying requirements for the efficiency, accuracy, and completeness of the reconstruction (Table 12). For example, Teng, Zhang (Teng et al., 2023) evaluated UAV-based SfM and 3D LiDAR SLAM for peach tree pruning. They found that the SLAM approach offers greater accuracy and higher efficiency than the SfM approach in modeling point clouds, which are also in line with our findings. Although our analyses pertain solely to visual SLAM, the results regarding efficiency (Table 4) and positional accuracy (Table 6) also indicate the potential for accurately and efficiently modeling pruned trees. For fertilization and irrigation, current research primarily focuses on the long-term monitoring of soil fertility on a large scale using remote sensing (Bulanon et al., 2016). Therefore, the requirements for efficiency, accuracy and completeness are relatively low.

Similarly, for weed and pest control, most studies collect image data in orchards using UAVs to generate orthophotos that map weeds and pests on a large scale (Kaivosoja et al., 2021). These monitoring methods for fertilization, irrigation and weed and pest control reflect global trends in orchard management. However, our results suggest there is potential to enhance accuracy and efficiency at a local level. Conversely, many studies have focused on fruit detection for harvest and yield estimation (Xiong et al., 2023), which aligns with our findings regarding efficiency, accuracy, and completeness. Overall, pruning and harvesting have similar requirements for precise identification of branches and plant structures since the fruit distribution is determined by the pruning methods, necessitating high position accuracy. Moderate efficiency and completeness are important for timely and thorough data collection during pruning and harvesting. Fertilization benefits from accurate placement but does not require extremely high accuracy. Moderate efficiency and completeness help ensure that the right areas are fertilized properly. Precise mapping of soil moisture and plant needs is essential, but the positional accuracy requirements are moderate. Completeness is less critical in irrigation than in other activities. Timely identification and treatment are crucial in weed and pest control, so high efficiency is needed. Accurate location of weeds and pests requires moderate accuracy and completeness.

Finally, after conducting a comprehensive evaluation of parameter sensitivity and illumination conditions in SLAM, we provide SLAM parameter and illumination condition recommendations based on various field management activities (Table 10, Table 12).

5. Conclusions

This study provides a comparative analysis of three leading visual localization and mapping methods—Colmap (SfM), ORB-SLAM3, and Droid-SLAM—applied to vineyard videos captured by UAV. Each method has distinct strengths and limitations: Colmap delivers high positional accuracy and reconstruction completeness but is computationally expensive and time-consuming; ORB-SLAM3 offers efficient processing and low hardware requirements, though its lower point cloud resolution limits its ability to capture fine crop features; Droid-SLAM, leveraging deep learning, excels in point cloud resolution and robustness but also demands substantial computational resources.

Two key insights emerge from this study: (1) the balanced performance of Droid-SLAM highlights its potential for real-time crop phenotyping, particularly in monocular-based applications; and (2) variations in illumination primarily affect point cloud completeness without significantly impacting positional accuracy or processing efficiency. In future, integrating deep learning into SLAM systems offers promising avenues for enhancing point cloud resolution and improving reconstruction efficiency, making these methods more suitable for remote sensing and agricultural applications. Further refinement of these techniques, especially in relation to specific orchard requirements, can facilitate more accurate and efficient phenotyping tasks.

Ethical approval

This article does not contain any studies with human participants or animals performed by any of the authors.

CRediT authorship contribution statement

Kaiwen Wang: Writing – review & editing, Writing – original draft, Visualization, Project administration, Methodology, Formal analysis, Conceptualization. **Lammert Kooistra:** Writing – review & editing, Supervision, Project administration, Methodology, Conceptualization. **Yaowu Wang:** Writing – review & editing, Visualization, Methodology, Formal analysis. **Sergio Vélez:** Writing – review & editing, Visualization, Methodology, Investigation, Data curation. **Wensheng Wang:** Writing – review & editing, Supervision, Project administration, Funding acquisition, Conceptualization. **João Valente:** Writing – review & editing, Supervision, Project administration, Methodology, Conceptualization.

Declaration of competing interest

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

Acknowledgements

The authors acknowledge the support of the Wageningen University, Chinese Academy of Agricultural Sciences joint Ph.D. program.

Appendix

Table 13
The time cost in Colmap, Droid-SLAM, and ORB-SLAM3 in three experiments under same configurations. "AT" means average time, "#" indicates how many times it has been tested.

Methods			Front_View_Light	Front_View_Dark	Side_Short_Light	Side_Short_Dark	Side_Loop	Side_Long
Colmap	Sparse Reconstruction	#1	35 mins 16 s	7 mins 33 s	15 mins 39 s	12 mins 51 s	168 mins 10 s	477 mins 19 s
		#2	37 mins 10 s	7 mins 15 s	15 mins 9 s	12 mins 56 s	164 mins 33 s	455 mins 56 s
		#3	34 mins 26 s	7 mins 8 s	14 mins 22 s	13 mins 53 s	160 mins 9 s	473 mins 59 s
		AT	35 mins 38 s	7 mins 19 s	15 mins 4 s	13 mins 14 s	164 mins 18 s	469 mins 5 s
	Dense Reconstruction	#1	169 mins 48 s	76 mins 42 s	118 mins 17 s	115 mins 10 s		1278 mins 42 s
		#2	174 mins 13 s	73 mins 45 s	113 mins 59 s	118 mins 17 s		1279 mins 39 s
		#3	171 mins 33 s	72 mins 20 s	115 mins 1 s	124 mins 26 s		1266 mins 44 s
		AT	171 mins 52 s	74 mins 16 s	115 mins 46 s	119 mins 18 s		1275 mins 2 s
ORB-SLAM3	#1		5 mins 3 s	Track Fail	4 mins 54 s	4 mins 52 s	22 mins 5 s	7 mins 41 s
	#2		4 mins 51 s	2 mins 15 s	4 mins 35 s	5 mins 8 s	22 mins 32 s	7 mins 49 s
	#3		4 mins 44 s	Track Fail	4 mins 44 s	4 mins 56 s	21 mins 34 s	7 mins 43 s
	AT		4 mins 53 s	2 mins 15 s	4 mins 47 s	4 mins 59 s	22 mins 4 s	7 mins 45 s
Droid-SLAM	#1		5 mins 31 s	3 mins 1 s	4 mins 39 s	5 mins 5 s	79 mins 27 s	28 mins 5 s
	#2		5 mins 30 s	2 mins 58 s	4 mins 32 s	5 mins 18 s	82 mins 31 s	28 mins 39 s
	#3		5 mins 54 s	3 mins 5 s	4 mins 33 s	4 mins 51 s	77 mins 48 s	29 mins 50 s
	AT		5 mins 39 s	3 mins 2 s	4 mins 35 s	5 mins 5 s	79 mins 56 s	28 mins 52 s

Table 14
Parameters setup in Colmap in the five steps in SfM.

Feature Extraction	
Parameter Name	**Value**
Camera model	Simple_pinhole
Image_size	3200
Num_features	8192
Feature Matching	
Matching method	Sequential
Overlap	10
Max_ratio	0.8
Max_distance	0.7
Max_num_matches	32,768
Use_gpu	true
Max_error	4.00
Confidence	0.99900
Max_num_trials	100,000
Min_inlier_ratio	0.250
Min_num_inliers	15
Stereo reconstruction	
Min_num_pixels	15
Max_num_pixels	10,000
Max_traversal_depth	100
Max_reproj_error	2
Max_depth_error	0.01
Max_normal_error	10
Check_num_images	50

Data availability

Data will be made available on request.

References

Abdelrasoul Y, Saman ABSH, Sebastian P, editors. A quantitative study of tuning ROS gmapping parameters and their effect on performing indoor 2D SLAM. 2016 2nd IEEE international symposium on robotics and manufacturing automation (ROMA); 2016: IEEE.

Ariza-Sentís, M., Vélez, S., Valente, J., 2023. Dataset on UAV RGB videos acquired over a vineyard including bunch labels for object detection and tracking. Data Brief 46, 108848.

Ariza-Sentís, M., Vélez, S., Valenti, R.G., Valente, J., 2023. Improving up-close remote sensing of occluded areas in vineyards through customized multiple-Unmanned-Aerial-Vehicle path planning. Environ. Sci. Proc. 29 (1), 57.

Ariza-Sentís, M., Vélez, S., Martínez-Peña, R., Baja, H., Valente, J., 2024. Object detection and tracking in Precision Farming: A systematic review. Comput. Electron. Agric. 219, 108757.

Ariza-Sentís, M., Wang, K., Cao, Z., Vélez, S., Valente, J., 2024. GrapeMOTS: UAV vineyard dataset with MOTS grape bunch annotations recorded from multiple perspectives for enhanced object detection and tracking. Data Brief 54, 110432.

Bulanon, D.M., Lonai, J., Skovgard, H., Fallahi, E., 2016. Evaluation of different irrigation methods for an apple orchard using an aerial imaging system. ISPRS Int. Geo-Inf. 5 (6), 79.

Burri, M., Nikolic, J., Gohl, P., Schneider, T., Rehder, J., Omari, S., et al., 2016. The EuRoC micro aerial vehicle datasets. Int. J. Robot. Res. 35 (10), 1157–1163.

Campos, C., Elvira, R., Rodríguez, J.J.G., Montiel, J.M., Tardós, J.D., 2021. Orb-slam3: An accurate open-source library for visual, visual–inertial, and multimap slam. IEEE Trans. Rob. 37 (6), 1874–1890.
Chen, L.-B., Huang, X.-R., Chen, W.-H., 2023. Design and implementation of an artificial intelligence of things-based autonomous mobile robot system for pitaya harvesting. IEEE Sens. J.
Costley A, Christensen R, editors. Landmark aided gps-denied navigation for orchards and vineyards. 2020 IEEE/ION Position, Location and Navigation Symposium (PLANS); 2020: IEEE.
DeTone D, Malisiewicz T, Rabinovich A, editors. Superpoint: Self-supervised interest point detection and description. Proceedings of the IEEE conference on computer vision and pattern recognition workshops; 2018.
Engel, J., Sturm, J., Cremers, D., 2014. Scale-aware navigation of a low-cost quadrocopter with a monocular camera. Robot. Auton. Syst. 62 (11), 1646–1656.
Gené-Mola, J., Llorens, J., Rosell-Polo, J.R., Gregorio, E., Arnó, J., Solanelles, F., et al., 2020. Assessing the performance of rgb-d sensors for 3d fruit crop canopy characterization under different operating and lighting conditions. Sensors 20 (24), 7072.
Gené-Mola, J., Sanz-Cortiella, R., Rosell-Polo, J.R., Escola, A., Gregorio, E., 2021. In-field apple size estimation using photogrammetry-derived 3D point clouds: Comparison of 4 different methods considering fruit occlusions. Comput. Electron. Agric. 188, 106343.
Hroob I, Polvara R, Molina S, Cielniak G, Hanheide M, editors. Benchmark of visual and 3D lidar SLAM systems in simulation environment for vineyards. Towards Autonomous Robotic Systems: 22nd Annual Conference, TAROS 2021, Lincoln, UK, September 8–10, 2021, Proceedings 22; 2021: Springer.
Islam, R., Habibullah, H., Hossain, T., 2023. AGRI-SLAM: A real-time stereo visual SLAM for agricultural environment. Auton. Robot. 47 (6), 649–668.
Jiang, S., Jiang, C., Jiang, W., 2020. Efficient structure from motion for large-scale UAV images: A review and a comparison of SfM tools. ISPRS J. Photogramm. Remote Sens. 167, 230–251.
Jiang, A., Noguchi, R., Ahamed, T., 2022. Tree trunk recognition in orchard autonomous operations under different light conditions using a thermal camera and faster R-CNN. Sensors 22 (5), 2065.
Kaivosoja, J., Hautsalo, J., Heikkinen, J., Hiltunen, L., Ruuttunen, P., Näsi, R., et al., 2021. Reference measurements in developing UAV systems for detecting pests, weeds, and diseases. Remote Sens. 13 (7), 1238.
Keller, M., 2020. The science of grapevines. Academic press.
Lawrence, J., Bernal, J., Witzgall, C., 2019. A purely algebraic justification of the Kabsch-Umeyama algorithm. J. Res. National Inst. Standards Technol. 124, 1.
Li, J., Gao, W., Wu, Y., Liu, Y., Shen, Y., 2022. High-quality indoor scene 3D reconstruction with RGB-D cameras: A brief review. Comput. Visual Media 8 (3), 369–393.
Marks E, Bömer J, Magistri F, Sah A, Behley J, Stachniss C. BonnBeetClouds3D: A Dataset Towards Point Cloud-based Organ-level Phenotyping of Sugar Beet Plants under Field Conditions. arXiv preprint arXiv:231214706. 2023.
Mirhaji, H., Soleymani, M., Asakereh, A., Mehdizadeh, S.A., 2021. Fruit detection and load estimation of an orange orchard using the YOLO models through simple approaches in different imaging and illumination conditions. Comput. Electron. Agric. 191, 106533.
Moreno, H., Andújar, D., 2023. Proximal sensing for geometric characterization of vines: A review of the latest advances. Comput. Electron. Agric. 210, 107901.
Moreno, H., Valero, C., Bengochea-Guevara, J.M., Ribeiro, Á., Garrido-Izard, M., Andújar, D., 2020. On-ground vineyard reconstruction using a LiDAR-based automated system. Sensors 20 (4), 1102.
Norzam W, Hawari H, Kamarudin K, editors. Analysis of mobile robot indoor mapping using GMapping based SLAM with different parameter. IOP Conference Series: Materials Science and Engineering; 2019: IOP Publishing.
Özyeşil, O., Voroninski, V., Basri, R., Singer, A., 2017. A survey of structure from motion*. Acta Numer. 26, 305–364.
Paulus, S., 2019. Measuring crops in 3D: Using geometry for plant phenotyping. Plant Methods 15 (1), 103.
Raman, M.G., Carlos, E.F., Sankaran, S., 2022. Optimization and evaluation of sensor angles for precise assessment of architectural traits in peach trees. Sensors 22 (12), 4619.
Santesteban, L.G., 2019. Precision viticulture and advanced analytics. A short review. Food Chem. 279, 58–62.
Schonberger JL, Frahm J-M, editors. Structure-from-motion revisited. Proceedings of the IEEE conference on computer vision and pattern recognition; 2016.
Schops T, Schonberger JL, Galliani S, Sattler T, Schindler K, Pollefeys M, et al., editors. A multi-view stereo benchmark with high-resolution images and multi-camera videos. Proceedings of the IEEE conference on computer vision and pattern recognition; 2017.
Schubert D, Goll T, Demmel N, Usenko V, Stückler J, Cremers D, editors. The TUM VI benchmark for evaluating visual-inertial odometry. 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS); 2018: IEEE.
Sossalla P, Hofer J, Rischke J, Busch J, Nguyen GT, Reisslein M, et al., editors. Optimizing Edge SLAM: Judicious parameter settings and parallelized map updates. GLOBECOM 2022-2022 IEEE Global Communications Conference; 2022: IEEE.
Su, W., Zhang, M., Bian, D., Liu, Z., Huang, J., Wang, W., et al., 2019. Phenotyping of corn plants using Unmanned Aerial Vehicle (UAV) images. Remote Sens. 11, 2021.
Tardaguila, J., Stoll, M., Gutiérrez, S., Proffitt, T., Diago, M.P., 2021. Smart applications and digital technologies in viticulture: A review. Smart Agric. Technol. 1, 100005.
Teed, Z., Deng, J., 2021. Droid-slam: Deep visual slam for monocular, stereo, and rgb-d cameras. Adv. Neural Inf. Proces. Syst. 34, 16558–16569.
Teng, P., Zhang, Y., Yamane, T., Kogoshi, M., Yoshida, T., Ota, T., et al., 2023. Accuracy evaluation and branch detection method of 3D modeling using backpack 3D Lidar SLAM and UAV-SfM for peach trees during the pruning period in winter. Remote Sens. 15 (2), 408.
Vélez, S., Ariza-Sentís, M., Valente, J., 2024. EscaYard: Precision viticulture multimodal dataset of vineyards affected by Esca disease consisting of geotagged smartphone images, phytosanitary status, UAV 3D point clouds and Orthomosaics. Data Brief 54, 110497.
Wallace, L., Lucieer, A., Malenovský, Z., Turner, D., Vopěnka, P., 2016. Assessment of forest structure using two UAV techniques: A comparison of airborne laser scanning and structure from motion (SfM) point clouds. Forests 7 (3), 62.
Wang W, Zhu D, Wang X, Hu Y, Qiu Y, Wang C, et al., editors. Tartanair: A dataset to push the limits of visual slam. In 2020 IEEE. RSJ International Conference on Intelligent Robots and Systems (IROS).
Wang K, Kooistra L, Pan R, Wang W, Valente J. UAV-based simultaneous localization and mapping in outdoor environments: A systematic scoping review. Journal of Field Robotics.n/a(n/a).
Wang, K., Kooistra, L., Pan, R., Wang, W., Valente, J., 2024. UAV-based simultaneous localization and mapping in outdoor environments: A systematic scoping review. J. Field Rob.
Xiong, J., Liang, J., Zhuang, Y., Hong, D., Zheng, Z., Liao, S., et al., 2023. Real-time localization and 3D semantic map reconstruction for unstructured citrus orchards. Comput. Electron. Agric. 213, 108217.
Xu, B., Zhang, L., Liu, Y., Ai, H., Wang, B., Sun, Y., et al., 2021. Robust hierarchical structure from motion for large-scale unstructured image sets. ISPRS J. Photogramm. Remote Sens. 181, 367–384.
Yang N, Stumberg Lv, Wang R, Cremers D, editors. D3vo: Deep depth, deep pose and deep uncertainty for monocular visual odometry. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition; 2020.
Yuan, W., Hua, W., Heinemann, P.H., He, L., 2023. UAV photogrammetry-based apple orchard blossom density estimation and mapping. Horticulturae 9 (2), 266.
Zhang, C., Valente, J., Kooistra, L., Guo, L., Wang, W., 2021. Orchard management with small unmanned aerial vehicles: A survey of sensing and analysis approaches. Precis. Agric. 22 (6), 2007–2052.