

UAV-based monocular 3D panoptic mapping for fruit shape completion in orchard

Kaiwen Wang^{a,d}^{*,1}, Yue Pan^b¹, Federico Magistri^b, Lammert Kooistra^c, Cyrill Stachniss^b, Wensheng Wang^d, João Valente^e

^a Wageningen University & Research, Information Technology Group, 6708 PB, Wageningen, The Netherlands

^b University of Bonn, Center for Robotics, 53115, Bonn, Germany

^c Wageningen University & Research, Laboratory of Geo-Information Science and Remote Sensing, 6708 PB, Wageningen, The Netherlands

^d Chinese Academy of Agriculture Sciences, Agricultural Information Institute, 10086, Beijing, China

^e Spanish National Research Council (CSIC), Center for Automation and Robotics (CAR), 28500, Madrid, Spain

ARTICLE INFO

Dataset link: [10.5281/zenodo.15635994](https://doi.org/10.5281/zenodo.15635994)

Keywords:

3D shape completion
Unmanned aerial vehicle (UAV)
3D reconstruction
Structure from motion (sfM)
Precision agriculture

ABSTRACT

Accurate fruit shape reconstruction under real-world field conditions is essential for high-throughput phenotyping, sensor-based yield estimation, and orchard management. Existing approaches based on 2D imaging or explicit 3D reconstruction often suffer from occlusions, sparse views, and complex scene dynamics as a result of the plant geometries. This paper presents a novel UAV-based monocular 3D panoptic mapping framework for robust and scalable fruit shape completion in orchards. The proposed method integrates (1) Grounded-SAM2 for multi-object tracking and segmentation (MOTS), (2) photogrammetric structure-from-motion for 3D scene reconstruction, and (3) DeepSDF, an implicit neural representation, for completing occluded fruit geometries with a neural network. We furthermore propose a new MOTS evaluation protocol to assess tracking performance without requiring ground truth annotations. Experiments conducted in both controlled laboratory conditions and an operational apple orchard demonstrate the accuracy of our 3D fruit reconstruction at the centimeter level. The Chamfer distance error of the proposed shape completion method using the DeepSDF shape prior reduces this to the millimeter level, and outperforms the traditional method, while Grounded-SAM2 enables robust fruit tracking across challenging viewpoints. The approach is highly scalable and applicable to real-world agricultural scenarios, offering a promising solution to reconstruct complete fruits with visibility higher than 10% for precise 3D fruit phenotyping at a large scale under occluded conditions.

1. Introduction

Fruit production is a relevant component of global agriculture, contributing to food supply, dietary diversity, and economic development (Mason-D'Croz et al., 2019). To support sustainable fruit cultivation and meet the increasing demand for high-quality produce, accurate and efficient phenotyping plays a vital role in assessing plant traits under field conditions (Tripodi et al., 2018). Among various phenotypic traits, fruit shape and size are particularly important, as they not only directly relate to market value and consumer preference but also serve as indicators of plant health and genetic performance (Onyekwelu et al., 2014). However, traditional phenotyping methods are often manual, labor-intensive, and difficult to scale (Reynolds et al., 2019). Developments in precision agriculture, including advances in robotics,

remote sensing, computer vision, and data-driven analysis have improved the efficiency and accuracy, thereby facilitating large-scale and high-throughput fruit phenotyping in orchards.

Terrestrial laser scanning (TLS), as one of the typical remote sensing tools, has been widely applied in orchard phenotyping due to its ability to capture high-resolution and geometrically accurate 3D data of fruit trees (Medic et al., 2023). While effective in controlled environments, TLS systems are often expensive, cumbersome to deploy across large areas, and sensitive to occlusions, limiting their practicality for extensive in-field phenotyping (Li et al., 2021).

Unmanned aerial vehicles (UAVs), equipped with RGB cameras, have emerged as a flexible and scalable alternative for remote sensing in precision agriculture (Maes and Steppe, 2019). Through structure-from-motion (SfM) techniques, UAV-based imagery can be used to reconstruct 3D point clouds that represent the spatial structure of

* Corresponding author at: Wageningen University & Research, Information Technology Group, 6708 PB, Wageningen, The Netherlands.

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

¹ Equal contribution

orchard environments for phenotyping (Huang et al., 2020). These reconstructions are more affordable (less than 1000€) and accessible than LiDAR (more than 10,000€). However, point clouds from UAV-based SfM also present limitations. Compared to LiDAR, which can receive second or third echoes to reconstruct the invisible areas, SfM relies on visible surface features, which restricts the reconstruction of shadow, occluded, or partially visible regions (Lingua et al., 2017). There exist methods including image strategies in image acquisition (e.g., choosing optimal lighting conditions, using long focal lengths, or adjusting flight paths), correction processes (e.g., using image centers, or advanced algorithms), and sensor selection (e.g., push-broom sensors to minimize directional occlusion) that can be used to reduce the impact of shadows and occlusions (Zhang et al., 2023). The unstructured, clustered, and dynamic agricultural environments, however, make these methods still challenging. In addition to the challenge of occlusions and shadows, the scale of orchard phenotyping is also a challenge for both LiDAR and conventional UAV remote sensing for SfM. Using the point clouds from LiDAR and UAV-based SfM for orchard phenotyping are commonly focused on tree structural traits such as canopy volume, tree height, and canopy diameter, which is limited to the tree level (Zhang et al., 2024; López-Granados et al., 2019). Up-close sensing with UAVs presents a promising solution to reduce the impact of occlusions and address the issue of scalability. Unlike conventional UAV-based remote sensing, which typically operates at altitudes above 10 m along predefined waypoints, up-close sensing allows UAVs to fly between crop rows at altitudes below 5 m, enabling the capture of more detailed crop information at fruit level and further minimizing occlusion and shadow effects by canopies (Wang et al., 2024).

Multi-view photogrammetry is important for orchard phenotyping through up-close sensing. To overcome the limitations of single-view or sparse-view reconstruction, multi-view image acquisition has been widely adopted to improve object completeness in 3D reconstructions (Zhu et al., 2015). By capturing the orchard scene from multiple viewpoints, it becomes possible to mitigate occlusions and recover a more comprehensive representation of the environment (Yu et al., 2024). However, due to the natural growth patterns of fruit trees, a significant number of fruits are often occluded by surrounding structures such as branches, leaves, or neighboring fruits (Amatya et al., 2016; Gené-Mola et al., 2023). These occlusions pose fundamental challenges in recovering the full geometry of individual fruits using conventional geometric methods, which rely heavily on visible surface features. (Marangoz et al., 2022) developed a fruit mapping method using superellipsoids for fruit completion to monitor fruit in greenhouses, but the shape completion with superellipsoids can have limitations in accuracy. Recent implicit 3D representations such as signed distance function (SDF) and neural radiance field (NeRF) offer promising alternatives to address occlusions and fruit shape completion (Mildenhall et al., 2021). Compared to explicit representations (e.g., point clouds, voxels), implicit representations can model continuous 3D shapes from sparse or partial observations by learning underlying geometric priors from training data (Park et al., 2019). The accurate fruit shape reconstructed by the model can improve the accuracy of grasping during harvesting (Magistri et al., 2024) and pruning the leaves to get better fruit shape completion (Yao et al., 2025). In the context of orchard environments, it becomes feasible to reconstruct the complete shape of partially occluded fruits by leveraging prior knowledge learned from complete 3D samples.

In this study, we propose solutions to three issues: (1) simple and efficient UAV-based up-closing sensing for orchard monitoring; (2) accurate fruit tracking and segmentation from complex orchard scenes for data association between 2D and 3D across frames; (3) 3D reconstruction of complete fruit geometries in real-world orchard environments from incomplete observations, aiming to assist fruit phenotyping and support data-driven orchard management. We introduce a multi-stage pipeline that integrates instance-level segmentation, multi-view data

association between 2D images and 3D point clouds, and implicit shape reconstruction for fruit completion in orchards. This novel approach enables fine-grained, instance-level fruit reconstitution in real-world orchard conditions, where occlusions and viewpoint limitations pose significant challenges. The main contributions of this work are (1) to present a unified fruit reconstruction pipeline that combines MOTs, DeepSDF, and photogrammetry methods by UAV-based up-close sensing in orchard environments, (2) to propose a novel method to evaluate MOTs without any annotations, and (3) to provide a highly accurate 3D apple dataset collected in a laboratory environment, along with UAV-captured high-resolution videos in the field. The dataset and codes for this research are publicly available at: <https://github.com/Kaiwen-Robotics/Mono3DOrchard>.

2. Study area and materials

This study contains two data collection areas: field data collection and laboratory data collection.

2.1. Field data collection

2.1.1. Study area

The field data collection was conducted within an apple orchard located in Randwijk, Overbetuwe, the Netherlands (51.9376, 5.703057 in WGS84 UTM 31U), as shown in Fig. 1. The selected areas of 0.083 ha in the orchard contain four rows of the apple variety Elstar, *Malus pumila* 'Elstar', with tree and row spacing of 1.1 m and 3.0 m, respectively. There were about 80 trees in each row in the targeted study area. The maturity of the apples was in the expanding growth stage during the data collection, these apples were harvested one month later.

2.1.2. UAV data collection and flight mode design

We used a commercial UAV equipped with a high-resolution RGB sensor for video data collection in the apple orchard, see Table 1. To ensure high-quality 3D reconstruction for fruit completion, four different UAV flight modes were carefully designed during data collection, see Fig. 2. Flight mode A followed a straight line from one edge of the row to another edge of the row, and the camera perspective was from the side to the trees, as shown in Fig. 2(a). Flight mode B followed an up-down trajectory with a closer distance to the trees compared to flight A, see Fig. 2(b). Flight modes C and D were flights between rows of fruit trees at a higher altitude for safety, but the camera perspectives were different, with C being a side view and D a front view, see Figs. 2(c) and 2(d). The four flight modes were designed to evaluate the influences of MOTs and reconstruction performance to vary altitude, camera perspective, and distance to the crops during flight between rows. The UAV collected video data at 29.98 FPS between tree rows at an altitude of around 2 to 4 m. For the approach, however, we used only 3 FPS to improve efficiency while keeping it robust.

2.1.3. Ground truth manual measurements

For ground truth manual measurements, we measured the heights of the wood poles in the orchard and the apple sizes as shown in Fig. 3(b). In a row of fruit trees, there are 14 wooden poles spaced apart, and between every two poles, there are around six to eight fruit trees arranged. These wood poles are set for supporting the growth of crops and organizing orchard management. The height of each wood pole is 2.7 m with our measurement. The height of the poles was the reference for metric scale recovery.

Apple size, the maximum diameter of the apple is one of the indicators for fruit grading, growth condition monitoring, and precise yield estimation at the fruit level. In row 1, we sampled 24 trees, and about six apples per tree, which leads to a total of 135 apples. We also recorded the maximum diameters of the sampled apples by manually measuring the diameters with dial calipers. To correspond the sampled apples to the video data, colored ribbons were tied on the apple stalks as colored labels depicted in Fig. 3(b).

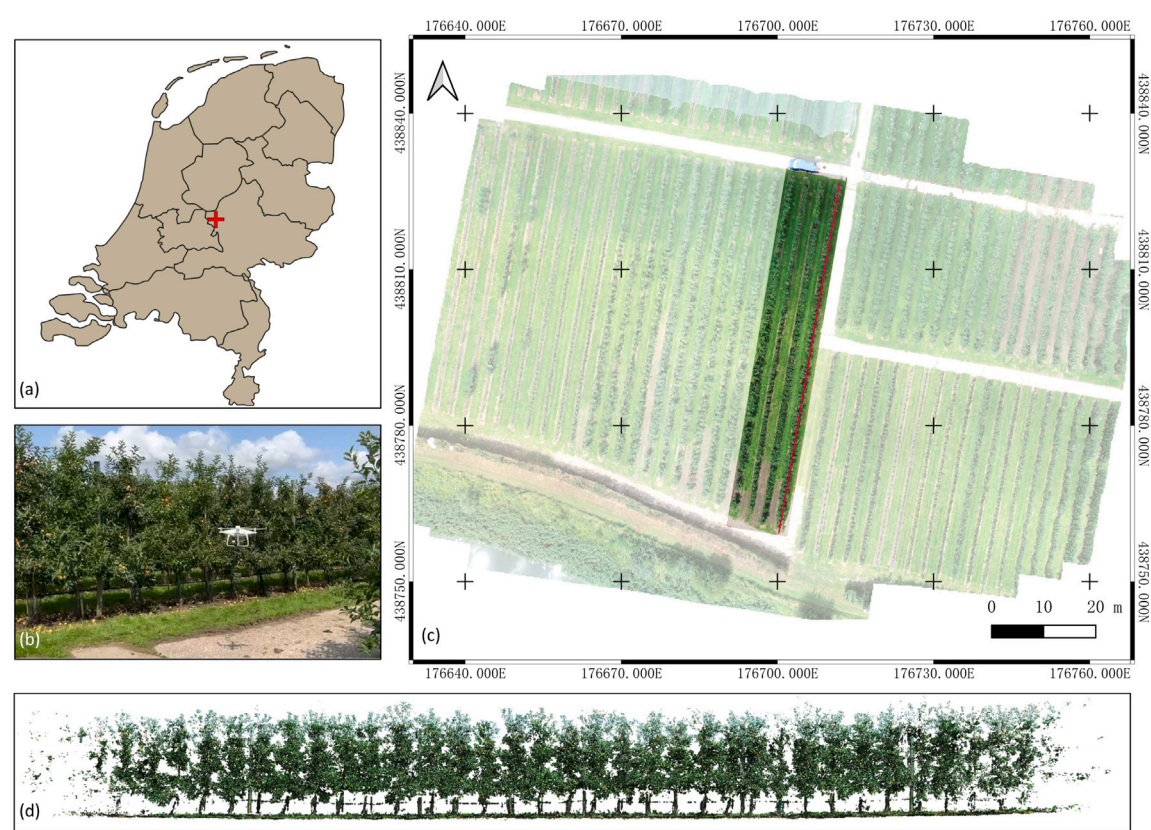


Fig. 1. The location of the apple orchard in Randwijk, Overbetuwe, Gelderland in the Netherlands. The red cross in (a) presents the location of the orchard. (b) shows the UAV flight conditions in the orchard during data collection. And the highlighted rectangular region in (c) marks the study area in the whole orchard orthomosaic, and the red line shows one of the targeted apple tree rows (row 1). (d) presents the overview of the panoptic point cloud of the targeted row in the orchards. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 1
Description of flight parameters and operation conditions during UAV data collection.

UAV platform	DJI Phantom 4 RTK, Shenzhen, China
Sensor	RGB FC6310R
Sensor Type	CMOS
Resolution	3840 × 2160
Focal length (mm)	8.8
Flight altitude (m)	Around 2 to 4
Flight velocity (m/s)	Around 0.1 to 1.5
Video Frame rate (fps)	29.98
Collection date & start time	July 24th, 2024, from 10:01 AM to 10:41 AM (before apple harvesting)
Wind Speed (m/s)	2.8
Illumination conditions	Sunny
Temperature (°C)	20

2.2. Lab setup and data collection

In the laboratory environment, the shapes of 100 individual apples of the same variety ‘Elstar’ as present in the orchard were scanned and collected with a 3D scanner as shown in Fig. 3(a). The type of 3D scanner was CR-Scan Ferret from Creality, China,² with 0.1 mm accuracy. The apples were fixed at the top of the stick. Two checkerboard grids were placed vertically on either side of the back of the apple for reference and camera calibration.

2.3. 3D fruit data preprocessing

For the lab dataset, all 3D apple scanning setups in the lab environment allow us to normalize 100 different apple data, which means that

the z-axis positive direction aligns the apple stem direction, and the center points of all apples were placed at the origin point, see Fig. 4. The meshes of 100 apples and sticks were segmented separately with CloudCompare. After that, the sticks can be considered as cylinders. We calculated the centers of mass of the cylinders and retained three principal components of the cylinders using PCA (principal component analysis), with the first direction of the principal component being the direction of the cylinder’s generatrix. Then, we aligned all the sticks with individual transformation matrices for each. In the meantime, each transformation matrix was applied to each corresponding apple. Finally, all apples were placed at the origin point in a common canonical coordinate system.

2.4. Apple diameter distribution in lab and field

After getting the aligned 3D apple dataset, we also calculated the maximum diameters of the 3D apples. Since the maturity of the apples

² <https://www.creality.com/products/cr-scan-ferret-3d-scanner>

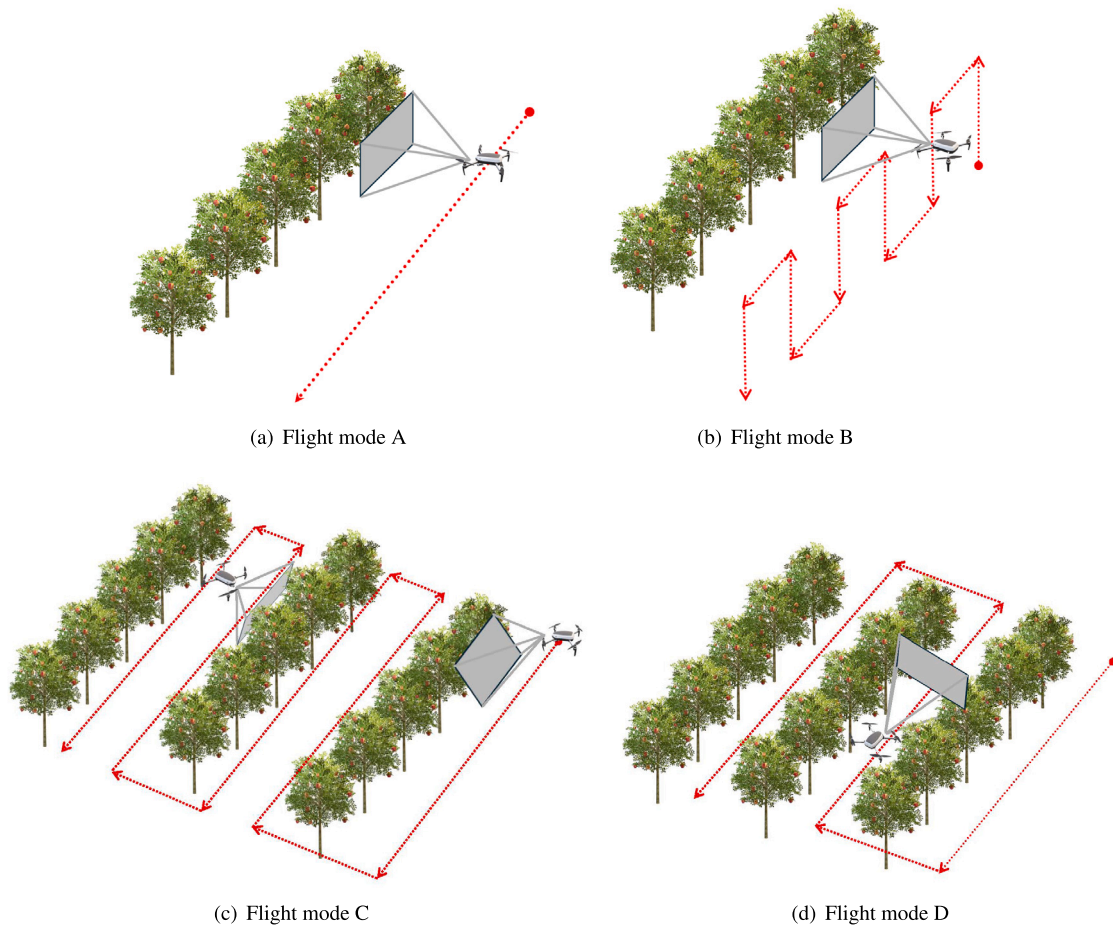


Fig. 2. UAV flight modes in apple orchard. Flight mode A shows the UAV flying at the edge of the orchard in a straight line, and the direction of the camera was a side perspective during video capture. Flight mode B shows the UAV following an up-down path at a closer distance than mode A, and the direction of the camera is a side perspective during video capture. Flight mode C shows the UAV flying between tree rows in a straight line with a downward side camera perspective view of 30 to 45 degrees at a higher altitude than mode A during video capture. Flight mode D shows the UAV flying between tree rows in a straight line with a downward front camera perspective view of 30 to 45 degrees at a higher altitude than mode A during video capture.

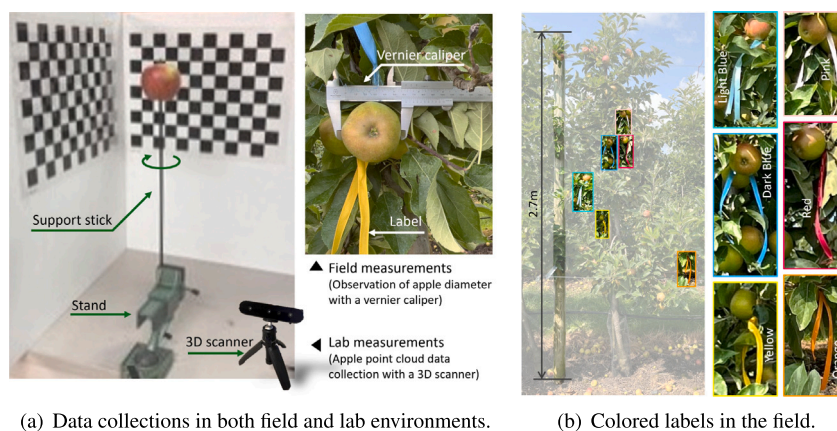


Fig. 3. Data collection in the apple orchard and a controlled laboratory environment. (a) shows the laboratory setup for data collection with a 3D scanner and field data measurement with a vernier caliper. (b) shows the six colored labels, including red, yellow, pink, orange, dark blue, and light blue in the orchards, and a 2.7-meter height of the wooden pole in the orchard, which supports the growth of the apple trees and organizes the orchards. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

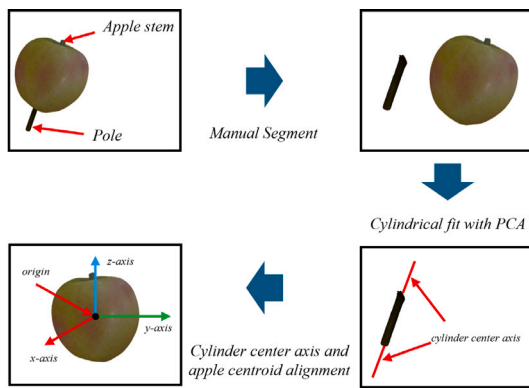


Fig. 4. Normalization procedure of the total 100 apple colored point clouds.

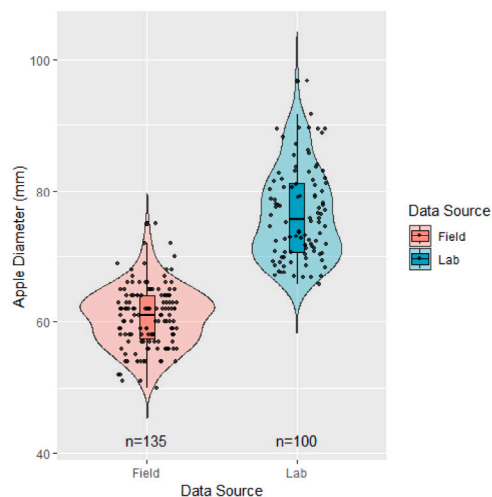


Fig. 5. Distribution of apple diameters in both lab and field environments.

in the field was in the expansion stage and the maturity of the apples in the lab was ripe, the distribution of apple diameters demonstrates a significant difference between apples in the laboratory and in the field. As shown in Fig. 5, the laboratory apples exhibited notably larger diameters with measurements predominantly ranging from 70 mm to 85 mm, while field apples displayed smaller diameters primarily between 55 mm and 65 mm.

3. Methods

The proposed UAV-based monocular 3D panoptic mapping approach for orchard environments is shown in Fig. 6. The approach integrates multi object tracking and segmentation (MOTS) with Grounded-SAM2 for apple tracking (Ravi et al., 2024; Ren et al., 2024), apple shape completion via DeepSDF for foreground reconstruction (Park et al., 2019), and background reconstruction of the orchard via structure-from-motion (SfM) (Schonberger and Frahm, 2016). The detailed methods are explained in the following sections.

3.1. Apple MOTS with grounded-SAM2

Data association across frames is important for linking 2D apple images and 3D point clouds. MOTS could be a solution to associate data since it can track objects. To achieve robust and consistent instance tracking of apples across multiple frames, we employ Grounded-SAM2 (Ren et al., 2024) for MOTS. The objective is to associate individual apples across 2D consecutive UAV-captured frames while

maintaining high segmentation accuracy. After that, the point cloud of the submaps will be unprojected by the 2D instance IDs and masks of each apple to get 3D individual partial point cloud of each apple with our panoptic mapping method. Unlike traditional bounding box-based tracking methods, MOTS enables pixel-wise instance tracking, which is crucial for accurate 3D fruit shape completion.

Leveraging the robust tracking capabilities of Segment Anything Model2 (SAM2), Grounded-SAM2 integrates it with open-set object detector Grounding DINO to enhance object tracking and segmentation. We initialize the tracking pipeline using the Grounding-DINO detector with a simple text prompt “apple” to generate class-aware object proposals. The predicted bounding boxes of apples from Grounding-DINO serve as box prompts for the SAM2 video predictor, enabling the extraction of high-quality apple instance masks with temporally consistent instance IDs, denoted as \mathcal{M} .

The raw UAV frames with 3 FPS were input to Grounded-SAM2. Then, the masks \mathcal{M} associated with the apple tracking IDs were predicted by a large SAM2 model and a tiny Grounding DINO model. The experiments for MOTS were implemented on an NVIDIA Quadro RTX A6000 GPU.

3.2. 3D reconstruction with SfM

We used Agisoft Metashape Pro³ (Agisoft LLC, St. Petersburg, Russia) to estimate the UAV camera poses and accurate dense depth information. We imported the raw UAV frames with 3 FPS into Agisoft and set a pair of marker points at the top and bottom of the poles, as illustrated in Fig. 3(a) as a scale reference. The distance between two points were set to 2.7 m. Then all frames were aligned with ‘highest accuracy’ and ‘source’ options under 40,000 key point limit and 4000 tie point limit. The reprojection error of the final result is 0.691 pixels. After that, point clouds with an ‘ultra high’ quality and ‘mild’ depth filtering were built. A depth export script⁴ was implemented to get the highly accurate depth maps D from SfM. Finally, the depth D and camera poses P with metric scale can be derived.

3.3. Apple shape completion with pretrained DeepSDF

The DeepSDF model (Park et al., 2019) takes a query position $\mathbf{x} \in \mathbb{R}^3$ and a latent shape code $\mathbf{z} \in \mathbb{R}^C$ as input, and predicts the corresponding SDF value $v \in \mathbb{R}$ at the query position \mathbf{x} with a decoder based on multi-layer perceptron (MLP), D_θ as $v = D_\theta(\mathbf{x}, \mathbf{z})$. We can compute a dense SDF volume by querying it at a regular 3D grid of points using a pre-trained DeepSDF model, which we use for a complete mesh reconstruction through marching cubes (Lorensen and Cline, 1998).

We followed the descriptions for DeepSDF, as described in Park et al. (2019), to prepare the dataset before training the model. We randomly sampled 200,000 points that contained positive values (outside the 3D apples) and negative (inside the 3D apples) values around all 3D apple point clouds. Then, we built eight-layer MLPs for training the DeepSDF model, with each layer comprising of 512 dimensions. The inputs of the DeepSDF model were the sampled points, which were uniformly in a sphere surrounding the 3D apples, including the sampled point positions (x, y, z) and the SDF values. The dataset was divided into a training set (88 apples) and a test set (12 apples). We calculate the SDF value of each sample point for supervision. We adopted a latent shape code size of $C = 32$ and used 3000 epochs and a 0.001 initial learning rate with a decrease of 0.0005 every 300 epochs for training. The DeepSDF model was trained on an HPC with an Intel Core i9-10940X CPU, an NVIDIA Titan RTX GPU, and 64 GB of memory.

³ <https://www.agisoft.com/>

⁴ <https://github.com/agisoft-llc/metashape-scripts>

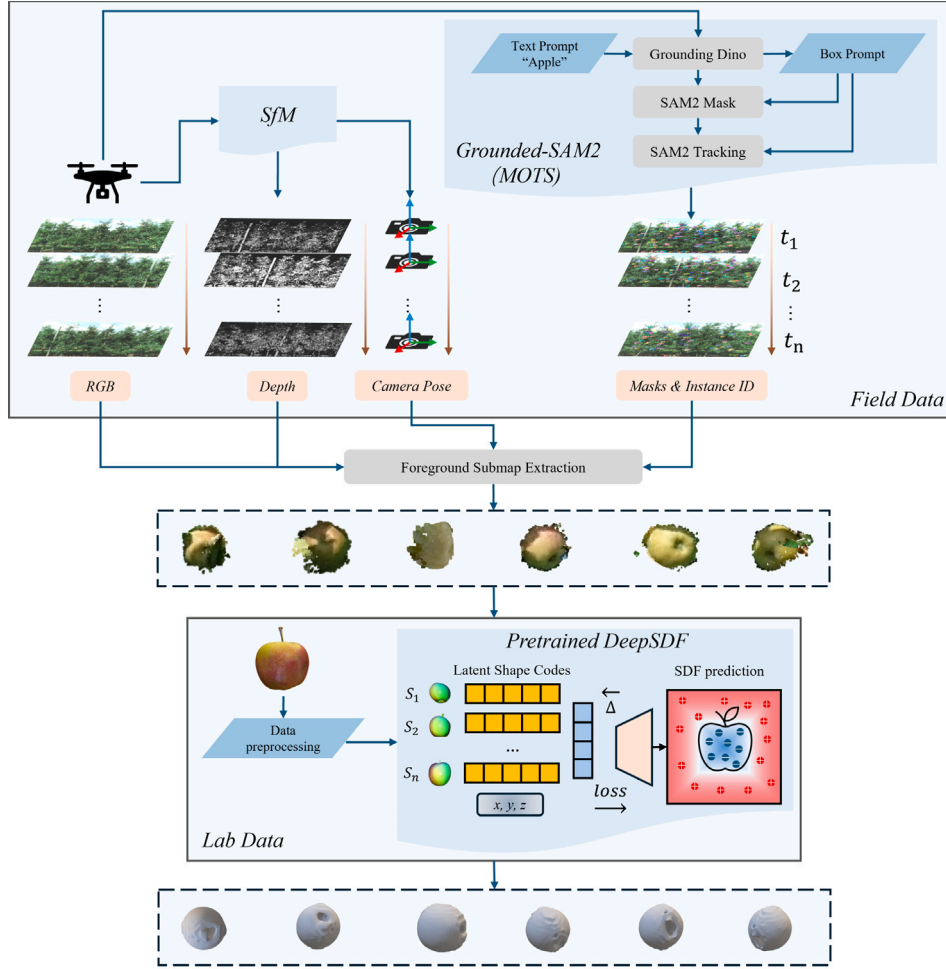


Fig. 6. UAV-based panoptic mapping framework for orchard environments. The video frames collected by UAV were input to both MOTS and SfM. From MOTS, the apple instance IDs and corresponding masks \mathcal{M} and bounding boxes across frames were output. From SfM, the depth D , camera poses \mathcal{P} , and the orchard point clouds were generated. After that, our mapping methods combined all camera poses \mathcal{P} , depth D , RGB images, and masks \mathcal{M} of each apple instance to get partial point clouds for each apple submap. Meanwhile, the pretrained DeepSDF was optimized for apple completion.

3.4. Panoptic mapping

In our previous work (Pan et al., 2023), we proposed a panoptic mapping pipeline using an RGB-D camera mounted on a ground robot for sweet pepper mapping and completion in a greenhouse environment, which aimed at online operation. In contrast to that, we use a monocular high-resolution camera mounted on a UAV in an outdoor apple orchard, aiming at offline panoptic mapping and reconstruction, i.e., a setup with different modalities.

We first built a panoptic point cloud map that decomposes the scene into the background (orchard) and individual foreground (apple instances) and then assigned submaps S to every panoptic entity. Then, the background submap was downsampled to a 0.01 resolution voxel size to save memory, while the apple instance submaps kept their original resolution. Each partial foreground point cloud was derived from the selected depth pixels by masks \mathcal{M} in each frame, which are then transformed to the world coordinate system using camera poses \mathcal{P} . Then, the foreground submaps of each apple entity were derived by accumulating the unprojected point cloud from multiple frames. Thus, the tracking robustness is crucial for the reprojection, which requires comprehensive evaluations.

Then, we conducted optimization by two loss functions. First we kept the points from the target point cloud \mathcal{P}_S of the foreground submap close to the iso-surface of the SDF predicted by D_θ , which

minimizes the surface reconstruction loss \mathcal{L}_s by Eq. (1).

$$\mathcal{L}_s = \frac{1}{|\mathcal{P}_S|} \sum_{p \in \mathcal{P}_S} D_\theta^2(\mathcal{T}_{ow} p^w, z^*) \quad (1)$$

$$\mathcal{L} = w_s \mathcal{L}_s + w_r \mathcal{L}_r, \quad (2)$$

where $\mathcal{L}_r = \|z\|^2$, w_s and w_r are the weights for each loss term. We then used Levenberg–Marquardt with analytical Jacobians to solve $\xi_{ow}^*, z^* = \arg \min \mathcal{L}$, where $\xi_{ow} \in \mathbb{R}^7$ belongs to $\text{sim}(3)$ Lie Algebra of \mathcal{T}_{ow} . We initialized the latent shape code z as $\mathbf{0}_C$. And the ξ_{ow} is initialized as an identity rotation, a scaling of 1, and a translation from the bounding box center of the submap point cloud to the origin point. At each iteration, with damping parameter λ , the update δx to the estimated parameter vector $[\xi_{ow}, z]^T$ is computed as

$$\delta x = [\delta \xi_{ow}, \delta z]^T = (H + \lambda \text{diag}(H))^{-1} g, \quad (3)$$

where $H = J^T P J$ is the approximate Hessian matrix, $g = J^T P b$ is the gradient of the target function, and J , P , b are the Jacobian matrix, weight matrix and residual vector, respectively. With all Jacobians and residuals, we can optimize Eq. (3) as

$$[\xi_{ow}^{(t+1)}, z^{(t+1)}]^T = [\xi_{ow}^{(t)}, z^{(t)}]^T + \delta x^{(t)} \quad (4)$$

until convergence. After Eq. (4) convergence, the complete 3D apple model can be reconstructed using marching cubes (Lorensen and Cline, 1998) with the optimized latent shape code z^* at 3D grid queries in

the apple's canonical coordinate system. Then the 3D model can be transformed into the world coordinate system using $T_{ow}^* = \exp(\mathcal{E}_{ow}^*)$. The panoptic mapping method was also implemented on an HPC with an Intel Core i9-10940X CPU, NVIDIA Titan RTX GPU, and 64 GB of memory.

3.5. Proposed method performance assessment

Three steps were conducted to comprehensively evaluate the performance of the panoptic mapping framework.

3.5.1. MOTS evaluation metrics

The MOTS performance was evaluated under four different UAV flight modes, see Fig. 2. We annotated 1890, 756, 470, and 1015 apple instances across 10 sequential frames in each flight mode, separately.

Three typical metrics were used to evaluate tracking and segmentation performance (Voigtlaender et al., 2019): MOTSA (multiple object tracking and segmentation accuracy), sMOTSA (soft MOTSA), and MOTSP (multiple object tracking and segmentation precision), as defined in Eqs. (7), (8), and (9).

$$c(h) = \begin{cases} \arg \max_{m \in M} \text{IoU}(h, m), & \text{if } \max_{m \in M} \text{IoU}(h, m) > 0.5 \\ \emptyset, & \text{otherwise.} \end{cases} \quad (5)$$

$$\widetilde{TP} = \sum_{h \in TP} \text{IoU}(h, c(h)) \quad (6)$$

$$\text{MOTSA} = \frac{|TP| - |FP| - |IDS|}{|N|} \quad (7)$$

$$s\text{MOTSA} = \frac{\widetilde{TP} - |FP| - |IDS|}{|N|} \quad (8)$$

$$\text{MOTSP} = \frac{\widetilde{TP}}{|TP|} \quad (9)$$

where $M = m_1, \dots, m_N$ with $m_i \in [0, 1]$ are the ground truth pixel masks, $H = h_1, \dots, h_K$ with $h_i \in [0, 1]$ are the non-empty hypothesis masks, TP are true positives, \widetilde{TP} are soft true positives, FP are false positives, IDS are instance ID switches, N is the total number of ground truth masks.

Based on the performance of the MOTS from the previous metrics in the four flight modes, we can get a preliminary idea of the optimal MOTS result for a particular flight mode. However, the conventional evaluation of MOTS requires annotations, which are labor-intensive and time-consuming. Thus, we also designed a novel evaluation algorithm to validate MOTS results without any prior ground truth annotations. As shown in Alg. 1, the total frames N are randomly divided into 10 parts S , each with 5 sequential frames. In each part, 10 apple instances are randomly selected to calculate their average centroid movement speed with masks \mathcal{M} , depth D , intrinsic matrix K , and camera poses \mathcal{P} . Meanwhile, the average movement speeds of the camera are also calculated and compared with the movement speeds of the apple instances. The average movement speeds of the camera can be considered as a reference for evaluating the tracking performance of apple instances. As another complementary evaluation, the method was implemented for the particular flight mode with the optimal performance of the MOTS.

3.5.2. Apple reconstruction quality assessment in lab

To quantitatively evaluate the accuracy of apple shape reconstruction, we compute the Chamfer distance (CD) between the reconstructed apple point clouds from the DeepSDF model and the ground truth scans acquired in a controlled laboratory setting. The CD measures the average nearest-neighbor distance between two point sets, providing an assessment of shape fidelity. Given a reconstructed apple point cloud P_r and a ground truth scanned apple P_g , the CD is computed as in Eq. (10).

$$D_C(P_r, P_g) = \frac{1}{|P_r|} \sum_{p \in P_r} \min_{q \in P_g} \|p - q\|^2 + \frac{1}{|P_g|} \sum_{q \in P_g} \min_{p \in P_r} \|q - p\|^2 \quad (10)$$

Algorithm 1 MOTS Performance Evaluation Without Ground Truth

Require: Total frames N , Masks \mathcal{M} , Depth D , Camera poses \mathcal{P} , Intrinsic matrix K

Ensure: MOTS Performance Score

- 1: **Extract camera positions** \mathcal{P} from XML file
- 2: **Select** 10 random frame windows $S = \{S_1, S_2, \dots, S_{10}\}$, where each window compose by 5 sequential frames
- 3: Initialize empty list \mathcal{R} for speed ratios
- 4: **for** each part $S_i \in S$ **do**
- 5: **Randomly select** 10 apple instances \mathcal{I} in the first frame of S_i
- 6: Initialize empty lists $\mathcal{V}_{\text{instance}}, \mathcal{V}_{\text{camera}}$
- 7: **for** each consecutive frame pair $(t, t+1) \in S_i$ **do**
- 8: Load mask centroids C_t and C_{t+1} from \mathcal{M}
- 9: Load depth maps Z_t and Z_{t+1} from D
- 10: **Extract depth values** z_t and z_{t+1} at centroids
- 11: **Compute apple instance speeds:**

$$v_{\text{instance},j} = \frac{\|X_{t+1}^j - X_t^j\|}{\Delta t}, \quad j \in \mathcal{I}$$

- 12: **Compute camera movement speed:**

$$v_{\text{UAV},t} = \frac{\|P_{t+1} - P_t\|}{\Delta t}$$

- 13: Append $v_{\text{UAV},t}$ to $\mathcal{V}_{\text{camera}}$
- 14: Append all $v_{\text{instance},j}$ to $\mathcal{V}_{\text{instance}}$
- 15: **end for**
- 16: **Compute average speed ratio:**

$$R = \frac{\mathbb{E}[\mathcal{V}_{\text{instance}}]}{\mathbb{E}[\mathcal{V}_{\text{camera}}]}$$

- 17: Append R to \mathcal{R}
- 18: **end for**
- 19: **Compute final MOTS performance score:**

$$\text{MOTS Score} = \mathbb{E}[\mathcal{R}]$$

- 20: **Return** MOTS Score

This metric computes the mean squared distance between each point in P_r and its closest counterpart in P_g , and vice versa. A lower CD indicates a higher reconstruction accuracy, as it reflects improved alignment and shape consistency between the reconstructed and ground truth apple models.

3.5.3. Field panoptic reconstruction assessment

To assess the effectiveness of the multi-resolution approach for the field panoptic reconstruction, we evaluate the memory consumption of both the original high-resolution background submap and the low-resolution background submap. In the proposed method, the resolution of the background (orchard) is reduced to mitigate the memory overhead while preserving the high resolution of the foreground (apple instances) to maintain the detailed representation of the objects of interest. It aims to optimize computational resources by focusing on areas where high precision is critical, while sacrificing some details of the background. The apple entity submaps were reconstructed as completed apple meshes and evaluated by comparing the maximum diameters of the completed meshes with manual measurements.

$$RMSE = \sqrt{\frac{\sum_{i=1}^n (\hat{y}_i - y_i)^2}{n}} \quad (11)$$

where the \hat{y}_i is the maximum diameter of the predicted apple mesh, y_i is the maximum diameter of the manual measurement, n is the number of the samples.

Table 2

Apple detection and tracking performance with AppleMOTS and Grounded-SAM2 under four different flight modes. The bold numbers indicate the best results.

Model	Flight mode	MOTSA (%)↑	sMOTSA (%)↑	MOTSP (%)↑	Avg Time (FPS)↑
TrackRCNN (de Jong et al., 2022; Voigtlaender et al., 2019)	A	−2.69	−2.73	58.71	0.15
	B	−6.70	−7.04	57.74	
	C	−8.30	−8.30	0.00	
	D	−4.43	−4.43	0.00	
PointTrack (de Jong et al., 2022; Xu et al., 2020)	A	8.47	6.38	64.98	3.24
	B	1.03	−3.79	62.19	
	C	−2.03	−3.85	14.21	
	D	−1.38	−2.43	6.98	
Grounded-SAM2 (Ren et al., 2024)	A	34.84	14.95	71.98	1.31
	B	2.72	−1.20	72.13	
	C	−40.18	−57.05	66.96	
	D	1.99	1.25	73.29	

The RMSE (root mean square error) of the maximum diameters between the predicted completed apple meshes and manual measurements was calculated to show the performance of our method in the field (Eq. (11)).

4. Results

4.1. Apple MOTS performance in orchard

Overall, the results in Table 2 and Fig. 8 show that the MOTS results of flight mode A through the Grounded-SAM2 model perform the best to track apples across frames in occluded orchard conditions. The apple instances were clustered closely together in each tree, often with overlapping shapes and partial occlusions.

The results of two MOTS methods for apple detection and tracking across are presented in Table 2. Four different flight modes (Fig. 2), and three MOTS metrics defined in Eqs. (5), (6), (7) are assessed to benchmark apple MOTS performance. Grounded-SAM2 substantially outperforms AppleMOTS (de Jong et al., 2022) using either TrackRCNN (Voigtlaender et al., 2019) and PointTrack (Xu et al., 2020) model across the four evaluated flight modes. AppleMOTS with TrackRCNN model yields negative scores for both MOTSA and sMOTSA in every mode (ranging from −2.69% to −8.30% and −2.73% to −8.30%, respectively), and its apple segmentation precision (MOTSP) never exceeds 58.71%, collapsing to 0% in modes C and D. The negative values indicate low tracking performance with too many ID switches of the instances.

By contrast, Grounded-SAM2 attains the highest MOTSA (34.84%) and sMOTSA (14.95%) in mode A, while also securing the top MOTSP (73.29%) in mode D. But the apple detection rate of Grounded-SAM2 in mode D is too low which limits the practical application performance (Fig. 2). Although performance for Grounded-SAM2 degrades markedly in mode C (MOTSA −40.18%, sMOTSA −57.05%), it still delivers superior precision compared to AppleMOTS in every mode, suggesting greater robustness to variations in flight dynamics.

In terms of inference efficiency, PointTrack achieved the highest processing speed at 3.24 FPS, followed by Grounded-SAM2 at 1.31 FPS, while TrackRCNN was slowest at 0.15 FPS. Combined with the tracking and detection performance, Grounded-SAM2 provides more practical potential for large-scale orchard applications.

To comprehensively assess the MOTS performance in the field, we also followed our proposed evaluation algorithm, which does not require any annotation (Alg. 1). The heatmap (Fig. 8) illustrates the speed ratios between instance movement and camera movement across multiple frame sequences by our random sampling methods. Most of the speed ratio values remain close to 1.0, indicating that most detected apple instances move at a rate similar to the camera, suggesting a stable tracking performance. However, a few extreme values are observed, where certain apple instances exhibit significantly higher or lower

Table 3

Apple reconstruction performance based on Chamfer distance. The comparison is between the DeepSDF model and sphere approximation (SA).

Method	Error ↓
SA	1.67 ± 0.08
DeepSDF	0.98 ± 0.05

speed ratios, as highlighted in the darker blue and orange regions of the heatmap. The high speed ratios may result from apple instance ID switches, occlusions, or tracking inconsistencies in challenging orchard conditions. And the zero value of speed ratios could be caused by the UAV hovering during flight. Overall, the apple instance tracking performance with Grounded-SAM2 in flight mode A is acceptable in the orchard for data association.

4.2. Evaluation of apple reconstruction in lab environment

A total of 88 individual scanned apples were trained by DeepSDF, another 12 scanned apples were used as the test set to evaluate the performance of the model. In apple orchards, apples are commonly approximated as spheres to simplify geometric modeling and analysis (Gené-Mola et al., 2021), thus, we also conduct a benchmark of apple reconstruction comparing the sphere approximation method and the DeepSDF model. 70 mm Tables 3 and A.6 present the reconstruction performance of individual apples using Chamfer distance (Eq. (9)) as the evaluation metric, where lower values indicate better shape reconstruction accuracy. Overall, DeepSDF outperforms SA in all cases, with lower Chamfer distances across all apples. For example, for Apple 10, 12, 13, 14, 100, and 101, DeepSDF achieves significantly lower errors which are lower than 1.00 mm, (e.g., 0.93 mm, 0.97 mm, 0.97 mm, 0.89 mm, 0.98 mm, respectively) compared to SA (1.59 mm, 1.70 mm, 1.65 mm, 1.62 mm, 1.50 mm, 1.73 mm, respectively). These consistent improvements suggest that DeepSDF can more accurately capture the geometric complexity of apple shapes compared to an SA model.

4.3. Panoptic mapping for apple shape completion in apple orchard

In the apple orchard, a total of 2729 individual apples were detected and reconstructed with our panoptic mapping method within row 1 under flight mode A. We classified the occluded apple on five levels of occlusion according to the visibility of pixels in the images in the real apple orchard (Du et al., 2023), following common categorization in orchard phenotyping studies (Table 5). As for the recorded 135 apples with artificial labels, 12.6% apples were at the A level, 19.3% apples were at B level, 27.4% apples were at C level, 30.4% apples were at D level, and 10.4% apples were at E level which were not detected or

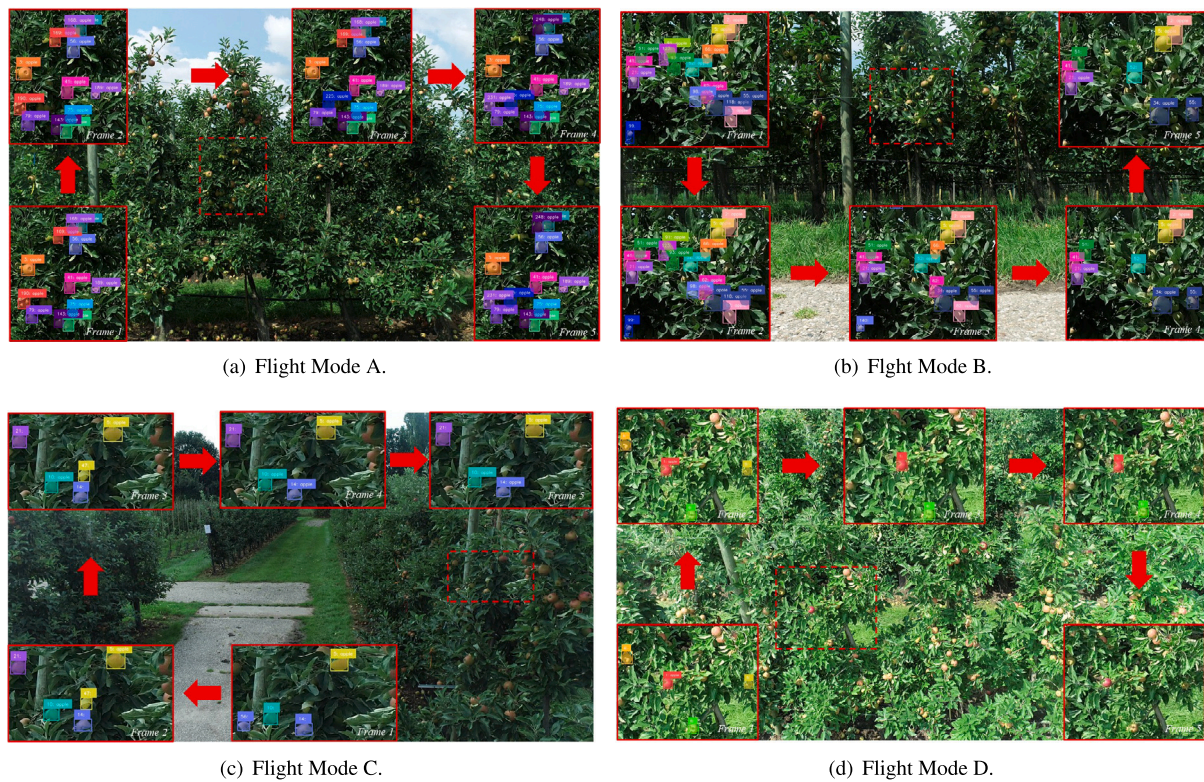


Fig. 7. MOTS examples of four different flight modes across five frames. The dashed red box shows the same targeted regions in the five frames. Different color masks and bounding boxes in the red boxes indicate different apple instances.

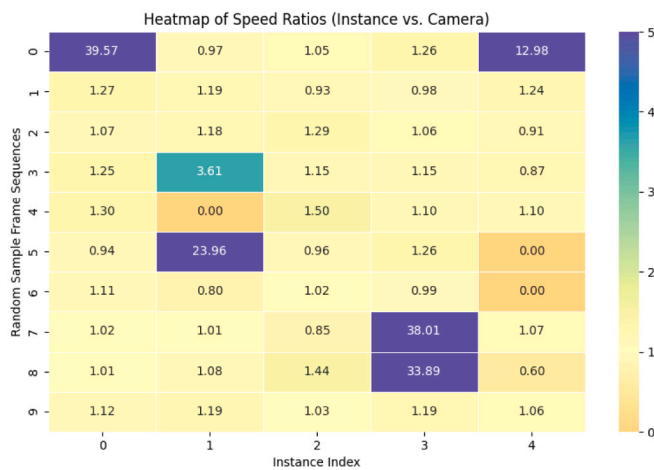


Fig. 8. The heatmap of speed ratios for the preliminary optimal flight mode A through evaluation with manual annotations (Instance Speed/Camera Speed).

tracked by our method. 120 from 135 sampled apples under different occlusion levels were detected and reconstructed their shapes (Fig. 9(b)). The level of apple occlusion showed a positive correlation with reconstruction error (Fig. 9(a)), while the error reduced at the level of D. This reduction of error at the level D may cause by the increase of the occlusion level leading to more accurate masks by Grounded-SAM2 than the level B and C. And the total RMSE of the 120 apples between predicted apple diameters and ground truth measurements was 1.85 cm.

The average computational time cost during the SfM steps for the four flight modes was around 24 min, as stated in Table 4. And for the apple shape completion in orchard environments, the average reconstruction time with pretrained DeepSDF was 0.58 sec for each

partial observed apples. The memory cost of the background submap was reduced by 31.9% from 1.75 GB to 571 MB after downsampling the resolution of the point cloud.

5. Discussion

5.1. Effect of dense and occluded homogeneous objects on UAV-based MOTS

MOTS techniques have been predominantly developed and benchmarked using datasets from urban driving, surveillance, or robotics domains, such as KITTI MOTS (Geiger et al., 2012), BDD100K (Yu et al., 2020), and TAO (Dave et al., 2020). These datasets typically feature scenes where foreground objects (e.g., vehicles or pedestrians) are relatively large, sparse in distribution, and exhibit distinct motion patterns from the background. Furthermore, foreground instances usually occupy a significant portion of the image and are well-separated from one another, allowing MOTS algorithms to leverage appearance features and temporal continuity for reliable instance association.

In contrast, orchard environments introduce a set of unique challenges that deviate significantly from the assumptions underlying conventional MOTS models. First, the foreground objects of interest (e.g., fruits, leaves) are small, visually homogeneous, and densely distributed (Fig. 7). Fruits with the same variety often have near-identical shape, color, and texture, making it difficult for MOTS algorithms to distinguish individual instances, especially when fruits overlap or cluster together. Second, unlike dynamic objects in urban scenes, fruits are largely stationary, and their motion is highly correlated with the background (i.e., the tree canopy). This diminishes the effectiveness of motion-based tracking and increases the risk of ID switches or tracking drift (de Jong et al., 2022). Moreover, the presence of occlusions, whether partial or complete, further complicates the association of fruit instances across frames, potentially causing tracking failures or object drift (Table 2).






Table 4

The computational costs of SfM steps under four different flight modes. “# Img” means the number of images for each flight mode.

Flight mode		A	B	C	D
# Img		210	271	60	135
SfM	Align Photo	1 min 57 s	2 mins 2s	1 min 11 s	1 min 29 s
Steps	Build Point Cloud	33 mins 8 s	28 mins 38 s	9 mins 19 s	19 mins 12 s

Table 5

Occlusion levels in the apple orchard. Five occlusion levels are defined according to the visibility.

Level	Visible area ratio	Examples	Description
A	≥90%		Fully visible or almost no occlusion
B	70%–89%		Mild occlusion
C	40%–69%		Medium occlusion
D	10%–39%		Heavily occlusion
E	<10%		Extremely occlusion or barely visible

The data acquisition process using UAV platforms in orchards imposes additional influences on MOTS performance. Compared to conventional MOTS scenarios, some studies reached lower performance of MOTS when using UAVs to track fruits (de Jong et al., 2022; Ariza-Sentís et al., 2023). Although they proposed a novel model or using multiple views to improve fruit tracking performance, some of the metrics, such as sMOTSA and MOTSA, presented even negative values. Instead of the unique features mentioned above in orchard environments, the UAV flight modes may be another factor that influences the MOTS performance. The studies mentioned above implement MOTS using only single flight mode, which may not be suitable for different crops and different applications. Our proposed method employs Grounded-SAM2 for MOTS, which demonstrates promising performance in the field, particularly under flight mode A (Table 2). Among the four UAV flight strategies tested, flight mode A achieved the highest MOTSA and sMOTSA scores (Table 2), indicating superior tracking accuracy and robustness across dense and partially occluded apple rows. The improved performance in this mode can be attributed to the favorable side-view perspective and moderate proximity to the canopy, which facilitates clearer object contours and minimizes severe occlusions. In contrast, flight mode D, despite achieving high segmentation precision (MOTSP), detected significantly fewer apple instances (Table 2). This suggests that while fewer, visually salient apples were well tracked in these modes, the lower detection rate limited their applicability for large-scale phenotyping.

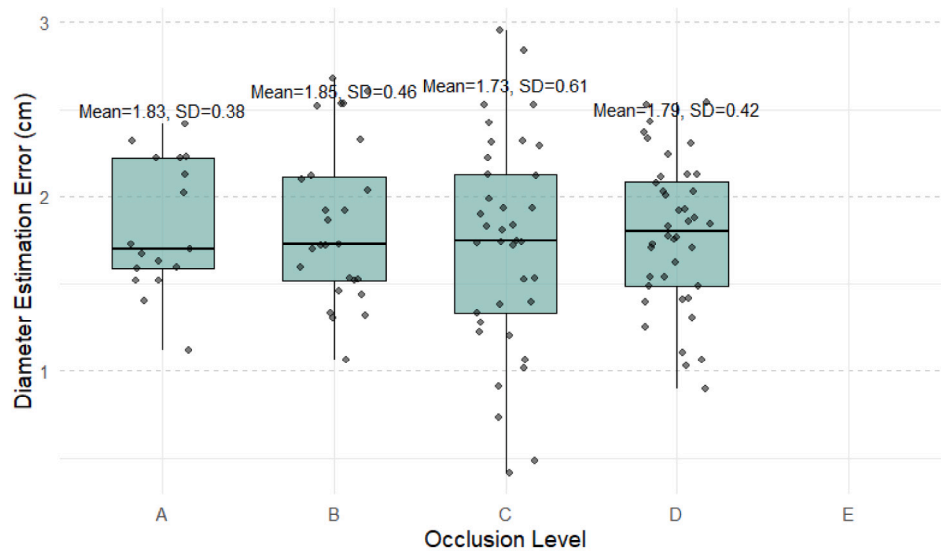
We also found that the manual annotation is crucial for MOTS evaluation in current research (Geiger et al., 2012; Yu et al., 2020; Dave et al., 2020). It is essential to evaluate the performance with a public dataset with annotations when developing a novel MOTS algorithm, but it is difficult for applications that only implement MOTS in a new domain. However, labeling for evaluating MOTS is challenging due to the need for consistent and precise instance-level annotations across frames in complex, dynamic scenes. To simplify the MOTS evaluation and boost MOTS applications, a novel evaluation algorithm was introduced, estimating object motion consistency without requiring prior annotations (Alg. 1). The “velocity ratio-based method” for MOTS evaluation only requires the velocity of the UAV, which is available

from state estimation (e.g., GNSS, IMU, visual odometry, etc.). Results from this unsupervised evaluation complement manual metrics, reinforcing the reliability of the proposed approach under field conditions. Nonetheless, extreme motion inconsistencies in some sequences suggest that Grounded-SAM2, while effective, remains sensitive to motion blur and abrupt occlusions, which are common in UAV-based videos in orchard environments (Table 2 and Fig. 8).

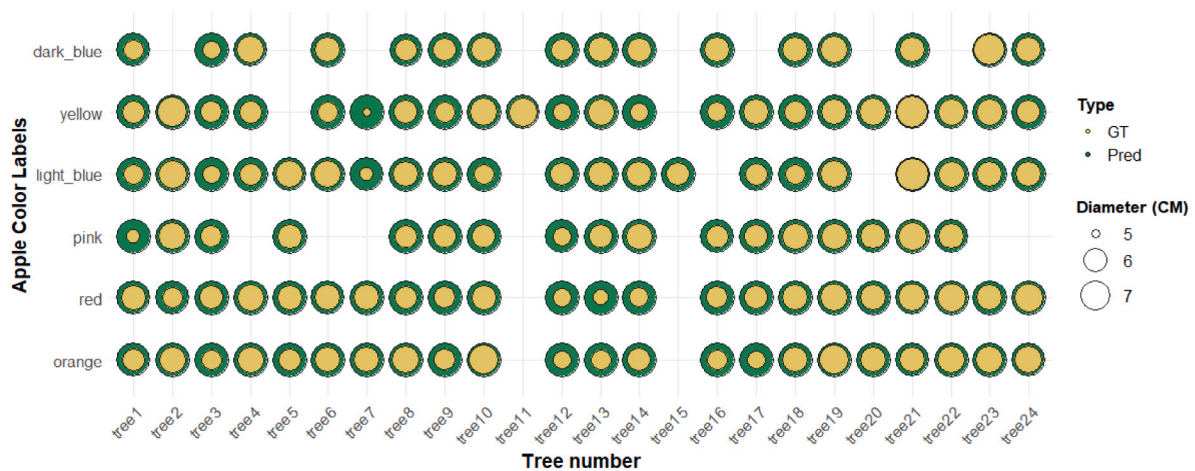
5.2. 3D fruit shape completion under occlusions

Accurate 3D shape completion of fruit from UAV-based monocular observations under occlusions remains a critical challenge in orchard environments due to complex scene layouts, heavy foliage, and limited perspectives from UAV. In our study, we addressed this challenge by integrating multi-view observations with a learning-based shape prior (DeepSDF) to infer complete 3D fruit geometries from sparsely visible surfaces.

Our results demonstrate that even when a significant portion of the fruit is occluded in individual frames, the reconstruction pipeline is capable of completing plausible and structurally consistent fruit shapes under occlusion conditions (Fig. 9 and 10). The accuracy of the reconstruction reaches centimeter level in our method. Aside from our 3D mapping method to address occlusion issues in the orchard, Gené-Mola et al. (2023) proposed a 2D amodal solution to predict the complete masks for the occluded apples. Their results from amodal completion for fruit size estimation have shown promising results, as they can obtain an MAE (mean absolute error) of 4.17 mm and an R^2 of 0.91 between the mean diameters measured manually and predicted by the model. But their method only detects and measures the apples with visibility higher than 60%, while according to our visibility level distribution (Table 4, and Fig. 9(a)), we can detect and reconstruct the apple with visibility higher than 10%. And their method cannot estimate fruit distribution and true geometric properties in the real orchards. Furthermore, their 2D amodal method typically requires extensive human labeling to train the model, which is time-consuming and limits scalability for large-scale orchard applications. Active Shape Models (ASM) are effective for capturing statistical shape variations,



(a) Prediction error across occlusion levels.



(b) The comparison of apple diameters between ground truth measurements and predicted reconstructions.

Fig. 9. Results of apple diameters between the predicted and ground truth and error analysis. (a) shows the prediction error of apple diameter across different occlusion levels. Each box shows the distribution of absolute prediction errors under a specific visibility condition, ranging from fully visible (Level A) to nearly invisible (Level E). Missing predictions result in empty boxes. (b) shows the comparison of apple diameters from 24 apple trees. The color labels indicate predicted values, and the yellow colors indicate ground truth values. The x- and y-axes illustrate the number of the fruit tree and the color label of the sampled apple in the orchard, respectively. The size of the circles represents the diameter of the apples. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

and can be applied in both 2D and 3D domains (Coates et al., 1995). However, the ASM requires manual landmark annotation and is less robust under strong occlusions, since missing landmarks cannot be reliably inferred, especially in complex orchard environments where fruits are frequently obscured by foliage and branches.

In addition, the minimum fruit size detectable by our method is influenced by three main factors. First, the UAV flight settings (altitude, camera resolution, and distance to the crops) determine the number of pixels covering each fruit, which directly affects segmentation reliability. Second, the diversity of training samples for DeepSDF plays an important role, as including a wider range of apple sizes and shapes improves the model's ability to reconstruct smaller fruits. Third, the phenological stage of the orchard (e.g., early growth vs. mature fruiting) also impacts detectability, since fruits in early stages are both smaller and more occluded by foliage. Together, these factors imply

that the minimum detectable size is not a fixed threshold but rather depends on sensing configuration, training data, and orchard growth stage. Thus, the minimum detectable fruits for our field experiment is around 45 mm according our flight settings, training dataset distribution and growth stage of fruit trees in the orchards.

5.3. Limitations and future work

Our DeepSDF model was pretrained on our lab 3D apple dataset where the apples were in the mature stage. While the apples in the field were in the expansion stage during data collection. The fact that apples were at different stages of growth leads to a systematic deviation of the diameter of apples collected in the laboratory from the diameter of apples collected in the field in the orchard (Fig. 5). The average deviation of the apple diameter was 1.6 cm between lab dataset and

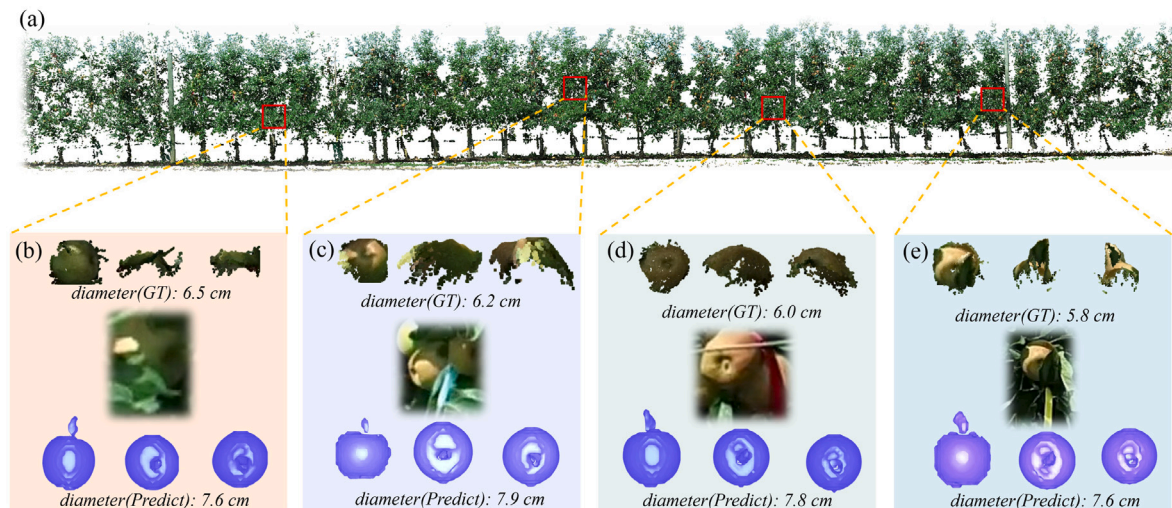


Fig. 10. The panoptic mapping of the orchard row and the apple shape completion results. (a) shows the panoptic mapping orchard with four different regions highlighted in red boxes. (b), (c), (d) and (e) show four partial apple point clouds with our measurements under three perspectives and their predicted complete meshes. The middle images show the occlusion conditions. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

field dataset, which was also in line with our predicted results (1.85 cm MAE).

Another limitation relates to MOTs performance. Because the data association between 2D images and 3D point clouds depends on the results from MOTs, the performance of MOTs is crucial for the whole system. But there are unavailable issues (e.g., instance ID switches and wrong detection) in the current MOTs frameworks, which lead to some failures of following apple 3D reconstructions. Thus, in our future study, we will first collect more fruit data covering more growth stages and more varieties to address the generalization issues. Then, we will continue to work on exploring and developing more robust MOTs algorithms that are suitable for agricultural scenarios.

Furthermore, our field experiments were conducted in row-planted orchards, which are the most common orchard production systems in Europe. However, there are other planting patterns (e.g., quincunx, hexagon, or triangle) which aim to maximize space and resource efficiency (Hendricks, 1996). In such scenarios, the UAV path would need to be adapted to acquire sufficient multi-view overlap. The active view selection proved to be one of the promising methods to improve the fruit reconstruction qualities, which can also adapt to different orchard planting patterns. Some studies developed next best view planning methods to improve the quality of fruit reconstruction and scene mapping for robot arms in greenhouses (Menon et al., 2023; Jose et al., 2025). Therefore, we believe that it would be possible for UAV to plan the next best view to increase the visibility of targeted fruits and the quality of reconstruction confidence in real orchards.

The external conditions during UAV flights, such as wind and illumination also influence the performance of the proposed method. In terms of wind, the modern UAV platform are commonly equipped with an advanced obstacle avoidance system and a flight control algorithm supported by high-accuracy RTK, RGB, and IMU sensors, which ensured stable flights under light wind conditions (5 m/s). Nevertheless, strong winds may introduce drift and motion blur in images, thereby affecting both SfM reconstruction and MOTs. For illumination, strong sunlight at low angles can cause glare, lens flare, and harsh shadows, which reduce MOTs quality and degrade feature matching in SfM. Non-uniform illumination across orchard rows may further contribute to inconsistent reconstruction results. To mitigate these effects, flights should ideally be scheduled under diffuse illumination (e.g., cloudy days) or at midday

when shadowing is minimal. Future work could also consider the use of high dynamic range (HDR) imaging to further reduce the sensitivity of the system to lighting variations or introduce LiDAR and other sensors for sensor fusion to make the system more robust.

Large digitization footprint is also a limitation of our framework. Due to the high-resolution of the images, the single flight with flight mode A of around 1 min for one row produces approximately 1800 frames, corresponding to 700 MB of raw video data. After SfM and MOTs, the generated depth images typically reach 1.27 GB for one row flight. For our outline process time is acceptable, around 20 mins for SfM reconstruction (Table 4), 2 mins for MOTs inference on around 200 frames (Table 2), and 9 mins for DeepSDF reconstruction of 2000 apple instances. Potential solutions for our future work may include onboard data compression, multi-modality fusion, lightweight implicit models, and GPU acceleration, which can reduce storage and processing bottlenecks.

6. Conclusion

This study presents a novel and fully automatic pipeline for UAV-based monocular 3D panoptic mapping for fruit shape completion in orchard environments. Our main contributions are integrating state-of-the-art techniques including Grounded-SAM2 for MOTs, DeepSDF for implicit 3D shape reconstruction, and SfM-based photogrammetry, the proposed framework effectively addresses key challenges in fruit phenotyping, such as occlusion, scale variation, and the dense distribution of homogeneous fruits.

Through comprehensive evaluations, the method demonstrates robust instance tracking under real-world field conditions and significantly improves the fidelity of reconstructed apple shapes compared to conventional sphere approximations. The proposed MOTs evaluation approach, capable of operating without ground truth annotations, further supports scalable and efficient model validation in complex agricultural settings.

The unified framework bridges high-resolution, instance-level reconstruction and scalable UAV-based data acquisition, offering a promising tool for large-scale, precise fruit phenotyping. Future work may explore the extension of this framework to other crop types and integrate multimodal sensing to further enhance reconstruction robustness and generalization under diverse field conditions.

CRedit authorship contribution statement

Kaiwen Wang: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Yue Pan:** Writing – review & editing, Visualization, Validation, Software, Methodology, Formal analysis. **Federico Magistri:** Validation, Software, Methodology. **Lammert Kooistra:** Writing – review & editing, Visualization, Supervision, Project administration, Methodology, Conceptualization. **Cyrril Stachniss:** Writing – review & editing, Visualization, Supervision, Project administration, Conceptualization. **Wensheng Wang:** Writing – review & editing, Supervision, Project administration, Funding acquisition, Conceptualization. **João Valente:** Writing – review & editing, Visualization, Supervision, Project administration, Methodology, Formal analysis, 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.

Acknowledgment

We would like to thank Zhen Cao, Tianyi Jia, Álvaro Lau Sarmiento, and Berry Onderstal for their support during data acquisition. We are also grateful to Pieter van Dalfsen for his support regarding the apple orchards.

Appendix. I

Algorithm 2 Mesh Alignment and Rescaling for Apple Models

```

1: procedure ALIGNANDRESCALEMESHERS
2:   Set rotation_matrix  $\leftarrow \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & -1 \\ 0 & 1 & 0 \end{bmatrix}$ 
3:   Set scale_factor  $\leftarrow 1000.0$ 
4:   Load reference mesh and apple meshes
5:   (ref_centroid, ref_axis)  $\leftarrow$  FitCylinderAxis(ref_mesh)
6:   ref_centroid_apple  $\leftarrow$  ComputeCentroid(ref_mesh_apple)
7:   for  $i \in [1, N]$  do
8:     Load stick_mesh and apple_mesh with index  $i$ 
9:     if stick or apple mesh does not exist then
10:       continue
11:     end if
12:     (src_centroid, src_axis)  $\leftarrow$  FitCylinderAxis(stick_mesh)
13:     T  $\leftarrow$  ComputeTransformationMatrix(src_axis, ref_axis)
14:     Apply T to apple_mesh
15:     src_centroid_apple  $\leftarrow$  ComputeCentroid(apple_mesh)
16:     T_apple  $\leftarrow$  ComputeAlignMatrix(src_centroid_apple, ref_centroid_apple)
17:     Apply T_apple to apple_mesh
18:     Save aligned apple_mesh
19:   end for
20:   for  $i \in [1, N]$  do
21:     Load aligned apple_mesh
22:     if mesh does not exist then
23:       continue
24:     end if
25:     Apply rotation_matrix and divide vertices by scale_factor
26:     Move transformed mesh to origin point
27:     Save transformed mesh
28:   end for
29: end procedure

```

See Table A.6.

Table A.6

Apple reconstruction performance based on Chamfer distance. The comparison is between the DeepSDF model and sphere approximation (SA).

Apple ID	SA (mm) ↓	DeepSDF (mm) ↓
Apple10	1.59	0.93
Apple11	1.63	1.03
Apple12	1.70	0.97
Apple13	1.65	0.93
Apple14	1.62	0.97
Apple15	1.78	1.05
Apple16	1.69	1.00
Apple17	1.74	1.03
Apple18	1.74	1.01
Apple19	1.66	1.01
Apple100	1.50	0.89
Apple101	1.73	0.98

Data availability

Our 3D apple data in the lab and UAV-based video data in the apple orchard can be found in: [10.5281/zenodo.15635994](https://doi.org/10.5281/zenodo.15635994).

References

- Amatya, S., Karkee, M., Gongal, A., Zhang, Q., Whiting, M.D., 2016. Detection of cherry tree branches with full foliage in planar architecture for automated sweet-cherry harvesting. *Biosyst. Eng.* 146, 3–15. [http://dx.doi.org/10.1016/j.biosystemseng.2015.10.003](https://doi.org/10.1016/j.biosystemseng.2015.10.003).
- Ariza-Sentís, M., Baja, H., Vélez, S., Valente, J., 2023. Object detection and tracking on UAV RGB videos for early extraction of grape phenotypic traits. *Comput. Electron. Agric.* 211, 108051. [http://dx.doi.org/10.1016/j.compag.2023.108051](https://doi.org/10.1016/j.compag.2023.108051).
- Cootes, T.F., Taylor, C.J., Cooper, D.H., Graham, J., 1995. Active shape models-their training and application. *Comput. Vis. Image Underst.* 61 (1), 38–59.
- Dave, A., Khurana, T., Tokmakov, P., Schmid, C., Ramanan, D., 2020. TAO: A large-scale benchmark for tracking any object. In: *Computer Vision – ECCV 2020*. Springer International Publishing, pp. 436–454. [http://dx.doi.org/10.1007/978-3-030-58558-7_26](https://doi.org/10.1007/978-3-030-58558-7_26).
- de Jong, S., Baja, H., Tamminga, K., Valente, J., 2022. APPLE MOTS: Detection, segmentation and tracking of homogeneous objects using MOTs. *IEEE Robot. Autom. Lett.* 7 (4), 11418–11425. [http://dx.doi.org/10.1109/lra.2022.3199026](https://doi.org/10.1109/lra.2022.3199026).
- Du, X., Cheng, H., Ma, Z., Lu, W., Wang, M., Meng, Z., Jiang, C., Hong, F., 2023. DSW-YOLO: A detection method for ground-planted strawberry fruits under different occlusion levels. *Comput. Electron. Agric.* 214, 108304.
- Geiger, A., Lenz, P., Urtasun, R., 2012. Are we ready for autonomous driving? The KITTI vision benchmark suite. In: *Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, pp. 3354–3361. [http://dx.doi.org/10.1109/cvpr.2012.6248074](https://doi.org/10.1109/cvpr.2012.6248074).
- Gené-Mola, J., Ferrer-Ferrer, M., Gregorio, E., Blok, P.M., Hemming, J., Morros, J.-R., Rosell-Polo, J.R., Vilaplana, V., Ruiz-Hidalgo, J., 2023. Looking behind occlusions: A study on amodal segmentation for robust on-tree apple fruit size estimation. *Comput. Electron. Agric.* 209, 107854. [http://dx.doi.org/10.1016/j.compag.2023.107854](https://doi.org/10.1016/j.compag.2023.107854).
- Gené-Mola, J., Sanz-Cortiella, R., Rosell-Polo, J.R., Escolà, 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. [http://dx.doi.org/10.1016/j.compag.2021.106343](https://doi.org/10.1016/j.compag.2021.106343).
- Hendricks, L., 1996. Orchard planning, design, and development. *Almond Prod. Man.* 47–51.
- Huang, Y., Ren, Z., Li, D., Liu, X., 2020. Phenotypic techniques and applications in fruit trees: a review. *Plant Methods* 16 (1), [http://dx.doi.org/10.1186/s13007-020-00649-7](https://doi.org/10.1186/s13007-020-00649-7).
- Jose, A.I., Pan, S., Zaenker, T., Menon, R., Houben, S., Bennewitz, M., 2025. GO-VMP: Global optimization for view motion planning in fruit mapping. [http://dx.doi.org/10.48550/arXiv.2503.03912](https://doi.org/10.48550/arXiv.2503.03912), URL [arXiv:2503.03912](https://arxiv.org/abs/2503.03912).
- Li, L., Mu, X., Soma, M., Wan, P., Qi, J., Hu, R., Zhang, W., Tong, Y., Yan, G., 2021. An iterative-mode scan design of terrestrial laser scanning in forests for minimizing occlusion effects. *IEEE Trans. Geosci. Remote Sens.* 59 (4), 3547–3566. [http://dx.doi.org/10.1109/tgrs.2020.3018643](https://doi.org/10.1109/tgrs.2020.3018643).
- Lingua, A., Noardo, F., Spanò, A., Sanna, S., Matrone, F., 2017. 3D model generation using oblique images acquired by UAV. *Int. Arch. Photogramm. Remote. Sens. Spat. Inf. Sci.* XLII-4/W2, 107–115. [http://dx.doi.org/10.5194/isprs-archives-xlii-4-w2-107-2017](https://doi.org/10.5194/isprs-archives-xlii-4-w2-107-2017).

- López-Granados, F., Torres-Sánchez, J., Jiménez-Brenes, F.M., Arquero, O., Lovera, M., de Castro, A.I., 2019. An efficient RGB-UAV-based platform for field almond tree phenotyping: 3-D architecture and flowering traits. *Plant Methods* 15 (1), <http://dx.doi.org/10.1186/s13007-019-0547-0>.
- Lorensen, W.E., Cline, H.E., 1998. Marching cubes: a high resolution 3D surface construction algorithm. In: *Seminal Graphics: Pioneering Efforts that Shaped the Field*. ACM, pp. 347–353. <http://dx.doi.org/10.1145/280811.281026>.
- Maes, W.H., Steppe, K., 2019. Perspectives for remote sensing with unmanned aerial vehicles in precision agriculture. *Trends Plant Sci.* 24 (2), 152–164. <http://dx.doi.org/10.1016/j.tplants.2018.11.007>.
- Magistri, F., Pan, Y., Bartels, J., Behley, J., Stachniss, C., Lehnert, C., 2024. Improving robotic fruit harvesting within cluttered environments through 3D shape completion. *IEEE Robot. Autom. Lett.* 9 (8), 7357–7364. <http://dx.doi.org/10.1109/ira.2024.3421788>.
- Marangoz, S., Zaenker, T., Menon, R., Bennewitz, M., 2022. Fruit mapping with shape completion for autonomous crop monitoring. In: *Proceedings of the 18th IEEE International Conference on Automation Science and Engineering*. IEEE, pp. 471–476. <http://dx.doi.org/10.1109/case49997.2022.9926466>.
- Mason-D'Croz, D., Bogard, J.R., Sulser, T.B., Cenacchi, N., Dunston, S., Herrero, M., Wiebe, K., 2019. Gaps between fruit and vegetable production, demand, and recommended consumption at global and national levels: an integrated modelling study. *Lancet Planet. Health* 3 (7), e318–e329. [http://dx.doi.org/10.1016/s2542-5196\(19\)30095-6](http://dx.doi.org/10.1016/s2542-5196(19)30095-6).
- Medic, T., Bömer, J., Paulus, S., 2023. Challenges and recommendations for 3D plant phenotyping in agriculture using terrestrial lasers scanners. *ISPRS Ann. Photogramm. Remote. Sens. Spat. Inf. Sci.* X-1/W1-2023, 1007–1014. <http://dx.doi.org/10.5194/isprs-annals-x-1-w1-2023-1007-2023>.
- Menon, R., Zaenker, T., Dengler, N., Bennewitz, M., 2023. NBV-SC: Next best view planning based on shape completion for fruit mapping and reconstruction. In: *Proceedings of the 2023 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, <http://dx.doi.org/10.1109/iros55552.2023.10341855>.
- Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R., 2021. NeRF: representing scenes as neural radiance fields for view synthesis. *Commun. ACM* 65 (1), 99–106. <http://dx.doi.org/10.1145/3503250>.
- Onyekwelu, J.C., Olusola, J.A., Stimm, B., Mosandl, R., Agbelade, A.D., 2014. Farm-level tree growth characteristics, fruit phenotypic variation and market potential assessment of three socio-economically important forest fruit tree species. *For. Trees Livelihoods* 24 (1), 27–42. <http://dx.doi.org/10.1080/14728028.2014.942386>.
- Pan, Y., Magistri, F., Läbe, T., Marks, E., Smitt, C., McCool, C., Behley, J., Stachniss, C., 2023. Panoptic mapping with fruit completion and pose estimation for horticultural robots. In: *Proceedings of the 2023 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, <http://dx.doi.org/10.1109/iros55552.2023.10342067>.
- Park, J.J., Florence, P., Straub, J., Newcombe, R., Lovegrove, S., 2019. DeepSDF: Learning continuous signed distance functions for shape representation. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 165–174.
- Ravi, N., Gabeur, V., Hu, Y.-T., Hu, R., Ryali, C., Ma, T., Khedr, H., Rädle, R., Rolland, C., Gustafson, L., Mintun, E., Pan, J., Alwala, K.V., Carion, N., Wu, C.-Y., Girshick, R., Dollár, P., Feichtenhofer, C., 2024. SAM 2: Segment anything in images and videos. <http://dx.doi.org/10.48550/arXiv.2408.00714>, URL [arXiv:2408.00714](https://arxiv.org/abs/2408.00714).
- Ren, T., Liu, S., Zeng, A., Lin, J., Li, K., Cao, H., Chen, J., Huang, X., Chen, Y., Yan, F., Zeng, Z., Zhang, H., Li, F., Yang, J., Li, H., Jiang, Q., Zhang, L., 2024. Grounded SAM: Assembling open-world models for diverse visual tasks. <http://dx.doi.org/10.48550/arXiv.2401.14159>, URL [arXiv:2401.14159](https://arxiv.org/abs/2401.14159).
- Reynolds, D., Baret, F., Welcker, C., Bostrom, A., Ball, J., Cellini, F., Lorence, A., Chawade, A., Khafif, M., Noshita, K., Mueller-Linow, M., Zhou, J., Tardieu, F., 2019. What is cost-efficient phenotyping? Optimizing costs for different scenarios. *Plant Sci.* 282, 14–22. <http://dx.doi.org/10.1016/j.plantsci.2018.06.015>.
- Schonberger, J.L., Frahm, J.-M., 2016. Structure-from-motion revisited. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.
- Tripodi, P., Massa, D., Venezia, A., Cardi, T., 2018. Sensing technologies for precision phenotyping in vegetable crops: Current status and future challenges. *Agronomy* 8 (4), 57. <http://dx.doi.org/10.3390/agronomy8040057>.
- Voigtlaender, P., Krause, M., Osep, A., Luiten, J., Sekar, B.B.G., Geiger, A., Leibe, B., 2019. MOTs: Multi-object tracking and segmentation. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 7942–7951.
- Wang, K., Kooistra, L., Wang, Y., Vélez, S., Wang, W., Valente, J., 2024. Benchmarking of monocular camera UAV-based localization and mapping methods in vineyards. *Comput. Electron. Agric.* 227, 109661. <http://dx.doi.org/10.1016/j.compag.2024.109661>.
- Xu, Z., Zhang, W., Tan, X., Yang, W., Huang, H., Wen, S., Ding, E., Huang, L., 2020. Segment as points for efficient online multi-object tracking and segmentation. In: *Proceedings of the 16th European Conference of Computer Vision*. pp. 264–281. http://dx.doi.org/10.1007/978-3-030-58452-8_16.
- Yao, S., Pan, S., Bennewitz, M., Hauser, K., 2025. Safe leaf manipulation for accurate shape and pose estimation of occluded fruits. In: *Proc. of the IEEE Intl. Conf. on Robotics & Automation*. ICRA, <http://dx.doi.org/10.48550/arXiv.2409.17389>, URL <https://arxiv.org/abs/2409.17389>.
- Yu, F., Chen, H., Wang, X., Xian, W., Chen, Y., Liu, F., Madhavan, V., Darrell, T., 2020. BDD100K: A diverse driving dataset for heterogeneous multitask learning. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.
- Yu, S., Liu, X., Tan, Q., Wang, Z., Zhang, B., 2024. Sensors, systems and algorithms of 3D reconstruction for smart agriculture and precision farming: A review. *Comput. Electron. Agric.* 224, 109229. <http://dx.doi.org/10.1016/j.compag.2024.109229>.
- Zhang, W., Peng, X., Bai, T., Wang, H., Takata, D., Guo, W., 2024. A UAV-based single-lens stereoscopic photography method for phenotyping the architecture traits of orchard trees. *Remote. Sens.* 16 (9), 1570. <http://dx.doi.org/10.3390/rs16091570>.
- Zhang, J., Xu, S., Zhao, Y., Sun, J., Xu, S., Zhang, X., 2023. Aerial orthoimage generation for UAV remote sensing: Review. *Inf. Fusion* 89, 91–120. <http://dx.doi.org/10.1016/j.inffus.2022.08.007>.
- Zhu, Z., Stamatopoulos, C., Fraser, C.S., 2015. Accurate and occlusion-robust multi-view stereo. *ISPRS J. Photogramm. Remote Sens.* 109, 47–61. <http://dx.doi.org/10.1016/j.isprsjprs.2015.08.008>.