UAV-assisted Multi-Object Tracking and Segmentation of Apples under Occlusions in Orchard Settings

Kaiwen Wang[1,*], Lammert Kooistra[2], Wensheng Wang[3], João Valente[4]

[1] Information Technology Group, Wageningen University & Research, 6706 KN, The Netherlands - kaiwen.wang@wur.nl
[2] Laboratory of Geo-Information Science and Remote Sensing,
Wageningen University & Research, 6708 PB, The Netherlands - lammert.kooistra@wur.nl
[3] Agricultural Information Institute, Chinese Academy of Agriculture Sciences, Beijing, 10086, China – wangwensheng@caas.cn
[4] Centre for Automation and Robotics (CAR), National Research Council (CSIC), Madrid, 28500, Spain – joao.valente@csic.es

Keywords: MOTS, Precision Agriculture, UAV, RGB imagery, Woody Crop.

Abstract

In precision agriculture, orchards present unique challenges for automated monitoring due to the dense foliage, complex tree structures, and frequent occlusions caused by branches, leaves, and overlapping fruits. To address these challenges, multi-object tracking and segmentation (MOTS) has been explored in general computer vision domains, aiming to simultaneously track and segment instance-level objects and maintain consistent identities across video frames. However, most existing studies focus on object-level detection without considering the temporal continuity and spatial consistency required for robust fruit monitoring over time. In this work, we implement one of the state-of-the-art MOTS methods, Grounded-SAM2, in an orchardian environment for tracking apples. In addition, four different UAV flight modes were conducted to explore the optimal solution for UAV-assisted MOTS. Our proposed evaluation framework, which relies on spatio-temporal consistency metrics and instance association heuristics, enabled the assessment of tracking performance without prior annotations.

1. Introduction

Precision horticulture increasingly relies on high-throughput and automated monitoring tools to assess plant and fruit conditions in the field. UAVs, as one of the efficient and significant tools, offer flexible, non-invasive, and scalable solutions in precision agriculture. Among various horticultural applications, orchards present unique challenges for automated monitoring due to the dense foliage, complex tree structures, and frequent occlusions caused by branches, leaves, and overlapping fruits (Wang et al., 2024). Thus, accurate detection and tracking of individual fruit instances across continuous spatial and temporal conditions from aerial imagery are essential for tasks such as yield estimation, growth assessment, quality assessment, and robotic harvesting (He et al., 2022).

Recent advances in computer vision, particularly in object detection and instance segmentation, have shown promising results in agricultural results. However, most existing studies focus on object-level detection without considering the temporal continuity and spatial consistency required for robust fruit monitoring over time. In dynamic orchard environments, occlusions and viewpoint variations further complicate the accurate association of fruit instances across consecutive frames, limiting the reliability of existing detection methods.

To address these challenges, multi-object tracking and segmentation (MOTS) have been explored in general computer vision domains, aiming to simultaneously track and segment instance-level objects and maintain consistent identities across video frames (Voigtlaender et al., 2019a). Nevertheless, the direct application of these methods to UAV-based orchard scenarios remains underexplored, primarily due to the unique characteristics of orchard environments, such as irregular object distribution, heavy occlusions, and non-uniform lighting conditions. Furthermore, the integration of aerial data acquisition with MOTS algorithms introduces additional complexities, including motion blur, perspective distortions, and variations in flight altitude and speed. Therefore, data acquisition with UAVs in the complex and clustered orchard environment requires an optimal flight mode for better performance. Moreover, MOTS evaluation with MOTSA (multiple object tracking and segmentation accuracy), sMOTSA (soft MOTSA), and MOTSP (multiple object tracking and segmentation precision) requires a large number of annotations, which is labor-intensive and time-consuming (Voigtlaender et al., 2019a).

In this work, we implement one of the state-of-the-art MOTS methods, Grounded-SAM2, in an orchard environment for tracking apples. In addition, four different UAV flight modes were conducted to explore the optimal solution for UAV-based MOTS in orchards. Our main contributions are (1) evaluation of the different flight modes for UAV-assisted MOTS, and (2) proposing a novel method for MOTS evaluation without prior annotations.

2. Study area and data collection

2.1 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 illustrated in Fig. 1. The study area of 0.083 ha, contains 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.

* Corresponding author

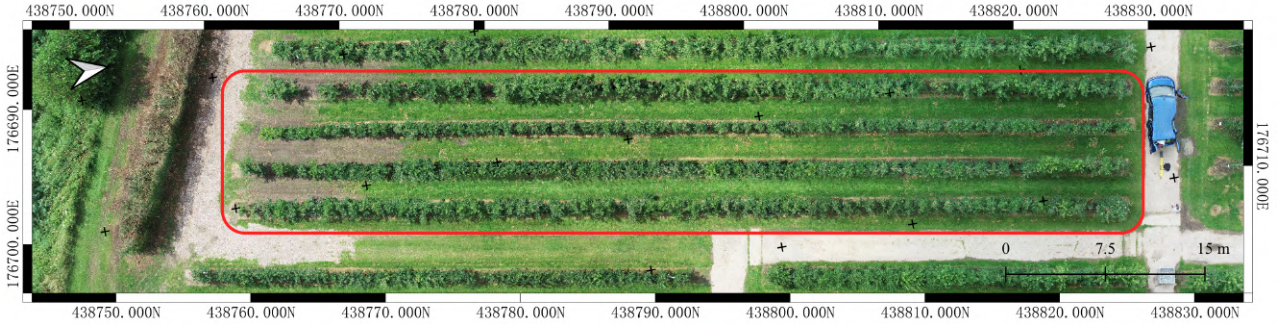

Figure 1. The study area of the apple orchard in Randwijk, Overbetuwe, Gelderland in the Netherlands. The red rectangle represents the selected region in the orchard.

2.2 UAV experimental setup

A commercial UAV equipped with a single high-resolution RGB sensor was used for video data collection in the apple orchard (Tab. 1). To find the optimal flight mode, four different UAV flight modes were implemented during data collection (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 to the side of the crops. Flight mode B followed an up-down trajectory with a closer distance to the crops compared to flight A. 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.

2.3 Data collection and preparation

In the apple orchards, we collected videos with a UAV as mentioned above. In addition, we collected the height of the wooden poles of the orchard, which support the growth of the apple trees and the structure of the rows. The height of the wooden poles was 2.7 meters, which were used to recover the scales during the reconstruction.

To evaluate the performance of MOTS in the apple orchards, we also annotated several frames in different flight modes. For flight mode A, 1890 apple instances were labeled across 10 sequential frames. For flight mode B, 756 apple instances were labeled within 10 frames. For flight mode C, 470 apple instances were labeled across 10 frames. And 1015 apple instances were annotated across 10 frames for flight mode D.

3. Method

The overall framework of this study contains Grounded-SAM2 for apple MOTS and structure-from-motion (SfM) for camera poses and depth estimation, as shown in Fig. 3.

3.1 MOTS with Grounded-SAM2

We implemented Grounded-SAM2 for the apple tracking and segmentation in the four flight modes (Ravi et al., 2024, Ren et al., 2024). To understand the preliminary MOTS results, we evaluated the MOTS performance throuth three evaluation metrics (Voigtlaender et al., 2019b), as shown in Eqs. 1, 2, 3, 4, and 5.

$$c(h) = \begin{cases} \arg\max_{m \in M} \text{IoU}(h, m), & \text{if } \max_{m \in M} \text{IoU}(h, m) > 0.5 \\ \emptyset, & \text{otherwise.} \end{cases} \quad (1)$$

$$\widetilde{TP} = \sum_{h \in TP} \text{IoU}(h, c(h)) \quad (2)$$

$$MOTSA = \frac{|TP| - |FP| - |IDS|}{|N|} \quad (3)$$

$$sMOTSA = \frac{\widetilde{TP} - |FP| - |IDS|}{|N|} \quad (4)$$

$$MOTSP = \frac{\widetilde{TP}}{|TP|} \quad (5)$$

where $M = m_1, ..., m_N$ with $m_i \in 0, 1$ are the ground truth pixel masks, $H = h_1, ..., h_K$ 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.

3.2 Camera poses and depth estimation with SfM

We used Agisoft Metashape Pro[1] (Agisoft LLC, St. Petersburg, Russia) to estimate the UAV camera poses and accurate dense depth information. We imported the raw UAV frames at 3 FPS into Agisoft and set a pair of marker points at the top and bottom of the poles as a scale reference. The distance between two points was set to 2.7 m. Then all frames were aligned with 'highest accuracy' and 'source' options under 40,000 key point limit and 4,000 tie point limit. The reprojection error of the final result is 0.691 pixels. After that, point clouds with 'ultra high' and 'mild' depth filtering were built. A depth export script[2] was implemented to get the highly accurate depth maps from SfM. Finally, the depth and camera poses with metric scale can be derived.

3.3 MOTS evaluation without prior-annotations

We propose a method to evaluate the MOTS performance without annotations. The basic principle of this method is the comparison of the speed between camera movement and instance movement. Therefore, we first extract camera poses from SfM (Structure from Motion). Then, we split the whole frame number into N groups, each group containing 5 consecutive frames. And we randomly sample 10 groups, which include 10 apple instances for the first frame in each group. Then we track each instance's centroid over frames using segmentation masks from Grounded-SAM2 and depth maps from SfM. Compute 3D instance displacements $v_{instance,j}$ using depth at centroids and camera intrinsics.

[1] https://www.agisoft.com/
[2] https://github.com/agisoft-llc/metashape-scripts

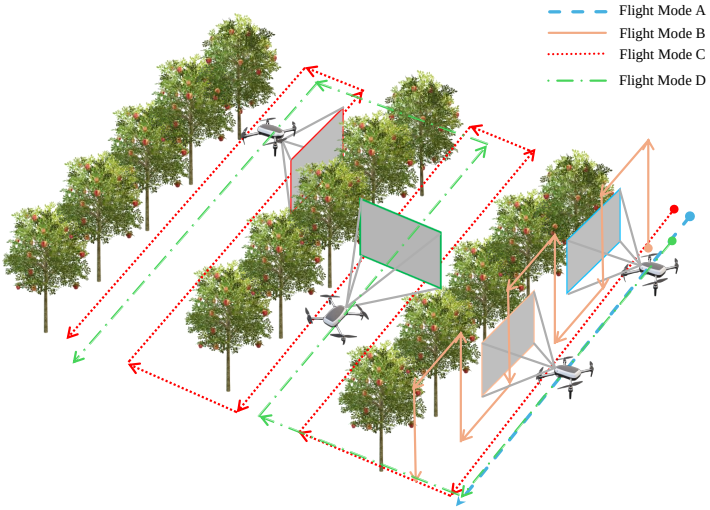

Figure 2. Four flight modes in orchards. Four different colors and line styles indicate four flight modes. The colors of the camera views are also in line with the colors of four flight modes.

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

4. Result

4.1 MOTS performance in four flight modes

Table 2 presents the evaluation results of apple detection and tracking performance using Grounding-SAM2 across four different UAV flight modes, assessed using the MOTS metrics defined in Eqs. 1, 2, 3 4, and 5. Among the four modes, Flight Mode A achieved the highest MOTSA score of 34.84%, indicating superior overall segmentation and tracking performance. It also demonstrated the best sMOTSA score (14.95%), highlighting its effectiveness in handling segmentation accuracy while penalizing ID switches and false positives. Flight Mode D outperformed others in terms of MOTSP (73.29%), signifying higher precision in object segmentation and tracking. However, instead of tens of apples, which can be detected in flight modes A, B, and C, only seven apples were detected in Flight Mode D (Fig. 4). Because the limited detected apple instances have more obvious features, which leads to better detection and tracking performance. Conversely, flight mode C exhibited the poorest performance across all metrics, with negative MOTSA (-40.18%) and sMOTSA (-57.05%) values, suggesting significant challenges in maintaining accurate object tracking and segmentation.

We also visualized the MOTS results in Fig. 4. In flight mode A and B, most of the apple instances were detected, as shown in Figs. 4(a) and 4(b). In flight mode C, more than half of the apple instances were not detected by Grounded-SAM2 (Fig. 4(c)). Flight mode D performed the worst, which almost cannot detect any apple instances (Fig. 4(d)).

Table 2. Apple detection and tracking performance with Grounding-SAM2 under four different flight modes. The bold numbers indicate the best results.

Flight Mode	MOTSA (%)↑	sMOTSA (%)↑	MOTSP (%)↑
A	**34.84**	**14.95**	71.98
B	2.72	-1.20	72.13
C	-40.18	-57.05	66.96
D	1.99	1.25	**73.29**

4.2 MOTS performance without prior-annotations

The bubble chart (Fig. 5) illustrates the speed ratios between instance movement and camera movement across sequential frames by our random sampling methods in flight mode A. Most of the speed ratio values remain close to 1.0, indicating that most detected 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 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.

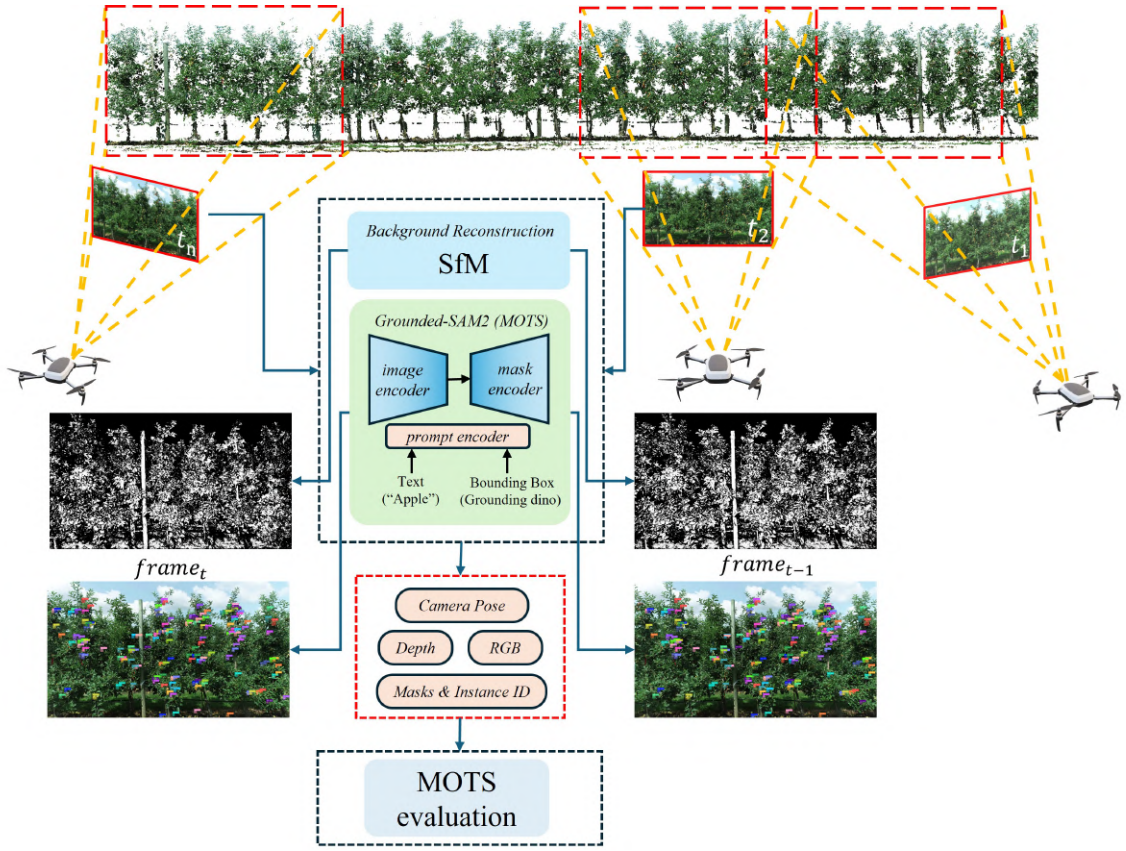

Figure 3. The overall framework of the data collection, MOTS implementation, and evaluation. The top image shows the reconstruction results of one apple row.

(a) Flight Mode A.

(b) Flght Mode B.

(c) Flight Mode C.

(d) Flight Mode D.

Figure 4. MOTS examples of four different flight.

We also evaluated three other flight modes with our evaluation method, but all of them provided a None value due to the low detection rate and high ID switches during the MOTS (Tab. 2 and Fig. 4). Thus, flight mode A would be an optimal solution

Algorithm 1 MOTS Performance Evaluation Without Annotations

Require: Total frames N, Masks $\mathcal{M}$, Depth $\mathcal{D}$, Camera poses $\mathcal{P}$, Intrinsic matrix K
Ensure: MOTS Performance Score
Extract camera positions $\mathcal{P}$ from XML file
Select 10 random frame windows $S = \{S_1, S_2, \ldots, S_{10}\}$, where each window compose by 5 sequential frames
Initialize empty list $\mathcal{R}$ for speed ratios
for each part $S_i \in S$ **do**
    Randomly select 10 apple instances $\mathcal{I}$ in the first frame of S_i
    Initialize empty lists $\mathcal{V}_{\text{instance}}$, $\mathcal{V}_{\text{camera}}$
    for each consecutive frame pair $(t, t+1) \in S_i$ **do**
        Load mask centroids $\mathcal{C}_t$ and $\mathcal{C}_{t+1}$ from $\mathcal{M}$
        Load depth maps Z_t and Z_{t+1} from $\mathcal{D}$
        Extract depth values $\mathcal{Z}_t$ and $\mathcal{Z}_{t+1}$ at centroids
        Compute **apple instance speeds**:

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

        Compute **camera movement speed**:

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

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

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

    Append R to $\mathcal{R}$
end for
Compute final MOTS performance score:

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

Return MOTS Score

to conduct MOTS in apple orchards with UAVs.

5. Discussion

5.1 Effectiveness of Grounded-SAM2 in orchard environment

The experimental results demonstrate that Grounded-SAM2 can detect and track apple instances across consecutive aerial frames under some specific conditions. Compared to previous CNN-based MOTS methods, such as PointTrack (Xu et al., 2020) and Track-RCNN (Voigtlaender et al., 2019b), which required a large amount of time for training and cannot easily adapt to other domains, Grounded-SAM2 can be implemented under several text prompts without any pre-training.

However, occlusions from the orchards, dynamic illumination conditions, and the dense homogeneous apple instances make it challenging to get a good MOTS performance (de Jong et al., 2022). In addition, the foregrounds (target objects) and backgrounds from the practical implementation of conventional MOTS methods have a large difference in motion patterns. Normally, the camera perspective and movement are fixed or slightly moving, which makes it easy to distinguish foregrounds and backgrounds. But when it comes to agricultural domains, the target objects, such as crops and fruits, move with the backgrounds in similar motion patterns, which makes it difficult. Thus, the intregration of autonomous inspection with path planning methods and MOTS solutions would be a feasible direction to address occlusion issues.

5.2 UAV flight modes impact

Our comparative analysis across different UAV flight modes revealed notable variations in tracking stability and segmentation consistency (Tab. 2). Specifically, the flight mode A with a straight line in a side perspective provided more consistent detection and tracking results (Figs. 5 and 4(a)). But for flight modes C and D, they obtained the worst results from Grounded-SAM2, which may be caused by the challenging camera perspectives. In flight mode C, the camera perspective leads to more occlusions by leaves for the apple instances. And in flight mode D, the higher flight altitude and slightly downward camera angle result in higher image contrast, which makes it hard to recognize the apple instances for the models. Therefore, the stable illumination conditions and simple flight mode are suitable for robust tracking performance in orchard settings. Ideally, combination of the four flight modes would optimize the visible areas to region of interest in the whole orchards.

5.3 MOTS annotations and evaluation

Currently, most of the MOTS algorithms were developed by large public datasets such as BDD100K (Yu et al., 2020) and KITTI MOTS (Voigtlaender et al., 2019b). Thus, the generalization of the pre-trained models faces challenges when it is implemented into other domains. The first challenge is the data acquisition and annotations, which is a labor-intensive and time-consuming task. The second challenge is, can current models adapt to the new domains with new datasets?

Our proposed evaluation framework, which relies on spatio-temporal consistency metrics and apple instances association heuristics, enabled the assessment of tracking and segmentation performance without requiring prior manual annotations. This approach is particularly valuable in large orchard deployments where annotating dense video data is prohibitively labor-intensive. While the method showed strong alignment with qualitative observations (Fig. 4), it is sensitive to initial detection quality and may overestimate tracking performance in sequences with static or slowly moving objects. Future work could integrate synthetic data augmentation or domain adaptation techniques to further calibrate the evaluation metrics.

6. Conclusion

In this study, we designed four UAV flight modes to explore the optimal solution for MOTS in apple orchards. Meanwhile, we implemented one of the state-of-the-art MOTS algorithms in orchard environments. Flight mode A, with a simple straight-line flight, obtained the best performance with our manual annotations (34.84% MOTSA, 14.95% sMOTSA, and 71.98% MOTSP). Due to the workload of data annotations during the MOTS evaluation, we also proposed a novel algorithm to evaluate MOTS performance without prior annotations.

References

de Jong, S., Baja, H., Tamminga, K., Valente, J., 2022. Apple mots: Detection, segmentation and tracking of homogeneous objects using mots. *IEEE Robotics and Automation Letters*, 7(4), 11418–11425.

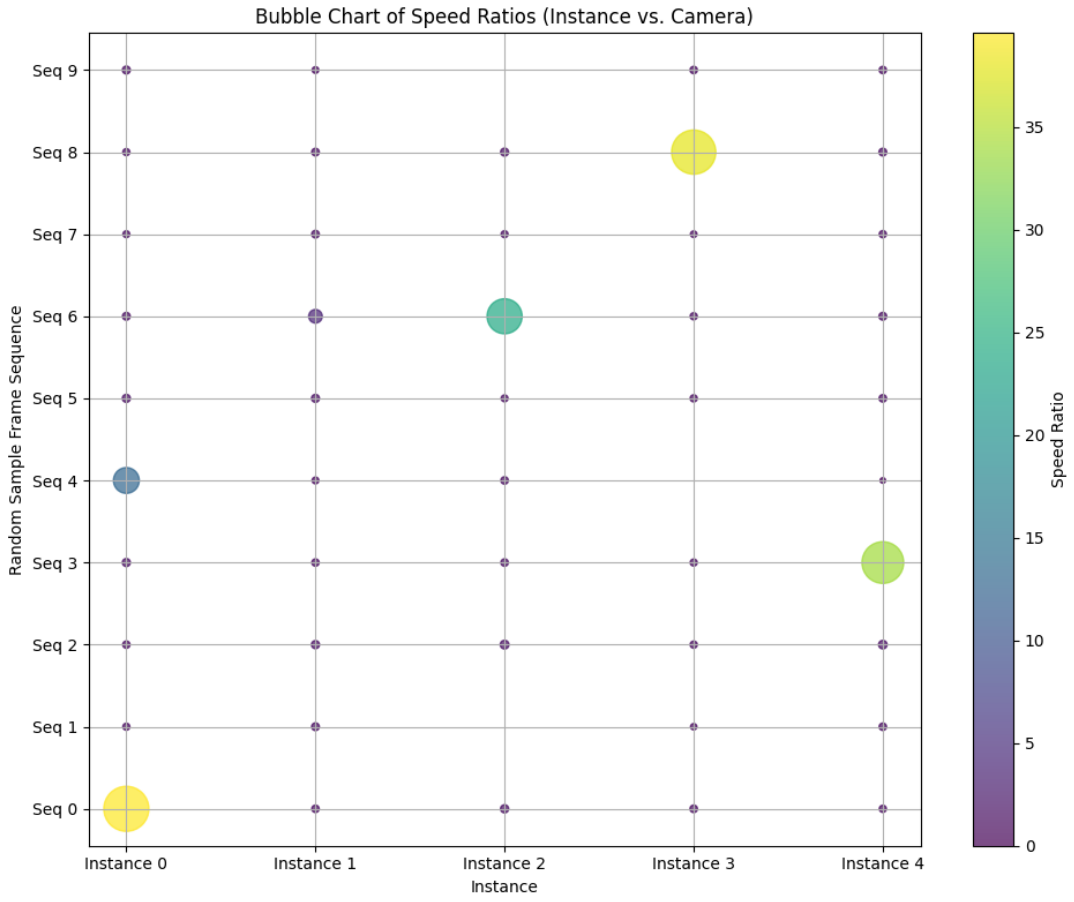

Figure 5. The bubble chart of Speed Ratios (Instance Speed/Camera Speed)

He, L., Fang, W., Zhao, G., Wu, Z., Fu, L., Li, R., Majeed, Y., Dhupia, J., 2022. Fruit yield prediction and estimation in orchards: A state-of-the-art comprehensive review for both direct and indirect methods. *Computers and Electronics in Agriculture*, 195, 106812. http://dx.doi.org/10.1016/j.compag.2022.106812.

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.

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.

Voigtlaender, P., Krause, M., Osep, A., Luiten, J., Sekar, B. B. G., Geiger, A., Leibe, B., 2019a. Mots: Multi-object tracking and segmentation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Voigtlaender, P., Krause, M., Osep, A., Luiten, J., Sekar, B. B. G., Geiger, A., Leibe, B., 2019b. MOTS: Multi-Object Tracking and Segmentation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 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. *Computers and Electronics in Agriculture*, 227, 109661. http://dx.doi.org/10.1016/j.compag.2024.109661.

Xu, Z., Zhang, W., Tan, X., Yang, W., Su, X., Yuan, Y., Zhang, H., Wen, S., Ding, E., Huang, L., 2020. Pointtrack++ for effective online multi-object tracking and segmentation.

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.