

Article

Occlusion-Aware Multi-Object Tracking in Vineyards via SAM-Based Visibility Modeling

Yanan Wang ^{1,†}, Hagsong Kim ^{1,†}, Muhammad Fayaz ¹ , Lien Minh Dang ², Hyeonjoon Moon ¹ 
and Kang-Won Lee ^{1,*} 

¹ Department of Computer Science and Engineering, Sejong University, Seoul 05006, Republic of Korea; muhammadfayaz@sju.ac.kr (M.F.)

² Department of Information and Communication Engineering and Convergence Engineering for Intelligent Drone, Sejong University, Seoul 05006, Republic of Korea

* Correspondence: kangwon.lee@sejong.ac.kr

[†] These authors contributed equally to this work.

Abstract

Multi-object tracking (MOT) in vineyard environments remains challenging due to frequent and long-term occlusions caused by dense foliage, overlapping grape clusters, and complex plant structures. These characteristics often result in identity switches and fragmented trajectories when using conventional tracking methods. This paper proposes OATSAM-Track, an occlusion-aware multi-object tracking framework designed for vineyard fruit monitoring. The framework integrates lightweight MobileSAM-assisted instance segmentation to estimate target visibility and occlusion severity. Occlusion-state reasoning is further incorporated into temporal association, appearance memory updating, and identity recovery. An adaptive temporal memory mechanism selectively updates appearance features according to predicted occlusion states, reducing identity drift under partial and severe occlusions. To facilitate occlusion-aware evaluation, an extended vineyard multi-object tracking dataset (GrapeOcclusionMOTS) with SAM-refined instance masks and fine-grained occlusion annotations is constructed. The experimental results demonstrate that OATSAM-Track improves identity consistency and tracking robustness compared to representative baseline trackers, particularly under medium and severe occlusion scenarios. These results indicate that explicit occlusion modeling is beneficial for reliable fruit monitoring in precision agriculture.

Keywords: multi-object tracking; occlusion modeling; vineyard monitoring; Segment Anything Model; temporal memory; precision agriculture



Academic Editor: Dah-Jye Lee

Received: 9 January 2026

Revised: 28 January 2026

Accepted: 30 January 2026

Published: 1 February 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and

conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

Robust multi-object tracking (MOT) is essential for precision agriculture, enabling tasks such as fruit counting, yield estimation, and growth monitoring in orchards [1,2]. In vineyard environments, MOT faces unique challenges: fruits are largely stationary, visually similar, and frequently occluded by dense foliage and overlapping grape clusters. These conditions violate common MOT assumptions based on motion continuity and appearance distinctiveness, leading to limited motion cues, severe appearance ambiguity, long-term visibility fluctuations, and frequent identity switches.

Most existing multi-object tracking methods adopt a tracking-by-detection paradigm, relying on motion and appearance cues for association. Methods such as DeepSORT [3], ByteTrack [4], and StrongSORT [5] perform well in generic benchmarks, but operate at the

bounding-box level and assume stable appearance, making them vulnerable to severe and long-term occlusions in agricultural environments. Although occlusion-aware tracking strategies have been explored [6–9], most are designed for urban or aerial scenarios and lack fine-grained visibility modeling. Recent segmentation foundation models such as SAM [10] offer new opportunities for pixel-level visibility reasoning, yet remain underutilized in agricultural multi-object tracking.

To address these challenges, we propose OATSAM-Track, an occlusion-aware MOT framework tailored for vineyard scenarios. The method leverages SAM-generated instance masks to model target visibility, integrates occlusion reasoning into temporal association and identity management, and introduces an adaptive temporal memory mechanism that selectively updates appearance features based on predicted occlusion states. A robust re-identification recovery strategy further restores lost tracks while minimizing false identity assignments. To support systematic evaluation, we constructed an extended vineyard tracking benchmark with SAM-refined instance masks and fine-grained occlusion annotations. Extensive experiments demonstrate that OATSAM-Track consistently outperforms state-of-the-art trackers under both standard and occlusion-specific evaluation metrics, especially in medium and severe occlusion scenarios.

The main contributions of this work are summarized as follows:

- We introduce OATSAM-Track, an occlusion-aware MOT framework that explicitly integrates SAM-based instance segmentation into temporal association and identity management for vineyard environments.
- We designed an adaptive temporal memory module guided by predicted occlusion states, enabling selective appearance updating to maintain identity consistency under partial and severe occlusions.
- We constructed an extended vineyard MOT dataset (GrapeOcclusionMOTS) with SAM-refined instance masks and fine-grained occlusion annotations [11], and demonstrate consistent performance improvements over state-of-the-art trackers.

The remainder of this paper is organized as follows: Section 2 reviews relevant literature on occlusion-aware multi-object tracking, segmentation-based visibility reasoning, and agricultural vision applications. Section 3 presents the proposed OATSAM-Track framework in detail, including SAM-based visibility estimation, hybrid occlusion state modeling, adaptive temporal memory update, and occlusion-aware re-identification recovery. Section 4 describes the experimental setup, dataset details, evaluation metrics, quantitative and qualitative results, and ablation studies analyzing the contribution of each module. Section 5 discusses the effectiveness, practical implications, and limitations of the proposed occlusion-aware tracking framework. Finally, Section 6 concludes the paper and outlines directions for future research.

2. Related Work

2.1. Occlusion-Aware Multi-Object Tracking

Occlusion is a major challenge in multi-object tracking, often causing identity switches and fragmented trajectories. Most contemporary trackers adopt a tracking-by-detection paradigm, associating detections across frames using motion and appearance cues. Representative methods such as DeepSORT [3], ByteTrack [4], and StrongSORT [5] have shown strong performance in structured scenes.

Despite their success, these methods primarily rely on bounding-box-level association and implicit temporal smoothing, making them vulnerable to partial and prolonged occlusions where fine-grained visibility information is unavailable. Several studies have explored occlusion-aware tracking by explicitly modeling visibility or occlusion states [6–9]. For example, Possegger et al. [6] introduced explicit occlusion reasoning to improve data

association, while later works incorporated state-aware appearance modeling or online occlusion prediction to mitigate identity drift under occlusion [7–9].

More recently, occlusion-aware tracking has also been investigated in aerial and UAV-based scenarios, where viewpoint changes and frequent overlaps further exacerbate visibility uncertainty [12]. However, most existing occlusion-aware designs are tailored to pedestrian or vehicle tracking and rely on heuristic rules or domain-specific assumptions. Consequently, their generalization to complex agricultural environments where occlusions are irregular, long-lasting, and highly asymmetric remains limited.

2.2. Segment-Based Tracking and Visibility Reasoning

Segmentation-based perception provides pixel-level information that is particularly valuable for reasoning about object visibility under occlusion. Recent advances in foundation segmentation models, most notably the Segment Anything Model (SAM) [10] and Mask2Former [13], enable high-quality instance segmentation across diverse domains without task-specific training. To improve practical deployability, lightweight variants such as MobileSAM [14] and MobileSAMv2 [15] have been proposed, significantly reducing computational overhead while maintaining competitive segmentation quality.

These segmentation advances open opportunities for occlusion-aware tracking by enabling fine-grained visibility reasoning that was previously unavailable in MOT pipelines. Despite this potential, most existing trackers continue to operate at the bounding-box level, ignoring pixel-level masks for visibility estimation and occlusion-state reasoning. In agricultural scenes characterized by dense foliage and frequent object overlap, this limitation often leads to identity fragmentation and degraded tracking performance under severe occlusion.

2.3. Multi-Object Tracking in Agricultural Vision

Multi-object tracking has gained increasing attention in agricultural vision, supporting applications such as fruit counting, yield estimation, and growth monitoring. Several studies have explored detection and tracking in orchards and vineyards, addressing challenges related to illumination variation, background clutter, and small object sizes [11,16–18]. Compared with generic MOT benchmarks, agricultural datasets remain limited in scale and occlusion annotation richness, restricting evaluation of occlusion-aware strategies.

Most existing agricultural tracking methods adopt general-purpose MOT frameworks without explicit adaptation for occlusion modeling. As a result, identity consistency often degrades in dense canopy environments with frequent and prolonged occlusions. These limitations highlight the need for methods that explicitly reason about occlusion and leverage fine-grained visibility information for robust long-term tracking.

3. Methodology

3.1. Overview of OATSAM-Track

The overall framework of OATSAM-Track is illustrated in Figure 1. The proposed method follows a tracking-by-detection paradigm [19–21] and is designed to explicitly handle severe and long-term occlusions in vineyard environments.

OATSAM-Track consists of five key components: (1) a detection and tracking backbone, (2) SAM-based visibility estimation, (3) hybrid occlusion state modeling, (4) adaptive temporal memory update, and (5) occlusion-aware re-identification recovery.

Given an input video sequence, fruit detections and initial tracklets are first generated using a YOLO-based detector [22] and a StrongSORT tracker [5]. These backbone components provide temporally consistent bounding boxes and identity hypotheses. The detection backbone is inspired by widely used YOLO variants that balance accuracy and

efficiency [19–21], providing robust initial detections under dense foliage and varying illumination conditions.

To overcome the limitations of bounding-box-level reasoning under occlusion, instance masks are extracted for each target using SAM. The resulting visibility cues are used to infer occlusion states, which explicitly guide appearance updating and identity recovery. Through this design, OATSAM-Track produces identity-consistent object trajectories under long-term occlusion in dense vineyard environments.

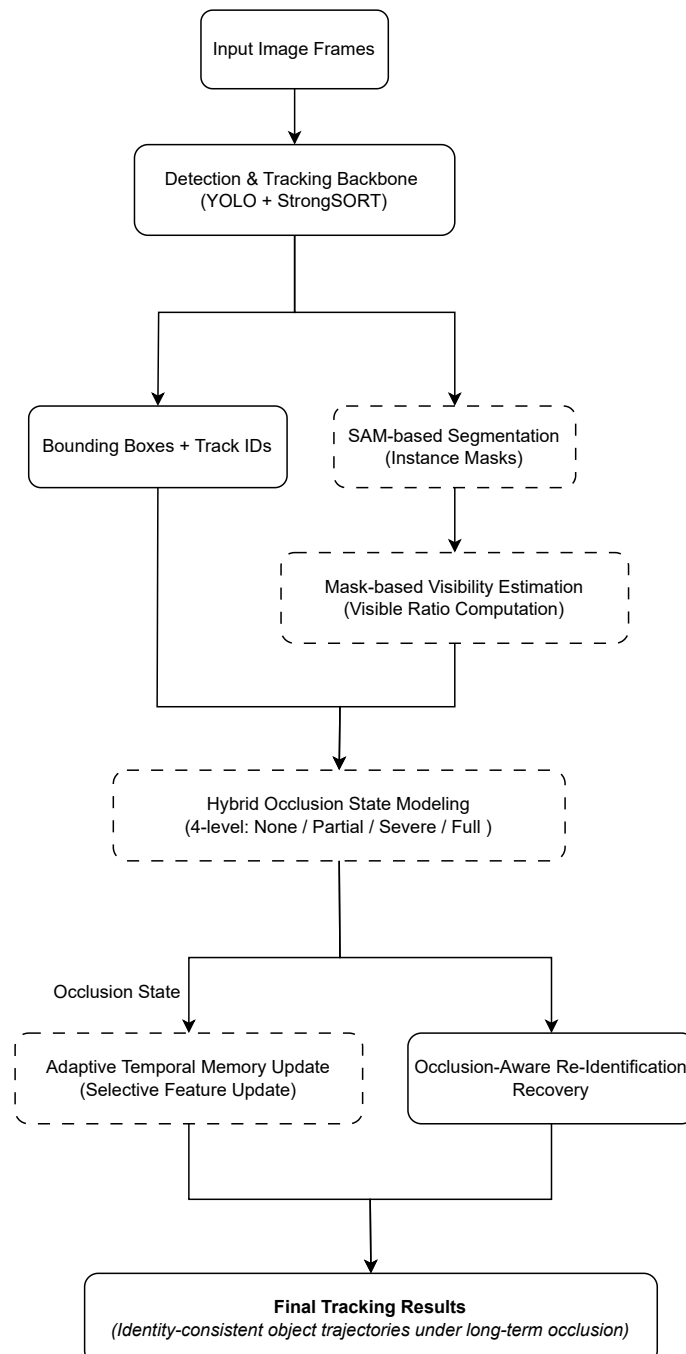


Figure 1. Overall framework of the proposed OATSAM-Track. SAM-generated instance masks enable explicit visibility estimation, which drives occlusion-aware memory update and re-identification recovery.

This occlusion-aware design is specifically tailored for vineyard fruit monitoring, where stationary targets are repeatedly occluded by foliage and structural elements in complex agricultural environments.

3.2. SAM-Based Visibility Estimation

To obtain fine-grained visibility information, OATSAM-Track leverages the Segment Anything Model (SAM) [10], specifically its lightweight MobileSAM variant [14], to generate instance-level masks for tracked objects. Compared with bounding boxes, segmentation masks provide pixel-level cues that are crucial for reasoning about partial and irregular occlusions commonly observed in vineyards.

For each tracked target i at frame t , the SAM mask is associated with the corresponding track bounding box through IoU matching, as defined in Equation (1):

$$j^* = \arg \max_j \text{IoU}(\mathbf{b}_i^t, m_j^t), \quad \text{s.t. } \text{IoU}(\mathbf{b}_i^t, m_j^t) \geq \tau_{\text{IoU}}, \quad (1)$$

where τ_{IoU} is set to 0.5 in all experiments. A local visible mask is obtained by intersecting the matched SAM mask with the corresponding bounding box region. The visibility ratio is then computed as shown in Equation (2):

$$v_i^t = \frac{|m_i^t|}{|\mathbf{b}_i^t|}, \quad (2)$$

where $|m_i^t|$ denotes the number of visible pixels and $|\mathbf{b}_i^t|$ denotes the bounding-box area. This ratio provides a compact and effective measure of occlusion severity and serves as the primary geometric cue for subsequent occlusion modeling.

Such pixel-level visibility estimation is particularly suitable for vineyard scenes. In these environments, fruit targets are frequently partially covered by leaves and branches. Moreover, bounding-box-level cues alone are insufficient for reliable occlusion reasoning in agricultural environments.

3.3. Hybrid Occlusion State Modeling

Based on the estimated visibility ratio and visual appearance cues, each target is assigned to one of four occlusion levels: no occlusion, partial occlusion, severe occlusion, and full occlusion. To improve robustness across diverse occlusion patterns, a hybrid modeling strategy is adopted.

The four occlusion levels are discretized into numerical occlusion states according to predefined visibility ratio thresholds, as summarized in Table 1. The visibility ratio thresholds are empirically determined based on qualitative observations of grape visibility patterns in vineyard scenes and validated through preliminary experiments. Specifically, higher values of v_i^t indicate unobstructed targets with reliable appearance cues, while decreasing v_i^t corresponds to increasing occlusion severity and higher identity uncertainty. Accordingly, the threshold ranges in Table 1 are designed to reflect practical visibility semantics in vineyard environments.

Table 1. Visibility ratio thresholds and corresponding occlusion definitions.

Visibility Ratio v_i^t	Occlusion State (ID)	Occlusion Level	Description
$v_i^t \geq 0.75$	0	No occlusion	Fully visible target
$0.40 \leq v_i^t < 0.75$	1	Partial occlusion	Partially occluded target
$0.05 \leq v_i^t < 0.40$	2	Severe occlusion	Heavily occluded target
$v_i^t < 0.05$	3	Full occlusion	Nearly or fully invisible target

As illustrated in Figure 2, the predicted occlusion state serves as a central control signal that regulates both appearance memory updating and identity recovery. A lightweight ResNet18-based classifier [23] is employed to predict the occlusion state based on visual

appearance cues. Given the input image region I corresponding to a tracked target, the classifier outputs a probability distribution over occlusion states, as defined in Equation (3):

$$p = \text{softmax}(f_{\theta}(I)), \quad (3)$$

where $f_{\theta}(\cdot)$ denotes the classifier with parameters θ .

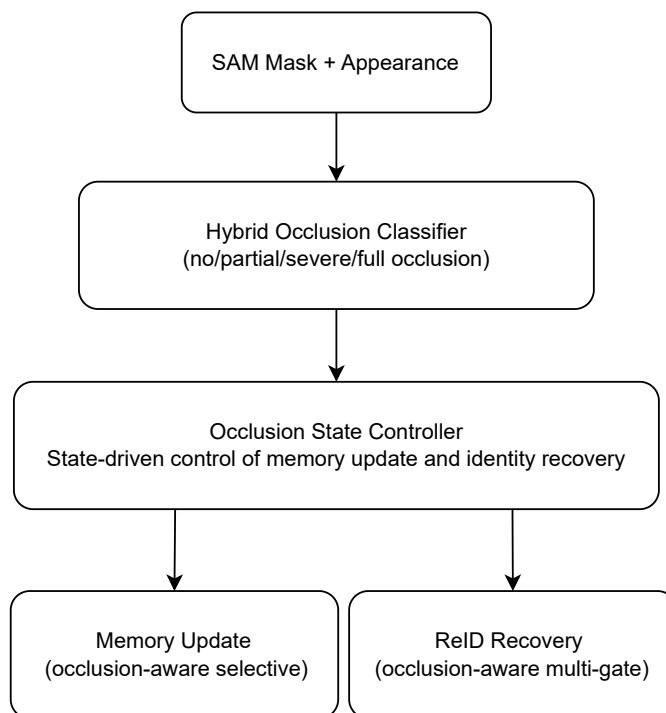


Figure 2. Occlusion-aware tracking mechanism of OATSAM-Track. The hybrid occlusion classifier integrates MobileSAM-based visibility cues and appearance information to estimate occlusion states [14]. The predicted occlusion state is further used by a state controller to selectively regulate appearance memory updates and to trigger strict multi-gate re-identification recovery.

When the classifier confidence is sufficiently high, the occlusion state is directly determined according to Equation (4):

$$\hat{s} = \arg \max_k p_k, \quad \text{if } \max_k p_k \geq \tau_c, \quad (4)$$

where τ_c is set to 0.65 in all experiments. The final occlusion state is determined by a hybrid decision rule that combines the classifier prediction and visibility-based estimation, as summarized in Equation (5):

$$s_i^t = \begin{cases} \hat{s}_{\text{ResNet}}, & \text{if } \max_k p_k \geq \tau_c, \\ s_{\text{visibility}}, & \text{otherwise.} \end{cases} \quad (5)$$

If the classifier confidence is low, the occlusion state falls back to a deterministic visibility-based estimation using predefined thresholds. This hybrid strategy combines the discriminative power of learned appearance features with the interpretability and stability of geometry-based visibility reasoning.

The state controller follows deterministic decision rules without introducing additional learnable parameters. This hybrid design combines the generalization ability of learned

features with the interpretability of geometric visibility reasoning. As a result, it enables stable occlusion estimation under cluttered and ambiguous conditions. Explicit occlusion state modeling provides interpretable and stable control signals for tracking behavior in vineyard environments. In these scenarios, fruit visibility changes are driven by structured plant geometry rather than target motion.

3.4. Adaptive Temporal Memory Update

Conventional trackers often rely on a motion model, typically implemented via a Kalman filter [24,25], to predict object positions across frames. However, continuous appearance updating without occlusion-aware control frequently leads to identity drift when targets are heavily occluded. In contrast, OATSAM-Track introduces an adaptive temporal memory mechanism that selectively updates appearance features based on the predicted occlusion state. When the adaptive memory module is disabled, the tracker degenerates to a standard appearance updating strategy that continuously replaces features without occlusion-aware gating, as adopted in StrongSORT [5].

The appearance feature of target i at frame t is extracted from the cropped image region, as defined in Equation (6):

$$\mathbf{f}_i^t = \phi(I_{\text{crop}}^{i,t}), \quad (6)$$

where $\phi(\cdot)$ denotes the appearance embedding network. Each track maintains a bounded temporal appearance memory that stores reliable features from recent frames, as summarized in Equation (7):

$$\mathcal{M}_i = \{(\mathbf{f}_i^{t_k}, t_k) \mid k = 1, \dots, K\}, \quad K \leq C, \quad (7)$$

where C denotes the maximum memory capacity.

When a target has no occlusion or only partial occlusion, its appearance representation is updated normally. In contrast, feature updates are partially suppressed or entirely frozen under severe or full occlusion. This selective update strategy preserves clean identity representations and prevents contamination from unreliable observations during occlusion. The aggregated memory representation is used for both online association and subsequent re-identification, providing stable appearance cues even after long-term occlusion.

This design is particularly important for fruit monitoring in vineyard environments, where targets typically remain stationary for long periods and are frequently re-occluded by leaves and branches. In such scenarios, unreliable appearance updates during partial or severe occlusion can permanently corrupt identity representations, making selective memory updating essential for long-term identity consistency.

3.5. Occlusion-Aware Re-Identification Recovery

To recover interrupted trajectories after prolonged occlusion, an occlusion-aware re-identification (ReID) recovery strategy is proposed. Unlike conventional recovery schemes that rely solely on appearance similarity, the proposed method incorporates occlusion state constraints and temporal consistency.

ReID recovery is activated only for targets previously classified as severely or fully occluded. This prevents premature identity reassignment during partial visibility. As illustrated in Figure 3, for each unmatched detection, candidate lost tracks are evaluated sequentially through a strict multi-stage validation process. First, a temporal proximity gate restricts recovery to recently lost tracks, ensuring temporal plausibility. Second, an appearance similarity gate compares the detection feature with the adaptive temporal memory embedding of each candidate track. This filters out visually inconsistent matches.

Third, a recency preference gate further suppresses recovery of long-lost tracks by favoring more recent identities. When applicable, geometric consistency based on motion prediction is further used as an auxiliary validation to suppress implausible recoveries.

Once a candidate passes all validation stages, the corresponding identity is recovered and initialized with a confidence score. The confidence is progressively accumulated through subsequent successful associations. The recovered track is considered stable only after reaching a predefined confidence level. By jointly considering visibility, temporal memory, and strict validation, the proposed strategy effectively balances identity recall and precision. This balance is critical in dense vineyard scenes with visually similar and stationary targets.

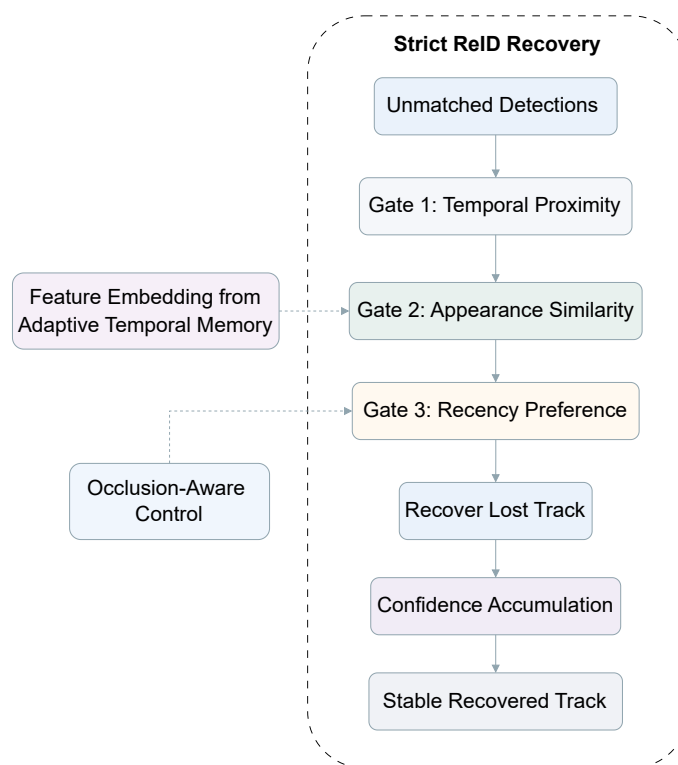


Figure 3. Occlusion-aware ReID recovery process with multi-gate validation and confidence accumulation.

4. Experimental Results

This section evaluates the effectiveness of the proposed OATSAM-Track framework under complex agricultural scenarios. We first introduce the dataset and experimental setup, followed by quantitative comparisons and an ablation study to analyze the contribution of each module.

4.1. Datasets

To evaluate the effectiveness of the proposed method under complex occlusion conditions, we constructed an extended vineyard multi-object tracking dataset, referred to as GrapeOcclusionMOTS, based on real-world vineyard video sequences. The dataset covers diverse challenging scenarios, including dense foliage, overlapping grape clusters, illumination variations, and long-term occlusions caused by leaves and branches.

Each frame is annotated with bounding boxes and unique track identities. In addition, the Segment Anything Model (SAM) [10] is employed to refine instance-level segmentation masks. It should be noted that all SAM-generated masks were manually verified and corrected by human annotators. During this process, annotators inspected each frame,

corrected object boundary errors, removed false positives, and ensured consistency of each object across frames. These masks are further used to estimate object visibility ratios and determine occlusion states, which are essential for occlusion-aware modeling and evaluation.

Following standard protocol, the dataset is split into training and test sets without scene overlap. Detailed statistics of the dataset, including the number of images, annotated instances, and their distribution across different occlusion states, are summarized in Table 2.

Table 2. Statistical description of the GrapeOcclusionMOTS dataset, including the number of images, annotated instances, and their distribution across four occlusion levels (no, partial, severe, full).

Subset	Total Images	Total Instances	No (0)	Partial (1)	Severe (2)	Full (3)
Train	927	4406	2403	1075	926	2
Test	400	5869	3960	1273	636	0
Total	1327	10,275	6363	2348	1562	2

Figure 4 shows representative samples from the GrapeOcclusionMOTS dataset. The top image is the original vineyard scene, and the bottom image shows the corresponding ground truth annotations, which are extended to include occlusion state labels for the targets.



Figure 4. Typical samples from GrapeOcclusionMOTS dataset. The original image (**top**) and the corresponding ground truth annotations (**bottom**). Different colors in the annotations indicate occlusion levels: green represents no occlusion, orange represents partial occlusion, and red represents severe occlusion.

The construction pipeline of the GrapeOcclusionMOTS extended dataset is illustrated in Figure 5. Representative detection results of the proposed OATSAM-Track on the GrapeOcclusionMOTS dataset are shown in Figure 6, covering four occlusion levels

ranging from no to full occlusion. Quantitative detection metrics are reported in Table 3, complementing the visual results shown in Figure 6.

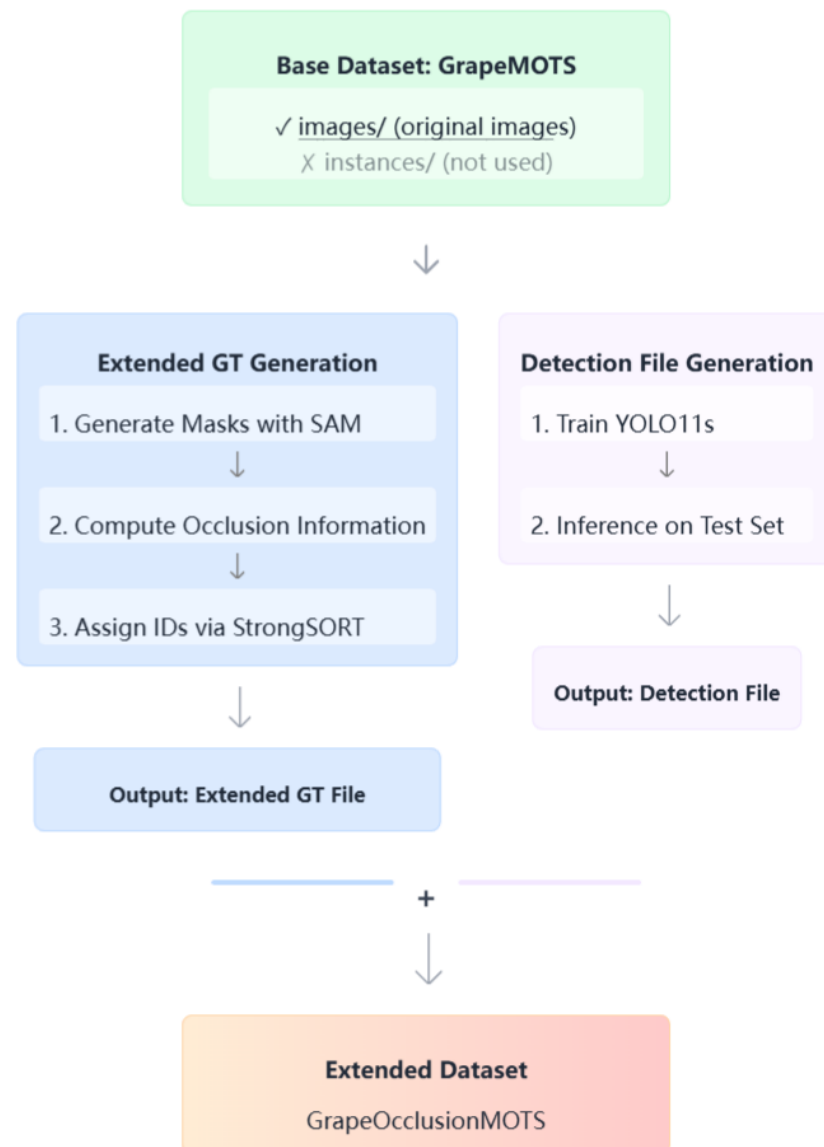


Figure 5. Illustration of the pipeline for constructing the GrapeOcclusionMOTS dataset, including mask generation, occlusion annotation, and detection file preparation for MOT evaluation. ✓ indicates items that are used, × indicates items that are not used, and + indicates the merging of two outputs to form the extended dataset.

Table 3. Performance comparison of YOLO11 detectors on the grape occlusion dataset.

Model	Precision	Recall	mAP50	mAP50-95
YOLO11n	0.852	0.763	0.824	0.463
YOLO11m	0.860	0.801	0.865	0.543
YOLO11s (Selected)	0.874	0.818	0.868	0.544

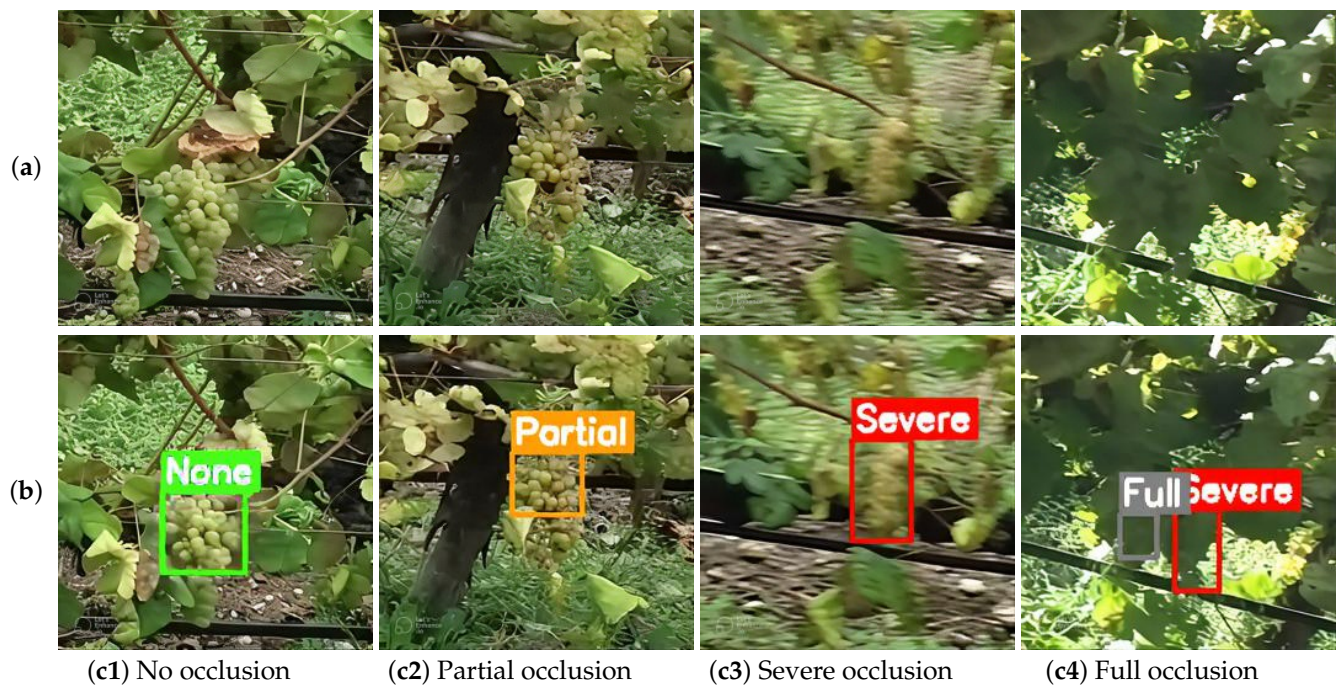


Figure 6. Representative detection results of the proposed OATSAM-Track on the GrapeOcclusion-MOTS dataset. (a) Input images. (b) Detection results. (c1–c4) Examples under different occlusion levels: no, partial, severe, and full occlusion.

4.2. Experimental Setup

4.2.1. Hardware and Software Environment

All experiments were conducted on a workstation summarized in Tables 4 and 5. The GPU used was an NVIDIA GeForce GTX TITAN X (12GB), with an Intel Core i7-5820K CPU, 62GB RAM, and Python 3.10 with PyTorch 1.12.1 + cu113 were used.

Table 4. Hardware Configuration (Actual Experiment Machine).

Component	Specification	Remark
GPU	NVIDIA GeForce GTX TITAN X	Actual experiment GPU
CPU	Intel Core i7-5820K CPU @ 3.30 GHz (6 cores, 6 threads)	Actual experiment CPU
RAM	62 GB DDR4	Actual experiment RAM
Storage	238.5 GB + 2 × 2.7 TB (NVMe/HDD)	Actual experiment storage

Table 5. Software Environment (Actual Experiment Machine).

Software	Version / Details	Remark
Operating System	Ubuntu 16.04.7 LTS	Actual OS
Python	3.10.18	Actual Python version
PyTorch	1.12.1 + cu113 (CUDA 11.3)	Actual PyTorch version
OpenCV	4.12.0	Actual OpenCV version
Key Libraries	ultralytics (v8.3.201), boxmot (v15.0.2), mobile_sam(v1.0)	Installed Python libraries used in experiments

4.2.2. Model Weights and Initialization

Detection and segmentation models, as well as the ReID and occlusion classifier modules, were initialized with the weights summarized in Table 6. YOLO11s was trained from scratch for 100 epochs, MobileSAM used pretrained weights, OSNet ReID was pretrained on MSMT17, and the occlusion classifier was trained for 20 epochs on the extended dataset.

Table 6. Model Weights (Actual Experiment).

Model	Weights / Training Details
YOLO11s [22]	Trained from scratch (100 epochs)
MobileSAM [14]	Pretrained (mobile_sam.pt)
ResNet18 [23] Occlusion Classifier	Trained on extended dataset (20 epochs)
OSNet [26] (StrongSORT ReID)	Pretrained on MSMT17 [27] (osnet_x0_25_msmt17.pt)

4.2.3. Tracker and Hyperparameter Settings

All baseline trackers were used with default hyperparameters, and OATSAM-Track employed the StrongSORT configuration summarized in Table 7.

Table 7. Final StrongSORT configuration used in the OATSAM-Track framework.

Parameter	Value	Description
max_age	30	Maximum number of frames a track is kept without detections
n_init	1	Immediate activation suitable for stable detector outputs
max_cos_dist	0.4	Appearance-matching threshold to prevent ID switches
max_iou_dist	0.8	Geometric matching threshold for slow-moving fruit clusters
min_conf	0.05	Allows low-confidence detections to avoid premature track deletion
nn_budget	100	Maximum number of stored ReID features per track
per_class	False	Class-agnostic tracking suitable for single-class fruit tracking
half	False	FP32 mode for numerically stable Kalman updates
reid_weights	OSNet-x0.25	Pretrained ReID model (MSMT17) used by StrongSORT

4.2.4. Training and Data Augmentation

During training, input images were resized to a fixed resolution, and standard data augmentation techniques, including random flipping and scaling, were applied. Unless otherwise specified, all experiments were conducted using the same training and evaluation protocol.

4.2.5. Computational Cost and Deployability

The computational overhead of SAM-based segmentation is acknowledged as a potential limitation for real-time deployment. In our experiments, the original SAM (ViT-H) requires approximately 0.8–1.0 s per frame on an NVIDIA TITAN X GPU with a batch size of 1, which is not suitable for on-field operation.

To address this issue, MobileSAM [14] is adopted in the proposed framework. With MobileSAM, the average runtime for visibility estimation is reduced to approximately 0.15–0.18 s per frame under the same hardware setting, providing sufficient speed for vineyard monitoring without compromising tracking accuracy. This significantly reduces the computational burden of occlusion-aware tracking.

It should be noted that OATSAM-Track is not designed as a high-frame-rate tracker. Instead, it targets precision agriculture scenarios where camera motion is slow and frame rates of 5–10 FPS are adequate for reliable analysis. Under these conditions, the proposed framework with MobileSAM can be deployed on GPU-equipped field hardware for vineyard monitoring. However, for resource-constrained edge devices, additional optimization may be required to achieve real-time performance, as shown in Table 8.

Table 8. Runtime comparison of different segmentation backends used for visibility estimation. All values report end-to-end per-frame costs measured within the tracking pipeline, including mask generation and necessary post-processing. The down arrow ($\downarrow$) indicates that lower values are better.

Segmentation Backend	GPU	Batch Size	Time per Frame (s) $\downarrow$
SAM (ViT-H) [10]	GTX TITAN X	1	>1.0 (impractical)
MobileSAM [14]	GTX TITAN X	1	~0.17

4.3. Evaluation Metrics

Following standard MOT evaluation protocols [28], we employ both widely used multi-object tracking metrics and occlusion-specific evaluation criteria to comprehensively assess tracking performance in complex vineyard environments.

4.3.1. Standard MOT Metrics

We first adopt widely used MOT metrics to evaluate overall tracking accuracy, localization precision, and identity consistency. Multi-Object Tracking Accuracy (MOTA) measures the combined effect of false negatives (FN), false positives (FP), and identity switches (IDSW), and is defined as shown in Equation (8):

$$\text{MOTA} = 1 - \frac{\sum_t (\text{FN}_t + \text{FP}_t + \text{IDSW}_t)}{\sum_t \text{GT}_t}, \quad (8)$$

where GT_t denotes the number of ground-truth objects at frame t . Higher MOTA values indicate better overall tracking performance.

Multi-Object Tracking Precision (MOTP) evaluates localization accuracy by measuring the average alignment between predicted and ground-truth bounding boxes, as defined in Equation (9):

$$\text{MOTP} = \frac{\sum_{t,i} d_{t,i}}{\sum_t c_t}, \quad (9)$$

where $d_{t,i}$ denotes the distance or overlap (e.g., IoU-based alignment) between the i -th matched object pair at frame t , and c_t represents the total number of matched object pairs. Lower MOTP values correspond to more precise localization.

To further assess identity preservation over time, we report the IDF1 score, which measures the harmonic mean of identity precision and identity recall [29], as defined in Equation (10):

$$\text{IDF1} = \frac{2 \cdot \text{IDTP}}{2 \cdot \text{IDTP} + \text{IDFP} + \text{IDFN}}, \quad (10)$$

where IDTP, IDFP, and IDFN denote identity true positives, false positives, and false negatives, respectively. A higher IDF1 score indicates stronger identity consistency throughout the sequence.

In addition, Precision and Recall are reported to assess detection and association quality, while Identity Switches (IDSW) and Fragmentations (Frag) are used to quantify identity stability and trajectory continuity.

4.3.2. Occlusion-Aware Evaluation

To specifically evaluate robustness under occlusion, we further perform occlusion-aware analysis by grouping frames according to ground-truth occlusion severity, including no occlusion, partial occlusion, and severe occlusion. For each occlusion subset, the F1 score is computed to evaluate the balance between precision and recall under different visibility conditions.

Moreover, occlusion-specific IDF1 scores are reported in the ablation study to analyze identity preservation under prolonged and severe occlusions. This occlusion-level

evaluation provides deeper insight into the effectiveness of the proposed occlusion-aware tracking design beyond standard MOT metrics.

4.4. Comparison with Baseline Method

We further compare the proposed OATSAM-Track with representative state-of-the-art trackers, including DeepSORT [3] and ByteTrack [4], which represent classical appearance-based and recent association-based MOT paradigms, respectively. We further include OC-SORT [30] and BoT-SORT [31] as additional modern baselines. These methods extend association-based tracking with enhanced motion modeling and robust data association strategies, and are widely regarded as strong state-of-the-art trackers under complex motion and occlusion. This extension enables a more comprehensive and up-to-date evaluation of the proposed occlusion-aware tracking framework.

It is worth noting that these baseline methods primarily rely on motion and appearance cues for data association, without explicitly modeling target visibility or occlusion states. This makes them suitable reference methods for evaluating the benefits of the proposed occlusion-aware tracking strategy.

As shown in Table 9, the proposed OATSAM-Track achieves substantial improvements over the baseline methods in key tracking metrics, including MOTA, IDF1, and recall. Notably, it maintains high precision while significantly enhancing identity consistency. These results indicate that the occlusion-aware modeling and temporal tracking strategy effectively improve robustness under challenging occlusion scenarios.

Table 9. Overall tracking performance on the full test set. Bold values indicate the best performance for each metric.

Method	MOTA	MOTP	IDF1	Precision	Recall	IDSW	Frag
DeepSORT [3]	0.126	0.144	0.301	0.602	0.418	94	26
ByteTrack [4]	0.286	0.044	0.302	0.992	0.321	183	165
OC-SORT [30]	0.362	0.041	0.387	0.971	0.398	142	138
BoT-SORT [31]	0.418	0.039	0.441	0.964	0.452	118	121
OATSAM-Track	0.712	0.031	0.642	0.918	0.756	61	74

To ensure a fair comparison, all trackers are evaluated using identical detection inputs. These inputs are generated by the same detector with fixed weights. The substantial MOTA improvement is primarily due to baseline trackers experiencing severe identity fragmentation under heavy occlusion. In contrast, OATSAM-Track explicitly models visibility and suppresses unreliable updates.

It is worth noting that MOTA is heavily influenced by detection recall. Therefore, it may not fully reflect identity preservation under prolonged occlusion. The primary advantages of OATSAM-Track are more clearly observed in identity-related metrics such as IDF1. These metrics are particularly relevant for long-term vineyard fruit monitoring. Specifically, the occlusion-specific F1 score is computed based on identity-level true positives, false positives, and false negatives, rather than frame-level detection accuracy.

As shown in Table 10, these occlusion-specific F1 scores highlight identity preservation under occlusion, rather than overall detection performance. OATSAM-Track exhibits clear advantages over DeepSORT and ByteTrack across all occlusion levels. While baseline methods degrade under severe occlusion, our approach maintains a high F1 score of 0.909, indicating robust identity consistency.

The most significant gains occur under partial occlusion, demonstrating the effectiveness of explicitly modeling occlusion states in guiding tracking. It is important to note that these metrics are computed only on subsets where targets are already detected and

tracked. Therefore, they reflect identity preservation capability under occlusion, not total detection completeness.

Table 10. Occlusion-specific F1 score comparison using ground-truth occlusion labels. These scores specifically measure identity preservation under occlusion, rather than overall detection completeness.

Method	No_F1	Partial_F1	Severe_F1
DeepSORT [3]	0.483	0.472	0.456
ByteTrack [4]	0.486	0.484	0.445
OC-SORT [30]	0.521	0.498	0.463
BoT-SORT [31]	0.548	0.517	0.482
OATSAM-Track	0.953	0.954	0.909

4.5. Ablation Study

To analyze the contribution of each occlusion-aware component in a controlled setting, we conduct an ablation study based on StrongSORT [5], a representative appearance-based tracking framework. StrongSORT is selected as the ablation baseline to ensure consistent backbone design while isolating the effects of the proposed occlusion-aware modules. The ablated components correspond to the key modules described in Sections 3.2–3.5, and are incrementally introduced to evaluate their individual and cumulative effects, as summarized in Table 11.

Table 11. Ablation study of occlusion-aware components corresponding to the proposed modules in Sections 3.2–3.5. ↑ and ↓ indicate that higher and lower values are better, respectively. ✓ and × denote whether the corresponding module is enabled or disabled.

Method	SAM	Occlusion State (ID)	Memory	ReID	IDF1 ↑	IDSW ↓	Severe-IDF1 ↑
Baseline (StrongSORT)	×	×	×	×	71.2	164	61.8
+ SAM visibility	✓	×	×	×	72.6	152	64.3
+ Occlusion state (rule)	✓	✓	×	×	74.1	138	67.5
+ Adaptive memory	✓	✓	✓	×	75.4	121	69.8
+ ReID recovery (Full)	✓	✓	✓	✓	76.8	109	72.6

Introducing SAM-based visibility estimation leads to a noticeable improvement in both IDF1 and Severe-IDF1, indicating that refined instance boundaries and visibility cues help reduce association ambiguity under partial occlusion. By further incorporating explicit occlusion-state modeling, the tracker achieves additional gains, particularly under severe occlusion, demonstrating the importance of occlusion-aware association rules. The adaptive memory module further improves identity consistency by stabilizing track representations during temporary visibility loss, resulting in a clear reduction in identity switches.

Finally, the integration of the ReID recovery mechanism yields the best overall performance, achieving the highest IDF1 and the lowest IDS. Although each individual module provides moderate improvements, their combination leads to a substantial performance gain, especially under severe occlusion. These results confirm that the proposed OATSAM-Track benefits from the complementary effects of visibility estimation, occlusion modeling, memory adaptation, and ReID-based recovery.

4.6. Qualitative Results

Figure 7 presents qualitative tracking results of the proposed method in representative vineyard scenes. The results show that OATSAM-Track maintains consistent target identities across consecutive frames, even when targets undergo partial or severe occlusion and subsequently reappear. This qualitative evidence highlights the robustness of the proposed occlusion-aware tracking strategy in real-world precision agriculture scenarios.

All experimental evaluations are designed to reflect realistic vineyard fruit monitoring conditions, with particular emphasis on identity consistency under long-term occlusion of stationary targets in agricultural environments.



Figure 7. Tracking continuity visualization across consecutive frames. The same track IDs are consistently maintained as targets move and reappear from partial or severe occlusion, demonstrating robust temporal association and ReID recovery in OATSAM-Track. Segmentation masks are generated using MobileSAM [14]. Representative examples are shown; similar tracking behavior was observed across multiple sequences and occlusion conditions.

To provide a balanced analysis, we further illustrate a representative failure case in Figure 8. As shown, under extreme foliage crowding at early tracking stages, dense leaf overlap significantly degrades the quality of SAM-generated masks, leading to unreliable visibility estimation. In such scenarios, the occlusion state may be misclassified, resulting in rapid identity switches and severe track fragmentation within a short temporal span.



Figure 8. Failure case of OATSAM-Track under extreme foliage crowding at early tracking stages. Consecutive frames (Frames 6 and 8) illustrate a representative failure scenario where dense leaf overlap and degraded SAM mask quality lead to unreliable visibility estimation. As a result, multiple targets undergo rapid identity switches within a short temporal span, causing severe identity fragmentation and unsuccessful re-identification recovery. This failure typically occurs during early track initialization under extreme occlusion conditions.

5. Discussion

5.1. Effectiveness of Occlusion-Aware Tracking in Vineyard Environments

The experimental results demonstrate that the proposed OATSAM-Track framework provides consistent improvements over representative baseline trackers, particularly under partial and severe occlusion conditions. These improvements are most pronounced in identity-related metrics, such as IDF1 and Severe-IDF1, highlighting the enhanced preservation of target identities under challenging vineyard conditions.

Unlike generic MOT benchmarks, vineyard environments are characterized by dense foliage, overlapping grape clusters, and long-term structural occlusions [1]. The experimental results indicate that the proposed occlusion-aware design is more effective than relying solely on bounding-box-level association and appearance similarity.

5.2. Practical Implications for Precision Agriculture

From a practical perspective, robust identity preservation is critical for downstream agricultural tasks such as fruit counting and yield estimation. Identity switches and trajectory fragmentation may introduce errors in fruit counting and compromise monitoring reliability.

The experimental results indicate that explicitly incorporating occlusion awareness into the tracking pipeline significantly improves long-term identity consistency in vineyard environments. In particular, reducing identity fragmentation under partial and severe occlusion is essential for stationary or low-motion targets commonly observed in precision agriculture scenarios [1].

5.3. Limitations

Despite its effectiveness, the proposed framework has several limitations. First, the current implementation assumes a relatively stable camera viewpoint and may be affected by extreme illumination changes or rapid camera motion. In rare cases of extreme occlusion or severe illumination variation, the quality of SAM-generated masks may degrade, which can negatively affect visibility estimation and occlusion state inference [32]. An example failure case is shown in Figure 8. Second, occlusion state classification relies partially on predefined visibility thresholds, which may require adaptation for different crop types or planting densities.

We note that OATSAM-Track is designed as a domain-specialized tracker targeting vineyard scenarios with frequent severe occlusions. While the integration of SAM provides strong object boundary priors [32], the current framework is not intended as a universal MOT solution. Generalization to other agricultural environments such as orchards or greenhouses, as well as adaptation to non-SAM-assisted datasets, remains an open challenge.

6. Conclusions and Future Direction

This paper presents OATSAM-Track, an occlusion-aware multi-object tracking framework designed for vineyard fruit monitoring. By integrating SAM-assisted visibility estimation with occlusion-guided temporal memory updating and re-identification recovery, the proposed method improves identity consistency under partial and severe occlusions in agricultural environments. The results demonstrate that explicit occlusion modeling combined with adaptive temporal memory substantially improves identity consistency in vineyard fruit tracking, highlighting the novelty of our approach. While the proposed method does not aim to replace general-purpose MOT frameworks, it demonstrates that explicitly modeling occlusion is particularly beneficial for vineyard fruit monitoring scenarios with stationary targets.

Future work will explore adaptive occlusion threshold learning and joint modeling of occlusion and illumination changes. In addition, extending the proposed framework to multi-camera vineyard monitoring and long-term seasonal analysis represents a promising direction to further improve robustness under diverse agricultural conditions.

Author Contributions: Conceptualization, Y.W. and H.K.; methodology, Y.W. and H.K.; writing—original draft preparation, Y.W. and H.K.; data curation, Y.W. and H.K.; validation, Y.W. and M.F.; visualization, Y.W., M.F. and L.M.D.; investigation, L.M.D. and M.F.; software, L.M.D. and M.F.; writing—review and editing, M.F. and L.M.D.; supervision, H.M.; project administration, H.M. and K.-W.L.; funding acquisition, H.M. and K.-W.L. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Korea Institute of Planning and Evaluation for Technology in Food, Agriculture, Forestry and Fisheries (IPET) through the Technology Commercialization Support Program, funded by Ministry of Agriculture, Food and Rural Affairs (MAFRA) (RS-2025-02218444), and by the “Regional Innovation System & Education (RISE)” through the Seoul RISE Center, funded by the Ministry of Education (MOE) and the Seoul Metropolitan Government (2025-RISE-01-019-04) and by the Culture, Sports and Tourism R&D Program through the Korea Creative Content Agency grant funded by the Ministry of Culture, Sports and Tourism in 2025 (Project Name: Training Global Talent for Copyright Protection and Management of On-Device AI Models, Project Number: RS-2025-02221620, Contribution Rate: 100%).

Data Availability Statement: Data available on request due to restrictions.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Kamilaris, A.; Prenafeta-Boldú, F.X. Deep learning in agriculture: A survey. *Comput. Electron. Agric.* **2018**, *147*, 70–90. [CrossRef]
2. Stein, M.; Bargoti, S.; Underwood, J. Image based mango fruit detection, localisation and yield estimation using multiple view geometry. *Sensors* **2016**, *16*, 1915. [CrossRef] [PubMed]
3. Wojke, N.; Bewley, A.; Paulus, D. Simple online and realtime tracking with a deep association metric. In Proceedings of the 2017 IEEE International Conference on Image Processing (ICIP), Beijing, China, 17–20 September 2017; IEEE: Piscataway, NJ, USA, 2017; pp. 3645–3649.
4. Zhang, Y.; Sun, P.; Jiang, Y.; Yu, D.; Weng, F.; Yuan, Z.; Luo, P.; Liu, W.; Wang, X. Bytetrack: Multi-object tracking by associating every detection box. In *Proceedings of the European Conference on Computer Vision, Tel Aviv, Israel, 23–27 October 2022*; Springer: Berlin/Heidelberg, Germany, 2022; pp. 1–21.
5. Du, Y.; Zhao, Z.; Song, Y.; Zhao, Y.; Su, F.; Gong, T.; Meng, H. Strongsort: Make deepsort great again. *IEEE Trans. Multimed.* **2023**, *25*, 8725–8737. [CrossRef]
6. Possegger, H.; Mauthner, T.; Roth, P.M.; Bischof, H. Occlusion geodesics for online multi-object tracking. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 1306–1313.
7. Li, P.; Zhang, J.; Zhu, Z.; Li, Y.; Jiang, L.; Huang, G. State-aware re-identification feature for multi-target multi-camera tracking. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, Long Beach, CA, USA, 16–17 June 2019.
8. Liu, Q.; Chen, D.; Chu, Q.; Yuan, L.; Liu, B.; Zhang, L.; Yu, N. Online multi-object tracking with unsupervised re-identification learning and occlusion estimation. *Neurocomputing* **2022**, *483*, 333–347. [CrossRef]
9. Xu, B.; He, L.; Liang, J.; Sun, Z. Learning feature recovery transformer for occluded person re-identification. *IEEE Trans. Image Process.* **2022**, *31*, 4651–4662. [CrossRef] [PubMed]
10. Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A.C.; Lo, W.Y.; et al. Segment anything. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, 1–6 October 2023; pp. 4015–4026.
11. Ariza-Sentís, M.; Wang, K.; Cao, Z.; Vélez, S.; Valente, J. GrapeMOTS: UAV vineyard dataset with MOTs grape bunch annotations recorded from multiple perspectives for enhanced object detection and tracking. *Data Brief* **2024**, *54*, 110432. [CrossRef] [PubMed]
12. Ho, T.; Bui, T.A.; Lee, P.J.; Lin, H.P.; Le, T.T.; Selva, D.; Tran, H. Mitigating Occlusion and Re-Identification Challenges in UAV Object Tracking. In Proceedings of the 2025 IEEE International Conference on Consumer Electronics-Taiwan (ICCE-Taiwan), Kaohsiung, Taiwan, 16–18 July 2025; IEEE: Piscataway, NJ, USA, 2025; pp. 21–22.
13. Cheng, B.; Misra, I.; Schwing, A.G.; Kirillov, A.; Girdhar, R. Masked-attention mask transformer for universal image segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 1290–1299.
14. Zhang, C.; Han, D.; Qiao, Y.; Kim, J.U.; Bae, S.H.; Lee, S.; Hong, C.S. Faster segment anything: Towards lightweight sam for mobile applications. *arXiv* **2023**, arXiv:2306.14289. [CrossRef]
15. Zhang, C.; Han, D.; Zheng, S.; Choi, J.; Kim, T.H.; Hong, C.S. Mobilesamv2: Faster segment anything to everything. *arXiv* **2023**, arXiv:2312.09579. [CrossRef]
16. Moonrinta, J.; Chaivivatrakul, S.; Dailey, M.N.; Ekpanyapong, M. Fruit detection, tracking, and 3D reconstruction for crop mapping and yield estimation. In Proceedings of the 2010 11th International Conference on Control Automation Robotics & Vision, Singapore, 7–10 December 2010; IEEE: Piscataway, NJ, USA, 2010; pp. 1181–1186.
17. Koirala, A.; Walsh, K.B.; Wang, Z. Attempting to estimate the unseen—correction for occluded fruit in tree fruit load estimation by machine vision with deep learning. *Agronomy* **2021**, *11*, 347. [CrossRef]

18. Matos, G.P.; Santiago, C.; Costeira, J.P.; Saldanha, R.L.; Morgado, E.M. Tracking and counting apples in orchards under intermittent occlusions and low frame rates. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 17–18 June 2024; pp. 5413–5421.
19. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You only look once: Unified, real-time object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 779–788.
20. Bochkovskiy, A.; Wang, C.Y.; Liao, H.Y.M. Yolov4: Optimal speed and accuracy of object detection. *arXiv* **2020**, arXiv:2004.10934. [CrossRef]
21. Wang, C.Y.; Bochkovskiy, A.; Liao, H.Y.M. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 7464–7475.
22. Jocher, G.; Qiu, J. Ultralytics YOLO11. *GitHub Repository*. Available online: <https://github.com/ultralytics/ultralytics> (accessed on 15 January 2025).
23. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016.
24. Kalman, R.E. A New Approach to Linear Filtering and Prediction Problems. *J. Basic Eng.* **1960**, *82*, 35–45. [CrossRef]
25. Bishop, G.; Welch, G. An introduction to the kalman filter. In Proceedings of the 28th Annual Conference on Computer Graphics and Interactive Techniques, SIGGRAPH 2001, Los Angeles, CA, USA, 12–17 August 2001; Volume 8, p. 41.
26. Zhou, K.; Yang, Y.; Cavallaro, A.; Xiang, T. Omni-scale feature learning for person re-identification. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 3702–3712.
27. Wei, L.; Zhang, S.; Gao, W.; Tian, Q. Person transfer gan to bridge domain gap for person re-identification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 79–88.
28. Bernardin, K.; Stiefelhagen, R. Evaluating multiple object tracking performance: The clear mot metrics. *EURASIP J. Image Video Process.* **2008**, *2008*, 246309. [CrossRef]
29. Ristani, E.; Solera, F.; Zou, R.; Cucchiara, R.; Tomasi, C. Performance measures and a data set for multi-target, multi-camera tracking. In *Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 8–10 and 15–16 October 2016*; Springer: Berlin/Heidelberg, Germany, 2016; pp. 17–35.
30. Cao, J.; Pang, J.; Weng, X.; Khiroudkar, R.; Kitani, K. Observation-centric sort: Rethinking sort for robust multi-object tracking. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 9686–9696.
31. Aharon, N.; Orfaig, R.; Bobrovsky, B.Z. BoT-SORT: Robust associations multi-pedestrian tracking. *arXiv* **2022**, arXiv:2206.14651.
32. Carraro, A.; Sozzi, M.; Marinello, F. The Segment Anything Model (SAM) for accelerating the smart farming revolution. *Smart Agric. Technol.* **2023**, *6*, 100367. [CrossRef]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.