Computers and Electronics in Agriculture 245 (2026) 111529

Contents lists available at ScienceDirect

Computers and Electronics in Agriculture

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

Original papers

Enhanced grape tracking (using deep neural networks) with an extended matching algorithm for SORT and DeepSORT

Jacob Piazolo [a], Benedikt Fischer [b,*], Robin Gruna [b], Jürgen Beyerer [a,b]

[a] Karlsruhe Institute of Technology (KIT), Kaiserstr. 12, Karlsruhe, 76131, Germany
[b] Fraunhofer IOSB, Fraunhoferstr. 1, Karlsruhe, 76131, Germany

ARTICLE INFO

Dataset link: https://github.com/Jax377/Grape-Detection-and-Tracking.git

Keywords:
Grape tracking
Extended matching cascade
Tracking comparison
SORT
DeepSORT

ABSTRACT

Vineyard managers traditionally count grape clusters manually for yield estimation, a process that is both time-consuming and labor-intensive. Recent advances in computer vision enable autonomous tracking, yet state-of-the-art methods often rely on re-identification networks that require expensive, hard-to-obtain instance ID annotations. This study addresses this challenge by evaluating the real-time tracking performance and counting accuracy of SORT, DeepSORT, ByteTrack, and the newly proposed SORT+ and DeepSORT+ algorithms. SORT+ and DeepSORT+ incorporate a novel matching cascade that leverages the complementary strengths of Mahalanobis, IoU, and Euclidean distances. Crucially, this approach allows SORT+ to achieve robust performance without the need for additional training data.

The extended matching cascade offers large improvements for SORT, making the training-free SORT+ comparable to the deep-learning-based DeepSORT. It increases MOTA and IDF1 by 5% to 6%, while decreasing ID switches by 62%. SORT+ improves the counting accuracy from 33% to 96%. DeepSORT+ shows further performance gains, decreasing ID switches by 12% compared to DeepSORT.

This work illustrates the feasibility of using unmanned aerial vehicles (UAVs) to autonomously track and count grape clusters in challenging real-world vineyard settings. By potentially improving yield estimation and non-destructive robotic farming, these findings support sustainable farming practices and economic growth.

1. Introduction

Wine not only plays an important role in culinary culture worldwide, but also generates approximately 339 billion dollars in revenue annually, with an expected growth rate of 1.5% each year (Statista, 2025). This market drives technical advances aimed at optimizing grape monitoring for precision agriculture and disease prevention, as well as improving grape yield estimation to inform logistical and financial decisions for vineyards and the wine industry at large. Traditional methods of estimating grape yield involve laborious and time-consuming manual counting, which is prone to human error. Emerging technologies, such as unmanned aerial vehicles (UAVs) for fast video capture and recent advances in computer vision offer promising solutions to these challenges (Khokher et al., 2023; Ariza-Sentís et al., 2023; Shen et al., 2023; Orlandi et al., 2025; Wu et al., 2023; Kumar, 2025). The challenge in automatic grape counting is to detect all unique grape clusters in a video sequence without missing or double-counting any clusters.

Two possible approaches for counting all unique clusters in a video sequence are object tracking and image stitching. The latter combines multiple video frames into a single, comprehensive panorama. This method allows for object detection on one static image, inherently avoiding the issue of double-counting instances. While image stitching has shown promise for yield estimation (Aquino et al., 2018), it can suffer from drawbacks such as the loss of grape clusters during panorama creation and a higher susceptibility to false positives compared to video-based methods (Khokher et al., 2023). Object tracking, by contrast, links detections of the same object across consecutive frames to establish its trajectory. Beyond fruit counting, tracking can be applied in agriculture for tasks such as robotic harvesting as well as in broader domains including surveillance, traffic monitoring and bulk material sorting (Maier et al., 2021). Given that tracking can filter transient detection errors over consecutive frames and offers greater flexibility, this study focuses on the advancement and evaluation of tracking algorithms.

* Corresponding author.
E-mail addresses: jacob.piazolo@outlook.com (J. Piazolo), benedikt.fischer@iosb.fraunhofer.de (B. Fischer), robin.gruna@iosb.fraunhofer.de (R. Gruna), juergen.beyerer@iosb.fraunhofer.de (J. Beyerer).

https://doi.org/10.1016/j.compag.2026.111529
Received 13 September 2025; Received in revised form 31 December 2025; Accepted 1 February 2026
Available online 12 February 2026

Tracking grape clusters, using a Kernelized Correlation Filter (Henriques et al., 2015) and Faster R-CNN (Ren et al., 2015) for detection, Jaramillo et al. (2021) demonstrate that automated counting outperforms manual methods and is more robust across sections with varying grape densities.

While the tracking of fruit, such as mangoes and apples, has been researched since the mid 2010s (Stein et al., 2016; Wang et al., 2019), automated grape tracking started only in 2020 with the introduction of the WGISD dataset (Santos et al., 2019) by Santos et al. (2020). Until 2023, studies on grape tracking and counting served primarily as a proof of concept and did not evaluate tracking performance.

In 2023, Ciarfuglia et al. (2023), Ariza-Sentís et al. (2023), and Saraceni et al. (2024) began to evaluate their tracking performance. Saraceni et al. (2024) was the first to rigorously benchmark multiple algorithms using standardized metrics, an effort that was only recently expanded by Zhai et al. (2025).

What is the best tracking algorithm to achieve high tracking performance and counting accuracy?

This study expands the current research by implementing and rigorously evaluating five tracking algorithms, each fine-tuned for the UAV RGB Video dataset from Sentís et al. (2021).

Several grape datasets annotated with bounding boxes (bboxes) and masks are available, while Sentís et al. (2021) and Saraceni et al. (2024) provide the only publicly accessible datasets that assign instance IDs to grape clusters. Therefore, in the absence of instance ID annotations, many researchers have relied on mathematical motion predictions, such as Simple-Online-Realtime-Tracking (SORT) (Bewley et al., 2016) or geometric similarities, for re-identification.

Shen et al. (2023) and Jaramillo et al. (2021) employ mathematical algorithms to predict the motion of the grape clusters and match them with detections in real time. Shen et al. (2023) track instances (grape clusters) with SORT (Bewley et al., 2016).

In Jaramillo et al. (2021), a Kernelized Correlation Filter (Henriques et al., 2015) correlates the appearance features of a grape cluster with the next frame, predicting its position. Afterwards, the predictions are matched to the detections using an Hungarian Algorithm (Kuhn, 1955).

Arlotta et al. (2023) simulate a virtual vineyard instead of using a dataset to test tracking. In their study, a virtual robot is tasked to track and harvest grapes using an Extended Kalman Filter (Kalman, 1960).

Saraceni et al. (2024) compare and evaluate the grape tracking results of SORT (Bewley et al., 2016), BoTSORT (Aharon et al., 2022), OC-SORT (Cao et al., 2023), ByteTrack (Zhang et al., 2022), StrongSORT (Du et al., 2023) and a new tracker named AgriSORT (Saraceni et al., 2024).

The tracking algorithm SORT (Bewley et al., 2016) and its further developments have been successfully used in fruit tracking tasks (Shen et al., 2023; Ciarfuglia et al., 2023; Saraceni et al., 2024; Stein et al., 2016; Li et al., 2022; Fang et al., 2022; He et al., 2022; Kirk et al., 2021). SORT (Bewley et al., 2016) is designed for speed and efficiency, making it suitable for real-time applications. It relies on a Kalman Filter (Kalman, 1960) that predicts the position of a tracked instance, which is then matched to the actual detection using the Hungarian algorithm (Kuhn, 1955).

DeepSORT (Wojke et al., 2017) uses the Mahalanobis distance (Mahalanobis, 1936) and a CNN as a re-identification network to match detections with track predictions. The CNN is trained to extract spatially close feature vectors from similar-looking instances. StrongSORT (Du et al., 2023) improves on DeepSORT (Wojke et al., 2017) by using the feature updating strategy proposed in Wang et al. (2020), camera movement compensation (Evangelidis and Psarakis, 2008), and a more robust Kalman Filter (Kalman, 1960). With small changes in DeepSORT (Wojke et al., 2017), StrongSORT (Du et al., 2023) achieves state-of-the-art results on multiple benchmarks.

ByteTrack (Zhang et al., 2022) recovers low confidence detections and matches them with track predictions, improving tracking performance in crowded videos especially. OC-SORT (Cao et al., 2023) estimates motion using the Taylor polynomial.

CSTMTrack (Xie et al., 2024) extends OC-SORT (Cao et al., 2023) by combining deep association metrics and motion information in a matching cascade. The extended matching process improves the tracking results in most metrics.

Zou et al. (2022) replace the Hungarian matching algorithm (Kuhn, 1955) in SORT (Bewley et al., 2016). Their method integrates the matching threshold directly into the cost matrix, rather than using it as a post-processing filter. Before applying the Hungarian algorithm (Kuhn, 1955), they assign a prohibitively high cost to any potential match exceeding the distance threshold, constraining the optimization to find an optimal assignment of exclusively valid pairings.

Some extensions of SORT (Bewley et al., 2016) aim to improve fruit tracking specifically. He et al. (2022) recognize that only the camera moves in agriculture. Therefore, their tracking algorithm CascadeSORT (He et al., 2022) replaces the re-identification network in DeepSORT (Wojke et al., 2017) with a traditional image retrieval method (Jégou et al., 2010) that does not require training data for feature matching, but is adequate for stationary instances.

AgriSORT (Saraceni et al., 2024) assumes that only the camera moves, while the instances remain stationary. To predict camera movement, a Kalman Filter (Kalman, 1960) first predicts the motion of the grape clusters and then "Lucas-Kanade sparse optical flow correspondences" (Lucas and Kanade, 1981) estimates the camera's motion. This estimated camera motion is then utilized to predict the coordinates of the grape bounding boxes, keeping height and width the same.

Kirk et al. (2021) add a variable to the Kalman filter (Kalman, 1960) in DeepSORT (Wojke et al., 2017) that represents the position of the camera in the row. Furthermore, they rework the CNN architecture, the appearance feature space, the matching cascade, and data preprocessing, achieving higher counting accuracy in strawberry counting compared to DeepSORT (Wojke et al., 2017).

Extending the matching process of SORT (Bewley et al., 2016) can greatly improve results (Wojke et al., 2017; Cao et al., 2023; Zhang et al., 2022; Du et al., 2023; Aharon et al., 2022; Xie et al., 2024; Zou et al., 2022; He et al., 2022; Saraceni et al., 2024; Kirk et al., 2021). However, state-of-the-art extensions (Wojke et al., 2017; Du et al., 2023; Aharon et al., 2022; Xie et al., 2024) often rely on deep re-identification networks, which require training datasets with annotated instance IDs. In the agricultural domain, creating such datasets is prohibitively labor-intensive and prone to human error compared to simple bounding box annotations. Furthermore, the original implementation of SORT (Bewley et al., 2016) uses only the IoU-overlap to match track predictions with detections, which is insufficient when detections are partial or fragmented due to foliage occlusion.

To address these limitations, we propose a systematic extension of the matching cascade. Leveraging the strengths of various distance measures, we introduce SORT+ and DeepSORT+. Both algorithms extend the matching process by incorporating the Mahalanobis distance, the IoU-overlap, and the Euclidean distance. Crucially, these similarity measures allow SORT+ to achieve robust tracking performance without requiring a trained re-identification network or ID-labeled training data, effectively bypassing the annotation bottleneck. For DeepSORT+, which utilizes appearance features, the extended cascade further refines association accuracy in cases where visual cues are ambiguous. Thus, the extended matching process improves DeepSORT (Wojke et al., 2017) and greatly improves SORT (Bewley et al., 2016), while remaining easy to implement.

Extending the matching process of SORT (Bewley et al., 2016) can greatly improve results (Wojke et al., 2017; Cao et al., 2023; Zhang et al., 2022; Du et al., 2023; Aharon et al., 2022; Xie et al., 2024; Zou et al., 2022; He et al., 2022; Saraceni et al., 2024; Kirk et al., 2021). Nevertheless, they are mostly difficult to implement and require training data, when using a re-identification network. The original implementation of SORT (Bewley et al., 2016) uses only the IoU-overlap to match the track predictions with the detections, while disregarding many other distance measures. Leveraging the strengths

Table 1
The table compares datasets available for grape detection, segmentation, and tracking. "Var" denotes the number of grape varieties included. "IDs" indicates whether instance ID annotations are used. Datasets with "Masks" are suitable for instance segmentation. "Occ" indicates occlusion (i.e., presence of partially occluded grape clusters).

Dataset	Labeled	Var	IDs	Masks	Occ	MOT(S)A
UAV RGB Sentís et al. 2021	681	1	x	x	x	−8.2%
WGISD Santos et al. 2019	300	5		x	x	46.7%
WSU Full Stages Zhang et al. 2020	459	2			x	
AI4Agriculture Morros et al. 2021	250	1			x	
GrapeCS-ML Seng et al. 2018	2078	15				
GrapeUNIPD-DL Sozzi et al. 2022	271	4			x	
CERTH Blekos et al. 2024	2502	1		x	x	
Lazio Saraceni et al. 2024	800	1	x		x	66.1%

of various distance measures, we propose SORT+ and DeepSORT+. They extend the matching process by incorporating the Mahalanobis distance, the IoU-overlap, and the Euclidean distance. These distance measures require no additional data or much computation. Thus, the extended matching process improves DeepSORT (Wojke et al., 2017) and greatly improves SORT (Bewley et al., 2016), while being easy to implement.

This work aims to serve as a reference point for the emerging field of grape tracking, counting, and yield estimation by making the following contributions:

1. Proposing an extended matching cascade for the trackers SORT (Bewley et al., 2016) and DeepSORT (Wojke et al., 2017), which are referred to as SORT+ and DeepSORT+.
2. Presenting a comprehensive evaluation of the tracking performance of SORT (Bewley et al., 2016), DeepSORT (Wojke et al., 2017), ByteTrack (Zhang et al., 2022), SORT+ and DeepSORT+, employing the uniform Mask R-CNN network (He et al., 2017) and utilizing bootstrapping (Mooney et al., 1993).
3. Evaluating the counting accuracy of the tracking algorithms to determine their suitability to estimate grape yield.

The algorithms were implemented, trained, and evaluated using Open-MMLab's MMTracking toolbox (https://github.com/open-mmlab/mmtracking) on a Windows 10 platform. MMTracking was modified, and new code was added. The altered code and detailed configuration of the detector and trackers is publicly accessible on GitHub (https://github.com/Jax377/Grape-Detection-and-Tracking.git).

2. Materials and methods

2.1. Dataset

A re-identification network, like the one used in DeepSORT (Wojke et al., 2017), classifies which grape clusters are the same (have the same instance ID) by comparing detected grapes across frames. Training such a network requires grape clusters in the training subset to be annotated with instance IDs. While trackers like SORT (Bewley et al., 2016) or Structure from Motion (SfM) (Hartley and Zisserman, 2003) do not require a re-identification network, evaluating their tracking performance necessitates instance ID labels in the test subset. (See Table 1.)

Only the UAV RGB Video dataset (Sentís et al., 2021) and the Lazio dataset (Saraceni et al., 2024) contain instance IDs and are therefore suitable to evaluate tracking. Because the camera does not capture the entire vine row, the dataset of Saraceni et al. (2024) is best suited for applications in robotic detection and tracking, rather than for estimating grape yield. For segmenting and counting grape clusters, the UAV RGB Video Dataset, published by Sentís et al. (2021) continues to be the most suitable dataset available and is therefore used in this study.

Fig. 1. Third annotated image from row 7.4.2. Red boxes: Grape clusters that are not annotated on the vine row of interest. Blue boxes: Grape clusters in the background. Orange boxes: Ground Truth data of annotated grape clusters.

UAV RGB Video dataset problems

The UAV RGB Video dataset (Sentís et al., 2021) faces two major issues that hinder detection and consequently tracking, as well as a third problem that presents difficulties specifically for grape yield estimation.

1. The grape clusters occupy a very small area in the frame, ranging from 0.01% to 0.17%, which is less than half the size typically seen in other available datasets. This significantly reduces tracking accuracy, as the algorithms depend on bounding box overlap across frames.
2. Some grape clusters were not identified and thus not annotated. Refer to Fig. 1. This negatively affects the detector's performance and, by extension, the tracker's accuracy. Moreover, the missing annotations cause false positives that should have been true positives, substantially lowering the MOTA and MOTSA scores.
3. Grape clusters present in vine rows in the background (see Fig. 2) are incorrectly detected as false positives. This not only further reduces the MOTA and MOTSA scores, but also falsely inflates yield calculations.

The UAV RGB Video dataset (Sentís et al., 2021) was partitioned into a training set and a test set. The training set consists of 528 annotated images, while the test set contains 136 images from distinct video sequences, as detailed in Table 2. Each video sequence contains between 250 and 2000 frames and, on average, approximately 22 of those frames are annotated.

2.2. Detection metrics

The Intersection over Union (IoU) measures the overlap between two bounding boxes, typically between a detection and a ground truth. If the IoU exceeds a specified threshold (e.g. 0.5), the prediction is classified as a true positive (TP) and otherwise as a false positive (FP). Precision is the model's accuracy in identifying TPs among all

Fig. 2. The image from row 8.1 is zoomed in on detected grape clusters in the background.

Table 2
For the training/test split, four vine rows of the UAV RGB Video dataset (Sentís et al., 2021) were split into subsequences.

Subset	Rows
Training set	4.3.2, 4.4.2, 4.4.4, 6.1.3, 6.1.4, 6.2.1, 6.2.2, 6.3, 7.1.3, 7.1.4, 7.2.1, 7.2.2, 7.2.3, 7.2.4, 7.3.1, 7.3.2, 7.3.3, 7.3.4, 7.4.1, 7.4.2, 8.2, 8.3, 8.4
Test set	4.2.1, 6.1.1, 6.1.2, 7.1.1, 7.1.2, 8.1

detections it makes. Recall is the percentage of detected TPs out of the total number of TPs that were possible to detect. A Precision–Recall curve can be created for a certain IoU-threshold by plotting Precision and Recall on a graph for confidence scores from 0.00 to 1.00 in increments of 0.01 ($[0.00, 0.01, \ldots, 0.99, 1.00]$). The mean Average Precision (mAP) for a certain IoU-threshold, such as 0.5, equals the area under the Precision–Recall curve. This area can be calculated using the trapezoidal rule. Traditionally in object detection, mAP refers to the average of the mAPs calculated at the IoU-thresholds ranging from 0.50 to 0.95 in increments of 0.05 ($[0.50, 0.55, \ldots, 0.90, 0.95]$). The average mAP usually serves to identify the optimal checkpoint (weights at a certain iteration/epoch) for a model. To calculate AP_{small}, mAP_{medium} and mAP_{large} the bounding boxes are categorized as small, medium, or large. The size classifications follow the COCO standard classifications (Lin et al., 2014). The mAP can be calculated similarly for masks using the same methodology.

2.3. Tracking metrics

Due to the high occlusion and the small size of the grape clusters in the Appendix (Sentís et al., 2021), the resulting detector often detects only parts of a grape cluster. The Hungarian Algorithm (Kuhn, 1955) makes an optimal one-to-one assignment by maximizing the average IoU-overlap. If matches exhibit less than 20% IoU-overlap, they are unmatched, making the detection a false positive (FP) and the ground truth (GT) a false negative (FN). The 20% IoU-threshold strikes a good balance between accurately assigning correct detections to GTs and minimizing the number of ID switches (IDsw) due to incorrect assignment. To evaluate tracking, we use the metrics IDsw, Multi Object Tracking Accuracy (MOTA), Multi Object Tracking Precision (MOTP), IDF1, and counting accuracy.

IDsw counts the number of times the tracker incorrectly switches the identity of a grape cluster from one instance ID to another. MOTA is one of the most important metrics to evaluate tracking performance. However, MOTA is influenced by the performance of the detector. Poor detection results cause poor MOTA scores. Multi Object Tracking and Segmentation Accuracy (MOTSA) closely resembles MOTA, but also incorporates segmentation masks to calculate TPs, FPs and FNs. MOTA is based on FP, FN, and IDsw. MOTP evaluates the precision with which the locations of the tracked objects are estimated. This happens independently of how accurately instance IDs are assigned to tracks. MOTP is based on IoU. The closer MOTP is to 1, the more accurate the Multi Object Tracking Precision is. Multi Object Tracking and Segmentation Precision (MOTSP) is very similar to MOTP, but it also includes the segmentation model's masks to calculate the IoU-overlap between the masks and their corresponding ground truths (GTs). IDF1 is a metric that assesses an algorithm's capability to accurately identify and maintain instance IDs of objects over time. It offers a single measure that reflects both the accuracy and the robustness of the tracking in maintaining consistent identities. Counting accuracy is a metric that evaluates the precision with which a tracking algorithm counts the number of grape clusters in a dataset.

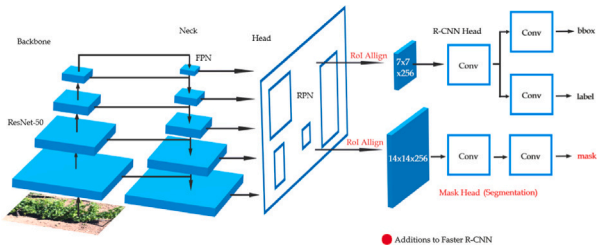

Fig. 3. Graphical illustration of the Mask R-CNN Architecture presented by He et al. (2017).

$$\text{Counting Accuracy} = 1 - \left| \frac{AG - CT}{AG} \right| \tag{1}$$

AG represents the number of annotated grape clusters in the dataset, while CT denotes the number of grape clusters counted by the tracking algorithm.

2.4. Mask R-CNN for detection

Research conducted by Poblete-Echeverria et al. (2025), Olenskyj et al. (2022), Lopes and Cadima (2021) explores which attributes most closely correlate with grapevine yield. Lopes and Cadima (2021) demonstrate that, during the final stages of berry ripening, the area of the grape clusters is a more accurate predictor of yield than the total number of berries counted. This finding is supported by Olenskyj et al. (2022), where the grape cluster area shows a stronger correlation with yield compared to the number of grape clusters. Consequently, a Mask R-CNN network (He et al., 2017) was chosen for detection, due to its ability to output the areas of the grape clusters. However, caution is advised when applying the findings from one grape species to another, as Lopes and Cadima (2021) observe differences in the results depending on the grape species.

A Mask R-CNN network, as proposed by He et al. (2017) extends the functionalities of Faster R-CNN (Ren et al., 2015) by introducing an additional branch that predicts segmentation masks for each Region of Interest (RoI). This enables the model to perform both object detection – identifying bounding boxes around objects – and instance segmentation – generating pixel-level masks for each object. The Mask R-CNN (He et al., 2017) framework is built from several components as illustrated in Fig. 3. Each component is customizable or interchangeable to best suit the task at hand. For details on configurations, the reader is referred to Bouraya and Belangour (2021).

The Mask R-CNN detector was trained on the 528 images of the training set. As this research focuses on evaluating trackers rather than the detector itself, we aimed to provide the trackers with the best possible inputs. Therefore, the model from the training epoch that achieved the highest performance on the test set was selected for all subsequent experiments, forgoing a separate validation set. This approach ensures a fair comparison of the trackers' performance.

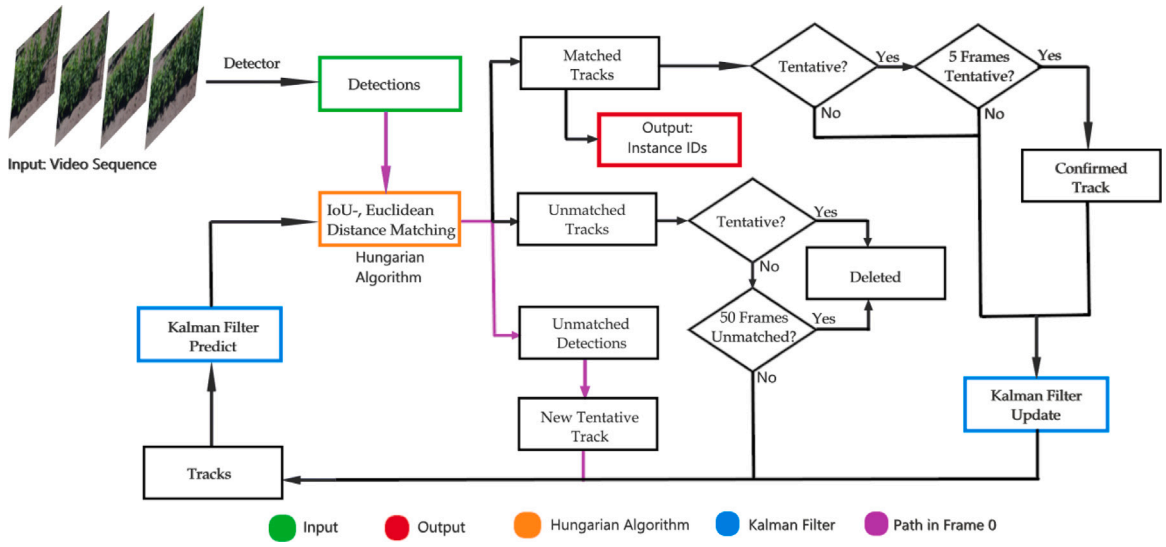

Fig. 4. Graphical illustration of the Simple Online and Realtime Tracking (SORT) algorithm presented by Bewley et al. (2016).

2.5. Tracking

This study examines two-step multi object tracking approaches, separating the process into distinct detection and tracking stages. As a result, detection functions independently of tracking. This modular framework allows for easy substitution of either the detector or the tracker.

To ensure full replicability, the complete source code, including all configuration files, hyper-parameters, and trained models used in this study, is publicly available at https://github.com/Jax377/Grape-Detection-and-Tracking.git.

This section introduces three algorithms for the tracking step.

2.5.1. Simple online and realtime tracking (SORT)

The multi object tracking algorithm SORT, as introduced by Bewley et al. (2016), is designed for speed and efficiency, making it well-suited for real-time applications. The mathematical tracker does not depend on training data, which allows the application of SORT (Bewley et al., 2016) on datasets lacking instance IDs in the training data.

An illustration of the SORT algorithm is shown in Fig. 4. Upon detecting all instances (grape clusters) in frame n, SORT (Bewley et al., 2016) employs a Kalman Filter (Kalman, 1960) to predict the position of each grape cluster in the subsequent frame $n + 1$. The Hungarian matching algorithm (Kuhn, 1955) is then used to optimally assign each predicted position to at most one detected instance in frame $n + 1$, within certain regional constraints. The Kalman Filter (Kalman, 1960) continues to predict the positions of unmatched tracks for up to 50 frames, until a detection match is found or the track is terminated. Detections that are not matched with any track initiate new tentative tracks, which are promptly removed if they remain unmatched. After a duration of five frames, a tentative track is confirmed. This helps to exclude false positives (FPs) from the count, as they are less likely to occur in five consecutive frames.

2.5.2. DeepSORT

Simple Online and Realtime Tracking with a Deep Association Metric (DeepSORT), developed by Wojke et al. (2017), extends the SORT (Bewley et al., 2016) algorithm by integrating a classifier to compare the appearances of objects (grape clusters) across frames. This feature makes DeepSORT (Wojke et al., 2017) potentially well suited for tracking grapes, because of the inanimate grape clusters and its capability to re-identify grapes that have been occluded or undetected for several frames.

As can be seen in Fig. 5, DeepSORT (Wojke et al., 2017) builds upon the matching mechanism of SORT (Bewley et al., 2016) by implementing the Mahalanobis distance (Mahalanobis, 1936) to filter out improbable matches (below 0.5% probability) and employing a classification network (Fig. 6) that extracts appearance features of objects.

DeepSORT (Wojke et al., 2017) aims to re-identify grape clusters when there is no overlap between detection and track prediction following a period of occlusion, potentially reducing instance ID switches (IDsw) and increasing tracking performance.

2.5.3. ByteTrack

In traditional vineyards, the occlusion of grape clusters significantly hinders detection efforts. Partly occluded objects receive lower confidence scores from the detection algorithm and are discarded if these scores fall below a specific threshold. This process results in the loss of true objects and leads to fragmented tracks.

ByteTrack, developed by Zhang et al. (2022), addresses this issue by recovering low-confidence detections that fit existing tracks.

The algorithm employs a matching cascade that prioritizes highly probable detections. Subsequently, the remaining track predictions are matched with the low-confidence detections. To handle cases where Mask R-CNN (He et al., 2017) detects two bboxes for the same track, ByteTrack (Zhang et al., 2022) scales the IoU-overlap between track predictions and detections based on the confidence scores of the detections. This adjustment ensures that more confident detections are prioritized for matching, potentially increasing matching accuracy.

2.5.4. Our proposed extended matching cascade

The original implementations of SORT (Bewley et al., 2016) and ByteTrack (Zhang et al., 2022) rely solely on IoU-overlap for matching tracks with detections. DeepSORT (Wojke et al., 2017) additionally incorporates appearance features in combination with the Mahalanobis distance (Mahalanobis, 1936). To leverage the complementary strengths of these similarity measures, we formalize our matching strategy as a recursive optimization process.

Formally, let $\mathcal{T} = \{T_1, \ldots, T_M\}$ be the set of predicted tracks from the Kalman Filter and $\mathcal{D} = \{D_1, \ldots, D_N\}$ be the set of new detections in the current frame. The matching process is organized into K sequential stages. Let $\mathcal{U}_{\mathcal{T}}^{(k)}$ and $\mathcal{U}_{\mathcal{D}}^{(k)}$ denote the sets of unmatched tracks and detections, respectively, at the start of stage k, initialized as $\mathcal{U}_{\mathcal{T}}^{(1)} = \mathcal{T}$ and $\mathcal{U}_{\mathcal{D}}^{(1)} = \mathcal{D}$.

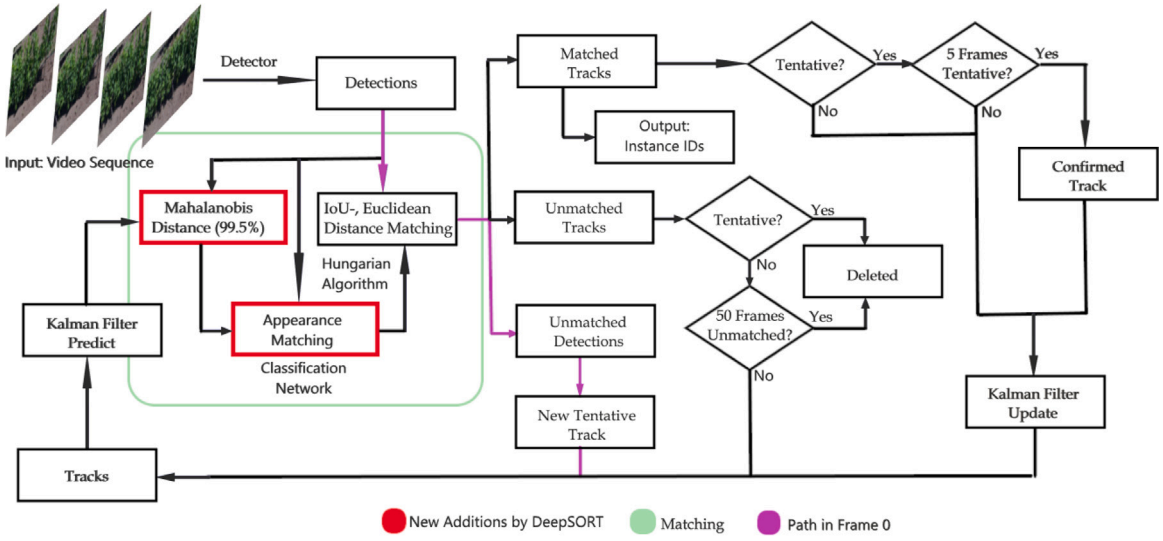

Fig. 5. Graphical illustration of the DeepSORT algorithm as an extension of SORT presented by Wojke et al. (2017).

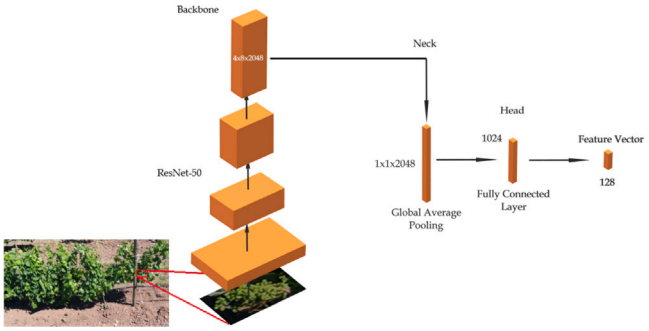

Fig. 6. Graphical illustration of the Classification Network in DeepSORT presented by Wojke et al. (2017).

At each stage k, we seek an optimal assignment matrix $\mathbf{X}^{(k)} \in \{0,1\}^{|\mathcal{U}_{\mathcal{T}}^{(k)}| \times |\mathcal{U}_{D}^{(k)}|}$ that minimizes the total cost based on a stage-specific distance metric $d^{(k)}$:

$$\min_{\mathbf{X}^{(k)}} \sum_{T_i \in \mathcal{U}_{\mathcal{T}}^{(k)}} \sum_{D_j \in \mathcal{U}_{D}^{(k)}} C_{i,j}^{(k)} \cdot x_{i,j} \tag{2}$$

subject to $\sum_i x_{i,j} \leq 1$ and $\sum_j x_{i,j} \leq 1$. To ensure the reliability of matches, we enforce a gating threshold τ_k. The cost matrix element $C_{i,j}^{(k)}$ is defined as:

$$C_{i,j}^{(k)} = \begin{cases} d^{(k)}(T_i, D_j) & \text{if } d^{(k)}(T_i, D_j) \leq \tau_k \\ \infty & \text{otherwise} \end{cases} \tag{3}$$

The linear assignment problem is solved using the Hungarian algorithm (Kuhn, 1955), yielding a set of matches $\mathcal{M}^{(k)}$. The sets of unmatched instances are updated recursively for the subsequent stage:

$$\mathcal{U}_{\mathcal{T}}^{(k+1)} = \mathcal{U}_{\mathcal{T}}^{(k)} \setminus \{T_i \mid \exists D_j, (T_i, D_j) \in \mathcal{M}^{(k)}\}. \tag{4}$$

This process repeats until all stages are completed. We employ three distinct distance metrics $d^{(k)}$ across the stages:

IoU Distance (d_{IoU}): Defined by the intersection over union of the track bounding box B_{T_i} and detection bounding box B_{D_j}:

$$d_{IoU}(T_i, D_j) = 1 - \frac{|B_{T_i} \cap B_{D_j}|}{|B_{T_i} \cup B_{D_j}|}. \tag{5}$$

Mahalanobis Distance (d_{Mah}): Measures the distance between the detection position $\mathbf{d}_j$ and the Kalman filter's predicted mean $\mathbf{y}_i$ and covariance $\mathbf{S}_i$:

$$d_{Mah}(T_i, D_j) = \sqrt{(\mathbf{d}_j - \mathbf{y}_i)^{\top} \mathbf{S}_i^{-1} (\mathbf{d}_j - \mathbf{y}_i)}. \tag{6}$$

Euclidean Center Distance (d_{Euc}): Defined by the Euclidean norm between the centers $\mathbf{c}$ of the bounding boxes:

$$d_{Euc}(T_i, D_j) = \|\mathbf{c}_{T_i} - \mathbf{c}_{D_j}\|_2. \tag{7}$$

2.5.5. SORT+

SORT+ implements this recursive cascade in four stages ($k = 1$ to 4) to leverage geometric similarities without requiring appearance features. The process is illustrated in Fig. 7.

Stage 1 (Strict IoU): All detections and track predictions are matched where the IoU-overlap exceeds 20% ($d_{IoU} \leq 0.8$). This prioritizes highly probable, overlapping matches.

Stage 2 (Mahalanobis): Previously unmatched detections are matched to the multivariate normal distribution of the closest track. Matches are accepted only if the Mahalanobis distance d_{Mah} is below the chi-squared 0.9 quantile, $\tau_2 = \chi^2_{0.9;2} = 4.605$.

Stage 3 (Relaxed IoU): The IoU-overlap is again utilized to match remaining unmatched detections with remaining track predictions, with a relaxed minimum overlap of 3.5% ($d_{IoU} \leq 0.965$).

Stage 4 (Euclidean): Finally, we employ d_{Euc} to recover matches where Mask R-CNN (He et al., 2017) detects only part of a grape cluster, resulting in no IoU-overlap. The threshold τ_4 is dynamic, based on the track's predicted width w_i and height h_i:

$$\tau_4 = w_i + \frac{h_i}{2}. \tag{8}$$

The Euclidean distance aims to recover fragmented tracks where the Kalman Filter (Kalman, 1960) causes the track estimate to converge onto the average size of the grape cluster.

2.5.6. DeepSORT+

DeepSORT+ extends the matching cascade of SORT+ by incorporating visual appearance information. It leverages the classification network to match visually similar grape clusters, identified by their spatially close feature vectors. The extended matching cascade of DeepSORT+ is shown in Fig. 8.

DeepSORT+ utilizes an initial **Stage 0** before the geometric cascade. In this stage, we utilize the Euclidean distance between the extracted

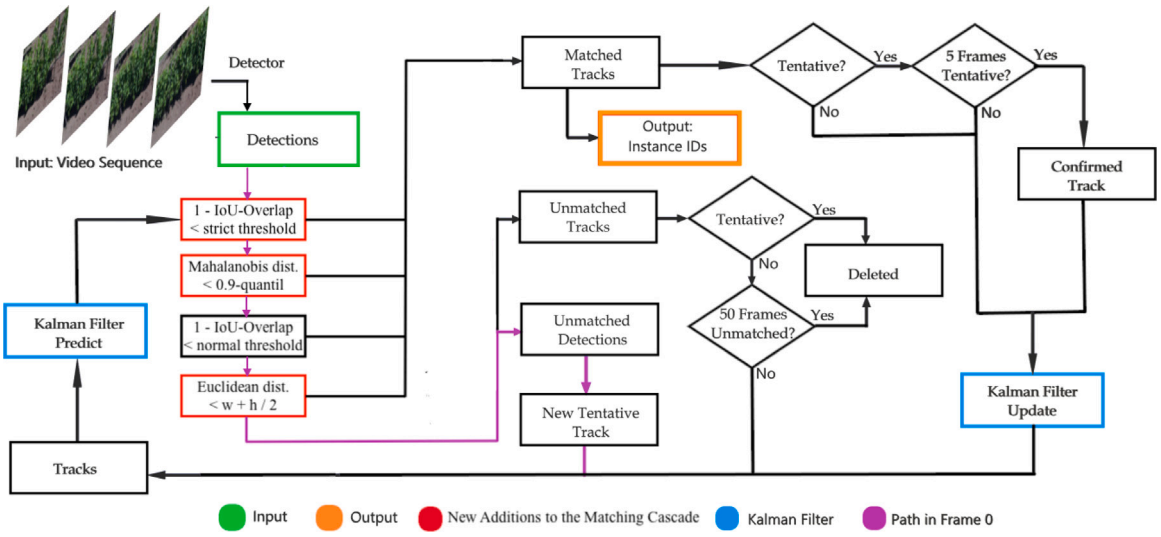

Fig. 7. An overview of the proposed SORT+ tracking framework. The diagram shows the overall flow, starting from video input and Mask R-CNN detections. Central to the framework is the multi-stage Extended Matching Cascade (red boxes). Within this cascade, valid 1-to-1 assignments that satisfy the specific threshold criteria exit to the right as matched tracks, while remaining unmatched tracks and detections propagate downward to the subsequent stage. The figure also explains how unmatched detections initiate tentative tracks that must survive for five consecutive frames to become confirmed instance IDs.

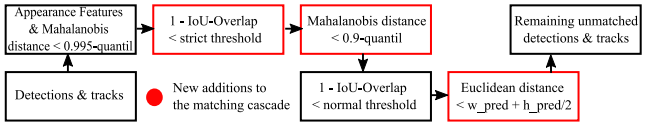

Fig. 8. The Extended Matching Cascade of DeepSORT+.

feature vectors of the track and detection. A match is confirmed only if two conditions are met: the Mahalanobis distance is smaller than the chi-squared 0.995 quantile ($\chi^2_{0.995;2} = 10.597$) and the Euclidean distance between feature vectors is below $\tau_0 = 1.5$. Following this appearance-based matching, the algorithm proceeds with the same four geometric stages ($k = 1$ to 4) defined in SORT+: strict IoU, Mahalanobis distance, relaxed IoU, and Euclidean center distance. This extended matching cascade aims to identify matches between detections and tracks that were missed by the original implementations, leveraging visual identity where available and falling back on robust geometric associations for occluded or ambiguous targets.

3. Results

All results presented in this chapter were evaluated on the test set defined in Table 2.

3.1. Detection results

A Mask R-CNN network (He et al., 2017) was optimized using the Adam Optimizer (Kingma and Ba, 2014) (lr = 0.000105) to achieve a bbox mAP of 19.4%. This model is utilized in all subsequent detection and tracking evaluations.

Table 3 shows the bounding box mAP alongside the mAP values using the IoU-thresholds of 0.5 and 0.75. Furthermore, Table 3 lists mAP_s, mAP_m and mAP_l by categorizing the detected bounding boxes into small (s), medium (m) and large (l). The sizes follow the COCO standard classifications (Lin et al., 2014).

Similarly, Table 4 evaluates the segmentation performance of the Mask R-CNN model by presenting the mAP metrics using the pixel-level segmentation masks instead of the bounding boxes.

Table 3
Mask R-CNN bounding box results.

↑ mAP	↑ mAP_{50}	↑ mAP_{75}	↑ mAP_s	↑ mAP_m	↑ mAP_l
19.4%	44.8%	14.1%	5.5%	19.2%	28.0%

Table 4
Mask R-CNN segmentation results.

↑ mAP	↑ mAP_{50}	↑ mAP_{75}	↑ mAP_s	↑ mAP_m	↑ mAP_l
14.5%	44.3%	4.5%	4.4%	13.8%	23.0%

3.2. Classification network results

The classification network used in DeepSORT (Wojke et al., 2017) is trained to generate appearance feature vectors from detected grape clusters. The feature vectors are spatially close for similar appearances. For the classification network a ResNet-50 (Koonce, 2021) backbone was incorporated, followed by Global Average Pooling (Hsiao et al., 2019) and a fully connected layer.

Using an Adam optimizer (Kingma and Ba, 2014) with a learning rate of 0.000105 and a weight decay of 0.0001, the convolutional neural network achieved a mAP of 80.6%.

4. Tracking results

Tables 5 and 6 fulfill one of the main objectives of this study by comparing the tracking results of SORT (Bewley et al., 2016), DeepSORT (Wojke et al., 2017), ByteTrack (Zhang et al., 2022), SORT+ and DeepSORT+. The best value for each metric is highlighted in bold. Since the MOTP depends solely on the detector, the variance in the results comes from bootstrapping (Mooney et al., 1993). Therefore, no values were highlighted for this metric.

To assess the statistical stability of the results, the metrics in Table 5 were estimated via bootstrapping (Mooney et al., 1993). From the original test set of 6 video sequences, 1000 resamples of the same size were generated by drawing with replacement. The table reports the mean performance across these 1000 trials and the 95% confidence interval, which quantifies the reliability of the results.

The included results of PointTrack from Ariza-Sentís et al. (2023) are in parentheses to emphasize the limited comparability of their

Table 5
Tracking results of the five algorithms.

Tracker	↑ MOTA	↑ IDF1	↓IDsw	↑ MOTP
SORT (Bewley et al., 2016)	37.69% ± 0.33%	60.42% ± 0.09%	73.50 ± 1.20	65.41% ± 0.08%
DeepSORT (Wojke et al., 2017)	42.25% ± 0.37%	66.93% ± 0.21%	17.05 ± 0.29	65.46% ± 0.07%
ByteTrack (Zhang et al., 2022)	41.53% ± 0.37%	66.01% ± 0.19%	26.09 ± 0.33	65.44% ± 0.08%
SORT+	41.50% ± 0.40%	66.48% ± 0.18%	28.24 ± 0.49	65.40% ± 0.07%
DeepSORT+	**42.71%** ± 0.37%	**67.10%** ± 0.19%	**14.90** ± 0.23	65.39% ± 0.08%
PointTrack (Ariza-Sentís et al., 2023)	(−8.2%)	–	(19)	(66.6%)

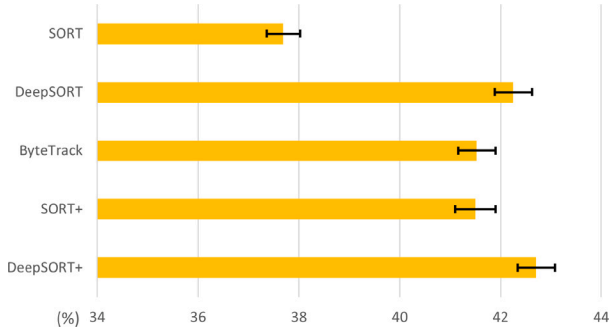

Fig. 9. MOTA scores with their corresponding 95% confidence interval.

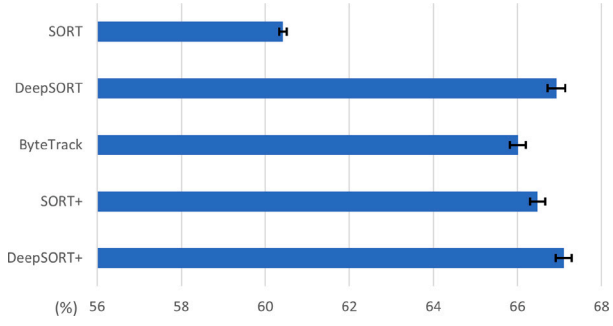

Fig. 10. IDF1 scores with their corresponding 95% confidence interval.

findings with this research. It is unknown what data was used for the testing by Ariza-Sentís et al. (2023). Depending on the size of the test subset, the number of instance ID switches (IDsw) can be interpreted positively or negatively. Furthermore, while Ariza-Sentís et al. (2023) employ multi-object tracking and segmentation accuracy (MOTSA) and Precision (MOTSP) in their evaluation, this study uses MOTA and MOTP. However, it is important to note that MOTSA and MOTSP are derived using the same foundational formulas as MOTA and MOTP. This similarity in the calculation methods provides a basis for comparison, albeit with the noted limitations.

Fig. 9 shows the MOTA scores of the five tracking algorithms, Fig. 10 the IDF1 scores, and Fig. 11 the number of ID switches made by the tracking algorithms. If the 95%-confidence intervals of two individual trackers overlap, it is inconclusive which tracking algorithm performs better.

Olenskyj et al. (2022) and Orlandi et al. (2025) show that the count of grape clusters together with the grape cluster area is a strong indicator of grape yield. As grape yield estimation is an important motivation for this research, the following Table 6 presents the number of grape clusters counted by the tracking algorithms.

A tracking algorithm counts the grape clusters within the test subset by tallying the number of confirmed tracks. The test subset contains 147 annotated grapes (AG). However, due to the problem of missing annotations within the dataset from Sentís et al. (2021) (see Fig. 1), there are additional visible grape clusters. The manual hand counting (HC) of the grapes in the video confirms at least 150 grape clusters

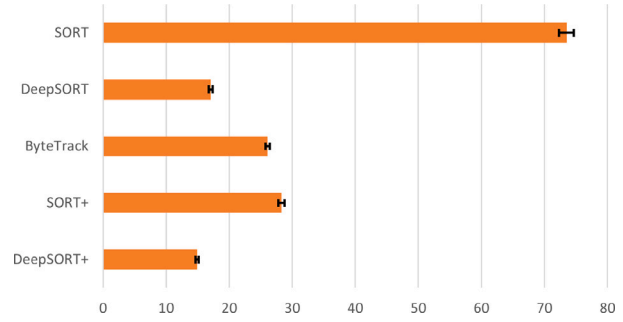

Fig. 11. Number of ID switches with their corresponding 95% confidence interval.

in the foreground (f) and 36 grape clusters in the background (b), resulting in a total of at least 186 visible grape clusters. Instead of the number of annotated grapes, the hand-counted grape clusters serve as a reference to evaluate the counting accuracy of the tracking algorithms.

5. Discussion

5.1. Evaluation of tracking algorithms

This study evaluates the performance of the enhanced matching process by comparing SORT+ and DeepSORT+ with existing tracking algorithms, such as SORT (Bewley et al., 2016), DeepSORT (Wojke et al., 2017) and ByteTrack (Zhang et al., 2022) in the tracking and counting of grape clusters. All trackers employ a uniform Mask R-CNN detector (He et al., 2017).

Summarizing the results, the selected Mask R-CNN network (He et al., 2017) achieves a bounding box mAP of 19.4% and a segmentation mAP of 14.5%. Despite the low detection accuracy, all trackers show robust tracking performance, with MOTA scores ranging from 37% to 43% and IDF1 scores from 60% to 67%.

The tracking algorithms tend to overcount grape clusters by up to 68%. However, DeepSORT (Wojke et al., 2017), SORT+ and DeepSORT+ maintain a more accurate count, with at most 5% overcount. This improved counting accuracy is likely due to the use of the Mahalanobis distance (Mahalanobis, 1936) as their only common similarity measure.

DeepSORT+ consistently outperforms the other tracking algorithms in almost every metric, achieving the highest MOTA score of 42.7% (±0.4%) and an IDF1 score of 67.1% (±0.2%), with only 15 ID switches over 186 counted tracks. Meanwhile, the original implementation of DeepSORT (Wojke et al., 2017) performs comparably, with overlapping 95%-confidence intervals in MOTA and IDF1 and 17 ID switches in 195 counted tracks. Due to overlapping confidence intervals, it remains inconclusive which tracking algorithm consistently achieves better tracking results on similar datasets. As only the classification network distinguishes the top-performing trackers DeepSORT (Wojke et al., 2017) and DeepSORT+, from the other algorithms, it proves highly effective in generating accurate feature vectors to match detections with their corresponding tracks. During evaluation, the classification network achieves a mAP of 80.6%.

Table 6
Grape Cluster (Track) Count and Counting Accuracy for the five algorithms. Number of annotated grapes (AG) and hand counted grapes (HC) in foreground (f) and background (b).

	AG	HC (f)	HC (b)	SORT	DSORT	ByteTrack	SORT+	DSORT+
Count	147	150	36	313	195	238	192	**186**
Counting Accuracy	–	–	–	32.7%	95.2%	72.0%	96.8%	**100%**

The original implementation of SORT (Bewley et al., 2016), being the precursor to the other implementations, exhibits the most identity switches; more than twice those of SORT+ and ByteTrack (Zhang et al., 2022), and more than four times those of DeepSORT (Wojke et al., 2017) and DeepSORT+. However, SORT+ achieves results comparable to ByteTrack (Zhang et al., 2022) and within the reach of DeepSORT(+), showing that integrating additional similarities in the matching process can make SORT (Bewley et al., 2016) a viable tracker and without requiring additional data and training.

Choosing a tracking algorithm for grape tracking involves considering not only the tracking performance and counting accuracy, but also the implementation complexity.

Regarding computational complexity, all evaluated algorithms rely on the Hungarian algorithm (Kuhn, 1955) for bipartite matching, scaling with $O(n^3)$ for n tracks and detections. However, a critical distinction lies in the feature extraction overhead. DeepSORT (Wojke et al., 2017) and DeepSORT+ incur a substantial computational cost due to the re-identification network. Specifically, the ResNet-50 (Koonce, 2021) backbone must perform a forward pass for every detected object proposal to generate appearance embeddings. This results in a heavy constant overhead per frame that scales linearly with the number of detections.

In contrast, SORT (Bewley et al., 2016), SORT+ and ByteTrack (Zhang et al., 2022) are purely geometric. They rely on efficient matrix operations, such as IoU and Mahalanobis distance (Mahalanobis, 1936) calculations, which are computationally negligible compared to the re-identification network inference. While the proposed SORT+ and DeepSORT+ algorithms introduce a fixed number of additional matching stages, these stages consist solely of these lightweight vector operations. Consequently, the extended matching cascade improves tracking robustness without altering the asymptotic time complexity or introducing the heavy latency associated with deep feature extraction.

Although DeepSORT (Wojke et al., 2017) and DeepSORT+ yield the best results in tracking and counting, ByteTrack (Zhang et al., 2022) and SORT+ present viable alternatives. These mathematical algorithms, which only require training of the detector, increase the number of ID switches by approximately ten, while only slightly underperforming DeepSORT(+) by around 1% in MOTA and IDF1. SORT+, in particular, achieves good counting accuracy, identifying 192 grape clusters of at least 186 visible clusters, suggesting that lost IDs are later recovered by another ID switch.

Because ByteTrack (Zhang et al., 2022) and SORT+ do not rely on a classification network, which requires training and labeled instance IDs in the dataset, both algorithms are easier to implement and feasible for broader applications. However, to achieve the best tracking performance, the advanced capabilities of DeepSORT (Wojke et al., 2017) and DeepSORT+ seem to be most promising.

5.2. Contextualizing the results

This study ties in with current research on grape tracking (Saraceni et al., 2024; Ariza-Sentís et al., 2023; Ciarfuglia et al., 2023), demonstrating that accurate grape tracking and counting is feasible on challenging datasets that resemble real-world conditions.

Ariza-Sentís et al. (2023) evaluate grape tracking using the same dataset (Sentís et al., 2021) as this study, achieving a MOTSP of 66.6%, which is comparable to the MOTP of 65.4% (±0.1%) reported here. However, their MOTSA score of −8.2% using PointTrack (Xu et al., 2022) indicates a high number of false positives (FPs) and false negatives (FNs). This is likely due to the small size of the grape clusters, which are not sufficiently distinct from the background for effective pattern matching. In contrast, SORT (Bewley et al., 2016), DeepSORT (Wojke et al., 2017), ByteTrack (Zhang et al., 2022), SORT+ and DeepSORT+ achieve MOTA scores between 37% and 43%. This suggests that Kalman Filter (Kalman, 1960) based tracking, which estimates the location and velocity of grape clusters, tends to outperform purely geometric based re-identification trackers on datasets where objects blend into the background.

(a) UAV RGB Video dataset

(b) WGISD dataset

(c) Lazio dataset

Fig. 12. Three datasets presenting unique challenges, impacting tracking performance: (*a*) Ninth annotated image in Row 4.2.1 in the Appendix (Sentís et al., 2021), (*b*) Annotated image CDY_2037 showing Chardonnay from the WGISD dataset (Santos et al., 2019), (*c*) Sixth annotated image from "CloseUp 2" in the Lazio dataset (Saraceni et al., 2024).

Fig. 12 shows example images from the three datasets that have been used to evaluate grape tracking in previous studies (Saraceni et al., 2024; Ariza-Sentís et al., 2023; Ciarfuglia et al., 2023).

Compared to this study, Ciarfuglia et al. (2023) achieve slightly higher MOTA scores of up to 46.7% and MOTP up to 72.9% on the WGISD dataset (Santos et al., 2019) using a Mask R-CNN model (He et al., 2017) with a mAP of 53.4% and DeepSORT (Wojke et al., 2017) as the tracking algorithm. The better performance of Mask R-CNN (He et al., 2017), with significantly higher mAP (53.4% vs. 19.4%), can likely be attributed to the larger grape clusters (0.2% to

7% of the frame area) and minimal foliage, which simplify detection. However, the close proximity of the clusters in the WGISD dataset (Santos et al., 2019) complicates the matching process during tracking and evaluation, which explains the modest improvement in tracking performance.

Saraceni et al. (2024) achieve MOTA scores up to 66.1% and IDF1 scores up to 73.7% on the Lazio dataset (Saraceni et al., 2024) using a YOLOv5 detector (Chen et al., 2022) and their AgriSORT (Saraceni et al., 2024) tracker. Although AgriSORT (Saraceni et al., 2024) performs exceptionally well on some subsets, it does not consistently outperform SORT (Bewley et al., 2016), which achieves MOTA scores up to 62.7%. In comparison, the higher MOTA scores (62.7% vs. 37.7%) are likely attributable to the dataset's (Saraceni et al., 2024) larger, more distinguishable red grape clusters which range from 0.3% to 20% of the frame area.

The Appendix (Sentís et al., 2021) used in this study contains at least 186 visible grape clusters, of which 147 are annotated in the test set. The grape clusters are small (0.01% to 0.17% of the frame area) and partially obscured in many cases. These challenges hinder the training of Mask R-CNN (He et al., 2017), which achieves 19.4% bbox mAP and 14.5% segmentation mAP. In particular, Mask R-CNN (He et al., 2017) shows better performances at lower IoU-thresholds, which are used to match detections with GTs during evaluation. This can be seen in Tables 3 and 4 in Section 3.1, where the model detects bboxes and masks with a mAP_{50} of 44.8% and 44.3% respectively, but only 14.1% and 4.5% at a mAP_{75}, using the 0.75 IoU-threshold. This discrepancy suggests that Mask R-CNN (He et al., 2017) regularly detects only parts of a grape cluster. However, these partial detections prove sufficient to track the corresponding grape cluster, which explains the comparably good tracking results of all algorithms, despite the low detection accuracy.

5.3. Limitations and further research

This section discusses the limitations of this study and proposes methods for future research to address these challenges.

DeepSORT (Wojke et al., 2017), SORT+ and DeepSORT+ achieve at least 95% counting accuracy, with DeepSORT+ counting the exact number of hand-counted visible grapes. However, these results should be interpreted with caution due to the unknown margin of error. For instance, it is improbable that DeepSORT+ maintains 100% counting accuracy for every other dataset. To confirm the results, bootstrapping could be utilized, as used for the tracking results (see Table 5 in Section 4, (Mooney et al., 1993)).

Previous studies by Orlandi et al. (2025) and Olenskyj et al. (2022) suggest that the counted grape clusters, together with the grape cluster mask area, are a reliable indicator for estimating grape yield. Using linear regression, various studies (Ahmedt-Aristizabal et al., 2024; Khokher et al., 2023; Olenskyj et al., 2022; Zabawa et al., 2022; Lopes and Cadima, 2021; Jaramillo et al., 2021) demonstrate a linear correlation between yield contributors, such as cluster count, cluster area, or berry count and the actual grape yield. By applying a linear regression function from one of these studies or an average of multiple correlations, the grape yield of the Appendix (Sentís et al., 2021) could be estimated. However, yield estimation could not be evaluated in this study, as the Appendix (Sentís et al., 2021) provides no data on grape weight or wine harvested from the vine rows.

As noted by Olenskyj et al. (2022), Jaramillo et al. (2021), grape bunch area is the most reliable individual yield indicator. Future research could reduce the occlusion error by using every track detection to potentially reconstruct the entire grape cluster mask. Furthermore, following Kierdorf et al. (2022), a generative adversarial network trained on defoliated vines can be used to estimate the number of occluded grapes.

The Appendix (Sentís et al., 2021) features four rows of "Vitis Vinifera" in its early ripening stage. The characteristics of these vine rows, such as color, size, camera perspective, and cluster proximity, vary depending on the maturity stage, grape variety, and specific environmental influences. Consequently, the detector and tracking algorithms fine-tuned for the Appendix (Sentís et al., 2021) are expected to perform worse on datasets with different characteristics.

To develop a grape tracking algorithm that achieves reliable results across all grape varieties, it is necessary to collect a comprehensive dataset that includes a large variety of grape species, colors, shapes, and ripening stages. Alternatively, future research could extend the work of Bellocchio et al. (2020), who successfully translate fruits in datasets into other fruits utilizing a C-GAN (Zhu et al., 2017). This approach could enable the training of a detector and tracker for multiple grape species using a dataset that contains only a few varieties.

For network training, the quality of the dataset is a deciding factor for the results. The Appendix (Sentís et al., 2021) presents three problems. First, it contains visible grape clusters in the background, which should not be counted. Adjusting the camera angle to exclude background grapes, using a depth sensor to distinguish between background and foreground (Wu et al., 2023), or capturing images at night with strong illumination (Olenskyj et al., 2022; Jaramillo et al., 2021) are potential solutions. Each of these approaches is either inconvenient or expensive. Hence, further research could explore the training of a network to differentiate foreground from background autonomously.

Second, the dataset contains visible grape clusters that were missed during annotation, complicating the training of reliable object detectors and thus limiting tracking performance.

Third, the grape clusters in the high-resolution images (2160×4096 pixels) represent only 0.01% to 0.17% of the frame area. Given the increasing use of UAVs in agriculture, with an annual market growth of 26% (GVR, 2025), and their ability to capture high-resolution videos, future object detectors should be able to capitalize on high resolutions. Since Mask R-CNN (He et al., 2017) is designed for resolutions of up to 1280×720 pixels, future research, such as Wu et al. (2021) should consider adapting Mask R-CNN (He et al., 2017) to handle higher resolutions or transition to newer architectures designed for small object detection in aerial imagery (Yuan et al., 2024).

Despite these limitations, the low detection accuracies of Mask R-CNN (He et al., 2017) (bbox mAP of 19.4% and segmentation mAP of 14.5%) are partially mitigated in tracking applications, because SORT (Bewley et al., 2016), DeepSORT (Wojke et al., 2017), ByteTrack (Zhang et al., 2022), SORT+ and DeepSORT+ maintain identification over intermittent detections.

This study introduces advanced versions of SORT (Bewley et al., 2016) and DeepSORT (Wojke et al., 2017), calling them SORT+ and DeepSORT+, respectively. Table 5 in Section 4 shows that SORT+ achieves significantly better results compared to the original SORT (Bewley et al., 2016) with MOTA increasing from 37.7% to 41.5% and ID switches decreasing from 73.5 to 28.2. DeepSORT+ exhibits minor improvements over its predecessor, reducing ID switches from 17.1 to 14.9. Using only the IoU-overlap for matching, ByteTrack (Zhang et al., 2022) achieves results comparable to SORT+ and approaches the performance of DeepSORT (Wojke et al., 2017) and DeepSORT+. Therefore, implementing and evaluating ByteTrack+ employing the same enhanced matching strategy as SORT+ and DeepSORT+ is worth investigating in future research.

6. Conclusion

This study introduces the enhanced tracking algorithms SORT+ and DeepSORT+. These trackers incorporate a novel matching process that leverages the complementary strengths of the IoU-overlap, the Mahalanobis distance (Mahalanobis, 1936) and the Euclidean distance to improve tracking performance and counting accuracy. To evaluate the extended matching algorithm SORT+ and DeepSORT+ are compared with SORT (Bewley et al., 2016), DeepSORT (Wojke et al., 2017) and ByteTrack (Zhang et al., 2022).

The enhanced matching process significantly reduces the ID switches in SORT (Bewley et al., 2016) from 74 to 28 in SORT+, improves the counting accuracy from 32.7% to 96.8% and greatly increases the tracking accuracy. This demonstrates the potential of the novel matching process, as it is easy to implement and does not require additional data, training, or extensive computation. Considering the complexity and additional requirements of the classification network for implementing DeepSORT(+), SORT+ presents a viable alternative with only a small decrease in tracking and counting accuracy.

This study demonstrates the feasibility of accurate grape tracking and counting in videos captured by UAVs, even at comparably large distances, despite a low detection accuracy. The uniform Mask R-CNN (He et al., 2017) detector achieves only a mAP of 19.4% due to the challenging nature of the Appendix (Sentís et al., 2021). All trackers could mostly maintain track identification over partial and missing detections. Furthermore, the trackers SORT+, DeepSORT (Wojke et al., 2017), and DeepSORT+ achieve over 95% counting accuracy, demonstrating their ability to estimate grape yield.

DeepSORT+, closely followed by DeepSORT (Wojke et al., 2017), achieves the highest MOTA score of 42.7% (±0.4%) and an IDF1 score of 67.1% (±0.2%), with only 15 ID switches over 186 tracks. However, it remains inconclusive which tracker consistently performs better on unseen similar data, due to the overlapping bootstrapped 95%-confidence intervals.

These findings not only align with, but also expand on current research in grape tracking (Saraceni et al., 2024; Ariza-Sentís et al., 2023; Ciarfuglia et al., 2023). As Ciarfuglia et al. (2023) use DeepSORT (Wojke et al., 2017) and Saraceni et al. (2024) compare SORT (Bewley et al., 2016), ByteTrack (Zhang et al., 2022) and StrongSORT (Du et al., 2023) (improved DeepSORT (Wojke et al., 2017)) and found that differences in tracking performance can largely be attributed to the characteristics of the dataset. Given that all current grape tracking research is focused solely on a single grape species, future research could be aimed at developing robust grape detection, tracking, and counting algorithms for various grape species. The performance increases of SORT+ and DeepSORT+ suggest the application of the extended matching process to other tracking algorithms such as ByteTrack (Zhang et al., 2022).

Accurate tracking and counting of small grape clusters that blend into the background is feasible using DeepSORT+, DeepSORT, (Wojke et al., 2017) or SORT+. Although DeepSORT+ demonstrates the best tracking and counting accuracy, SORT+ is the better choice, where the implementation of the classification network proves difficult. Using a UAV to count the number of visible grape clusters and determine their mask area offers a fast and inexpensive method for estimating grape yield. Therefore, the implications of this research extend beyond academic interest, offering practical insight to vineyard managers. By potentially improving yield estimation and non-destructive robotic agriculture, these findings support sustainable farming practices and economic growth.

CRediT authorship contribution statement

Jacob Piazolo: Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Benedikt Fischer:** Writing – review & editing, Supervision, Methodology, Conceptualization. **Robin Gruna:** Supervision, Funding acquisition. **Jürgen Beyerer:** Supervision, Funding acquisition.

Declaration of generative AI and AI-assisted technologies in the manuscript preparation process

During the preparation of this work the authors used Gemini 3 Pro in order to improve language and readability. After using this tool/service, the authors reviewed and edited the content as needed and take full responsibility for the content of the published article.

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.

Appendix A. UAV RGB Video dataset specifications

For the UAV RGB Video Dataset (Sentís et al., 2021) a drone captures the whole vine, therefore the green grape clusters take up merely 0.01% to 0.17% of the frame area. Furthermore, a dense canopy with significant occlusion makes detection a challenging task. Since no pruning nor leave removal was performed, this dataset reflects real world conditions for fast data collection (see Table 7).

Table 7
UAV RGB video dataset specifications.

Source	Dataset on UAV RGB videos acquired over a vineyard including bunch labels for object detection and tracking (Sentís et al., 2021)
UAV	DJI Matrice 210 RTK
Flight Speed	0.7 m/s
Flight Altitude	3 m above ground level
Frame Width	4096
Frame Height	2160
Frame Rate	59.94 frames/s
Annotated Frames	664
Train Split	528
Test Split	136
Data Collection	"The four flights were executed on June 28th, 2021 over four rows of the vineyards. The flights were carried out on a sunny day with wind velocity lower than 0.5 m/s. The four rows were selected according to the ripening stage of the grape clusters, to have a representative sample of the development status over the four rows. The grape bunch annotations were labeled using the CVAT software and are available in the MOTS style" (Sentís et al., 2021).
Vineyard Location	41° 57'18.3"N 8° 47'41.9"W Tomiño, Spain

Appendix B. List of Abbreviations

See Table 8.

Table 8
List of abbreviations.

Abbreviation	Definition
AG	Annotated Grapes
b	background
bbox	bounding box
COCO	Common Objects in Context
CVAT	Computer Vision Annotation Tool
DeepSORT	Simple Online and Realtime Tracking with a Deep Association Metric
f	foreground
FN	False Negative
FP	False Positive
FPN	Feature Pyramid Network
GNSS	Global Navigation Satellite System
HC	Hand Counted
IDF1	Identification F1 Score
IDFN	False Negative Identities
IDFP	False Positive Identities
IDP	Identification Precision
IDR	Identification Recall
IDTP	True Positive Identities
IDsw	ID-switches
IoU	Intersection over Union
IR	Induction Requirement

(continued on next page)

Table 8 (*continued*).

Abbreviation	Definition
MAE	Mean Average Error
mAP	mean Average Precision
MOTA	Multi Object Tracking Accuracy
MOTP	Multi Object Tracking Precision
MOTS	Multi Object Tracking and Segmentation
MOTSA	Multi Object Tracking and Segmentation Accuracy
MOTSP	Multi Object Tracking and Segmentation Precision
pos. dif.	positive definite
R-CNN	Region Based Convolutional Neural Network
ReLU	Rectified Linear Unit
ResNet	Residual Network
RoI	Region of Interest
RPN	Region Proposal Network
SfM	Structure from Motion
SORT	Simple Online and Realtime Tracking
TP	True Positive
UAV	Unmanned Aerial Vehicle

Data availability

The code is publicly accessible on GitHub (https://github.com/Jax377/Grape-Detection-and-Tracking.git).

References

Aharon, N., Orfaig, R., Bobrovsky, B.-Z., 2022. [Preprint] BoT-SORT: Robust associations multi-pedestrian tracking. http://dx.doi.org/10.48550/ARXIV.2206.14651, arXiv preprint arXiv:2206.14651, URL: https://arxiv.org/abs/2206.14651, Publisher: arXiv Version Number: 2.

Ahmedt-Aristizabal, D., Smith, D., Khokher, M.R., Li, X., Smith, A.L., Petersson, L., Rolland, V., Edwards, E.J., 2024. An in-field dynamic vision-based analysis for vineyard yield estimation. IEEE Access 12, 102146–102166. http://dx.doi.org/10.1109/ACCESS.2024.3431244, URL: https://ieeexplore.ieee.org/document/10604813.

Aquino, A., Millan, B., Diago, M.-P., Tardaguila, J., 2018. Automated early yield prediction in vineyards from on-the-go image acquisition. Comput. Electron. Agric. 144, 26–36. http://dx.doi.org/10.1016/j.compag.2017.11.026, URL: https://www.sciencedirect.com/science/article/pii/S0168169917303964.

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, URL: https://www.sciencedirect.com/science/article/pii/S0168169923004398.

Arlotta, A., Lippi, M., Gasparri, A., 2023. An EKF-based multi-object tracking framework for a mobile robot in a precision agriculture scenario. In: 2023 European Conference on Mobile Robots. ECMR, pp. 1–6. http://dx.doi.org/10.1109/ECMR59166.2023.10256338, URL: https://ieeexplore.ieee.org/abstract/document/10256338, ISSN: 2767-8733.

Bellocchio, E., Costante, G., Cascianelli, S., Fravolini, M.L., Valigi, P., 2020. Combining domain adaptation and spatial consistency for unseen fruits counting: A quasi-unsupervised approach. IEEE Robot. Autom. Lett. 5 (2), 1079–1086. http://dx.doi.org/10.1109/LRA.2020.2966398, URL: https://ieeexplore.ieee.org/document/8957569, Conference Name: IEEE Robotics and Automation Letters.

Bewley, A., Ge, Z., Ott, L., Ramos, F., Upcroft, B., 2016. Simple online and realtime tracking. In: 2016 IEEE International Conference on Image Processing. ICIP, pp. 3464–3468. http://dx.doi.org/10.1109/ICIP.2016.7533003, URL: http://arxiv.org/abs/1602.00763, arXiv:1602.00763 [cs].

Blekos, A., Chatzis, K., Kotaidou, M., Chatzis, T., Solachidis, V., Konstantinidis, D., Dimitropoulos, K., 2024. CERTH grape dataset, http://dx.doi.org/10.5281/zenodo.10777647, https://zenodo.org/records/10777647. Zenodo (Version v1), Zenodo.

Bouraya, S., Belangour, A., 2021. Deep learning based neck models for object detection: A review and a benchmarking study. Int. J. Adv. Comput. Sci. Appl. 12 (11), http://dx.doi.org/10.14569/IJACSA.2021.0121119, URL: http://thesai.org/Publications/ViewPaper?Volume=12&Issue=11&Code=IJACSA&SerialNo=19.

Cao, J., Pang, J., Weng, X., Khirodkar, R., Kitani, K., 2023. Observation-centric SORT: Rethinking SORT for robust multi-object tracking. In: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. CVPR, IEEE, Vancouver, BC, Canada, pp. 9686–9696. http://dx.doi.org/10.1109/CVPR52729.2023.00934, URL: https://ieeexplore.ieee.org/document/10204818/.

Chen, Y., Yang, J., Wang, J., Zhou, X., Zou, J., Li, Y., 2022. An improved YOLOv5 real-time detection method for aircraft target detection. In: 2022 27th International Conference on Automation and Computing. ICAC, pp. 1–6. http://dx.doi.org/10.1109/ICAC55051.2022.9911114.

Ciarfuglia, T.A., Motoi, I.M., Saraceni, L., Fawakherji, M., Sanfeliu, A., Nardi, D., 2023. Weakly and semi-supervised detection, segmentation and tracking of table grapes with limited and noisy data. Comput. Electron. Agric. 205, 107624. http://dx.doi.org/10.1016/j.compag.2023.107624, URL: https://www.sciencedirect.com/science/article/pii/S0168169923000121.

Du, Y., Zhao, Z., Song, Y., Zhao, Y., Su, F., Gong, T., Meng, H., 2023. StrongSORT: Make DeepSORT great again. IEEE Trans. Multimed. 25, 8725–8737. http://dx.doi.org/10.1109/TMM.2023.3240881, URL: https://ieeexplore.ieee.org/abstract/document/10032656?casa_token=BnHb7NfaLIgAAAAA:Xi6Q2ITEzlTB9tQEmfB1y18MTVT8X_1yVxMr6K8VgF4Nh1nGnPIDWuGkvSzZrLHaH82JJ-vRQiQ, Conference Name: IEEE Transactions on Multimedia.

Evangelidis, G.D., Psarakis, E.Z., 2008. Parametric image alignment using enhanced correlation coefficient maximization. IEEE Trans. Pattern Anal. Mach. Intell. 30 (10), 1858–1865. http://dx.doi.org/10.1109/TPAMI.2008.113, URL: https://ieeexplore.ieee.org/document/4515873, Conference Name: IEEE Transactions on Pattern Analysis and Machine Intelligence.

Fang, W., Sun, X., Wu, Z., Zhao, G., Li, G., Li, R., Fu, L., Zhang, Q., 2022. A novel apple fruit detection and counting methodology based on deep learning and trunk tracking in modern orchard. Comput. Electron. Agric. 197, 107000. http://dx.doi.org/10.1016/j.compag.2022.107000, URL: https://www.sciencedirect.com/science/article/pii/S0168169922003179.

GVR, 2025. Agriculture drones market size, share | industry report 2033. URL: https://www.grandviewresearch.com/industry-analysis/agriculture-drones-market.

Hartley, R., Zisserman, A., 2003. Multiple View Geometry in Computer Vision. Cambridge University Press, Google-Books-ID, si3R3Pfa98QC.

He, K., Gkioxari, G., Dollár, P., Girshick, R.B., 2017. Mask R-CNN. In: 2017 IEEE International Conference on Computer Vision. ICCV, URL: https://www.semanticscholar.org/paper/Mask-R-CNN-He-Gkioxari/84d440eac8a3fb52ea5708e4943d02fc3fcfe009.

He, L., Wu, F., Du, X., Zhang, G., 2022. Cascade-SORT: A robust fruit counting approach using multiple features cascade matching. Comput. Electron. Agric. 200, 107223. http://dx.doi.org/10.1016/j.compag.2022.107223, URL: https://www.sciencedirect.com/science/article/pii/S0168169922005373.

Henriques, J.F., Caseiro, R., Martins, P., Batista, J., 2015. High-speed tracking with kernelized correlation filters. IEEE Trans. Pattern Anal. Mach. Intell. 37 (3), 583–596. http://dx.doi.org/10.1109/TPAMI.2014.2345390, URL: https://ieeexplore.ieee.org/abstract/document/6870486?casa_token=T0wgcd63t0MAAAAA:tjMMc-8rC7kXtd1I2ZfEBk5MA4P7kXZdt8OsCjTBLpdkHOC5sE0AMx32fNvrXh-Bd7aQYLKbOFA, Conference Name: IEEE Transactions on Pattern Analysis and Machine Intelligence.

Hsiao, T.-Y., Chang, Y.-C., Chou, H.-H., Chiu, C.-T., 2019. Filter-based deep-compression with global average pooling for convolutional networks. J. Syst. Archit. 95, 9–18. http://dx.doi.org/10.1016/j.sysarc.2019.02.008, URL: https://www.sciencedirect.com/science/article/pii/S1383762118302340.

Jaramillo, J., Vanden Heuvel, J., Petersen, K.H., 2021. Low-cost, computer vision-based, prebloom cluster count prediction in vineyards. Front. Agron. 3, URL: https://www.frontiersin.org/articles/10.3389/fagro.2021.648080.

Jégou, H., Douze, M., Schmid, C., Pérez, P., 2010. Aggregating local descriptors into a compact image representation. In: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. pp. 3304–3311. http://dx.doi.org/10.1109/CVPR.2010.5540039, URL: https://ieeexplore.ieee.org/abstract/document/5540039, ISSN: 1063-6919.

Kalman, R.E., 1960. A new approach to linear filtering and prediction problems. J. Basic Eng. 82 (1), 35–45. http://dx.doi.org/10.1115/1.3662552.

Khokher, M.R., Liao, Q., Smith, A.L., Sun, C., Mackenzie, D., Thomas, M.R., Wang, D., Edwards, E.J., 2023. Early yield estimation in viticulture based on grapevine inflorescence detection and counting in videos. IEEE Access 11, 37790–37808. http://dx.doi.org/10.1109/ACCESS.2023.3263238, URL: https://ieeexplore.ieee.org/document/10087292/.

Kierdorf, J., Weber, I., Kicherer, A., Zabawa, L., Drees, L., Roscher, R., 2022. Behind the leaves: Estimation of occluded grapevine berries with conditional generative adversarial networks. Front. Artif. Intell. 5, URL: https://www.frontiersin.org/articles/10.3389/frai.2022.830026.

Kingma, D.P., Ba, J., 2014. Adam: A method for stochastic optimization. CoRR, URL: https://www.semanticscholar.org/paper/Adam%3A-A-Method-for-Stochastic-Optimization-Kingma-Ba/a6cb366736791bcccc5c8639de5a8f9636bf87e8.

Kirk, R., Mangan, M., Cielniak, G., 2021. Robust counting of soft fruit through occlusions with re-identification. In: Vincze, M., Patten, T., Christensen, H.I., Nalpantidis, L., Liu, M. (Eds.), Computer Vision Systems. Springer International Publishing, Cham, pp. 211–222. http://dx.doi.org/10.1007/978-3-030-87156-7_17.

Koonce, B., 2021. ResNet 50. In: Koonce, B. (Ed.), Convolutional Neural Networks with Swift for Tensorflow: Image Recognition and Dataset Categorization. A Press, Berkeley, CA, pp. 63–72. http://dx.doi.org/10.1007/978-1-4842-6168-2_6.

Kuhn, H.W., 1955. The Hungarian method for the assignment problem. Nav. Res. Logist. Q. 2 (1–2), 83–97. http://dx.doi.org/10.1002/nav.3800020109, URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/nav.3800020109, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/nav.3800020109.

Kumar, A., 2025. Advancements in computer vision: Bridging perception and intelligence. Acta Sci. Comput. Sci. 7, 1, URL: https://actascientific.com/ASCS/pdf/ASCS-07-0572.pdf.

Li, Y., Bao, Z., Qi, J., 2022. Seedling maize counting method in complex backgrounds based on YOLOV5 and Kalman filter tracking algorithm. Front. Plant Sci. 13, URL: https://www.frontiersin.org/journals/plant-science/articles/10.3389/fpls.2022.1030962.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L., 2014. Microsoft COCO: Common objects in context. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (Eds.), In: Computer Vision – ECCV 2014, vol. 8693, Springer International Publishing, Cham, pp. 740–755. http://dx.doi.org/10.1007/978-3-319-10602-1_48, URL: http://link.springer.com/10.1007/978-3-319-10602-1_48, Series Title: Lecture Notes in Computer Science.

Lopes, C., Cadima, J., 2021. Grapevine bunch weight estimation using image-based features: comparing the predictive performance of number of visible berries and bunch area. OENO One 55 (4), 209–226. http://dx.doi.org/10.20870/oeno-one.2021.55.4.4741, URL: https://oeno-one.eu/article/view/4741, Number: 4.

Lucas, B.D., Kanade, T., 1981. An iterative image registration technique with an application to stereo vision. In: IJCAI'81: 7th International Joint Conference on Artificial Intelligence, vol. 2, Vancouver, Canada, pp. 674–679, URL: https://hal.science/hal-03697340.

Mahalanobis, P.C., 1936. On the generalized distance in statistics. Proc. Natl. Inst. Sci. India URL: http://localhost:8080/xmlui/handle/10263/6765, Accepted: 2017-04-03T08:54:00Z Publisher: National Institute of Science of India.

Maier, G., Pfaff, F., Pieper, C., Gruna, R., Noack, B., Kruggel-Emden, H., Längle, T., Hanebeck, U.D., Wirtz, S., Scherer, V., Beyerer, J., 2021. Experimental evaluation of a novel sensor-based sorting approach featuring predictive real-time multiobject tracking. IEEE Trans. Ind. Electron. 68 (2), 1548–1559. http://dx.doi.org/10.1109/TIE.2020.2970643.

Mooney, C.Z., Duval, R.D., Duvall, R., 1993. Bootstrapping: A Nonparametric Approach to Statistical Inference. SAGE, Google-Books-ID, ZxaRC4I2z6sC.

Morros, J.R., Lobo, T.P., Salmeron-Majadas, S., Villazan, J., Merino, D., Antunes, A., Datcu, M., Karmakar, C., Guerra, E., Pantazi, D.-A., Stamoulis, G., 2021. AI4agriculture grape dataset, (Version v1). http://dx.doi.org/10.5281/zenodo.5660081, https://zenodo.org/records/5660081. Zenodo.

Olenskyj, A.G., Sams, B.S., Fei, Z., Singh, V., Raja, P.V., Bornhorst, G.M., Earles, J.M., 2022. End-to-end deep learning for directly estimating grape yield from ground-based imagery. Comput. Electron. Agric. 198, 107081. http://dx.doi.org/10.1016/j.compag.2022.107081, URL: https://www.sciencedirect.com/science/article/pii/S0168169922003982.

Orlandi, G., Matese, A., Ulrici, A., Calvini, R., Berton, A., Gennaro, S.F.D., 2025. Automated yield prediction in vineyard using RGB images acquired by a UAV prototype platform. OENO One 59 (1), http://dx.doi.org/10.20870/oeno-one.2025.59.1.8133, URL: https://oeno-one.eu/article/view/8133.

Poblete-Echeverria, C., Berry, A., Venter, T., Velez, S., Pavez, M.I.G., Iñiguez, R., 2025. Morphological image analysis for estimating grape bunch weight under different irrigation regimes in Cabernet-Sauvignon: This article is part of the special issue of the GiESCO 2025 meeting. OENO One 59 (2), http://dx.doi.org/10.20870/oeno-one.2025.59.2.9309, URL: https://oeno-one.eu/article/view/9309.

Ren, S., He, K., Girshick, R., Sun, J., 2015. Faster R-CNN: Towards real-time object detection with region proposal networks. In: Advances in Neural Information Processing Systems, vol. 28, Curran Associates, Inc., pp. 91–99, URL: https://proceedings.neurips.cc/paper_files/paper/2015/hash/14bfa6bb14875e45bba028a21ed38046-Abstract.html.

Santos, T., de Souza, L., Andreza, d.S., Avila, S., 2019. Embrapa wine grape instance segmentation dataset – Embrapa WGISD, http://dx.doi.org/10.5281/zenodo.3361736, https://zenodo.org/records/3361736. Zenodo.

Santos, T.T., de Souza, L.L., dos Santos, A.A., Avila, S., 2020. Grape detection, segmentation, and tracking using deep neural networks and three-dimensional association. Comput. Electron. Agric. 170, 105247. http://dx.doi.org/10.1016/j.compag.2020.105247, URL: https://www.sciencedirect.com/science/article/pii/S0168169919315765.

Saraceni, L., Motoi, I.M., Nardi, D., Ciarfuglia, T.A., 2024. AgriSORT: A simple online real-time tracking-by-detection framework for robotics in precision agriculture. In: 2024 IEEE International Conference on Robotics and Automation. ICRA, pp. 2675–2682. http://dx.doi.org/10.1109/ICRA57147.2024.10610231, URL: https://ieeexplore.ieee.org/document/10610231/.

Seng, K.P., Ang, L.-M., Schmidtke, L.M., Rogiers, S.Y., 2018. Grape image database, Version v1. http://dx.doi.org/10.26189/5da7a8603c55c, https://researchoutput.csu.edu.au/en/datasets/grape-image-database/. Charles Sturt University Research Output.

Sentís, M.A., Vélez, S., Valente, J., 2021. Dataset on UAV RGB videos acquired over a vineyard property of Bodegas Terras Gauda at an early stage of Botrytis cinerea infection in 2021, (Version v1). http://dx.doi.org/10.5281/zenodo.7330951, https://zenodo.org/records/7330951. Zenodo.

Shen, L., Su, J., He, R., Song, L., Huang, R., Fang, Y., Song, Y., Su, B., 2023. Real-time tracking and counting of grape clusters in the field based on channel pruning with YOLOv5s. Comput. Electron. Agric. 206, 107662. http://dx.doi.org/10.1016/j.compag.2023.107662, URL: https://www.sciencedirect.com/science/article/pii/S0168169923000509.

Sozzi, M., Cantalamessa, S., Cogato, A., Kayad, A., Marinello, F., 2022. wGrapeUNIPD-DL: an open dataset for white grape bunch detection, (Version v1). http://dx.doi.org/10.5281/zenodo.4066730, https://zenodo.org/records/4066730. Zenodo.

Statista, 2025. Wine - Worldwide | Statista market forecast. Statista, URL: https://www.statista.com/outlook/cmo/alcoholic-drinks/wine/worldwide.

Stein, M., Bargoti, S., Underwood, J., 2016. Image based mango fruit detection, localisation and yield estimation using multiple view geometry. Sensors 16 (11), 1915. http://dx.doi.org/10.3390/s16111915, URL: https://www.mdpi.com/1424-8220/16/11/1915, Number: 11 Publisher: Multidisciplinary Digital Publishing Institute.

Wang, Z., Walsh, K., Koirala, A., 2019. Mango fruit load estimation using a video based MangoYOLO—Kalman Filter—Hungarian algorithm method. Sensors 19 (12), 2742. http://dx.doi.org/10.3390/s19122742, URL: https://www.mdpi.com/1424-8220/19/12/2742, Number: 12 Publisher: Multidisciplinary Digital Publishing Institute.

Wang, Z., Zheng, L., Liu, Y., Li, Y., Wang, S., 2020. Towards real-time multi-object tracking. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (Eds.), In: Computer Vision – ECCV 2020, vol. 12356, Springer International Publishing, Cham, pp. 107–122. http://dx.doi.org/10.1007/978-3-030-58621-8_7, URL: https://link.springer.com/10.1007/978-3-030-58621-8_7, Series Title: Lecture Notes in Computer Science.

Wojke, N., Bewley, A., Paulus, D., 2017. Simple online and realtime tracking with a deep association metric. In: 2017 IEEE International Conference on Image Processing. ICIP, pp. 3645–3649. http://dx.doi.org/10.1109/ICIP.2017.8296962, URL: https://ieeexplore.ieee.org/document/8296962/, ISSN: 2381-8549.

Wu, Q., Feng, D., Cao, C., Zeng, X., Feng, Z., Wu, J., Huang, Z., 2021. Improved Mask R-CNN for aircraft detection in remote sensing images. Sensors 21 (8), 2618. http://dx.doi.org/10.3390/s21082618, URL: https://www.mdpi.com/1424-8220/21/8/2618, Number: 8 Publisher: Multidisciplinary Digital Publishing Institute.

Wu, Z., Sun, X., Jiang, H., Mao, W., Li, R., Andriyanov, N., Soloviev, V., Fu, L., 2023. NDMFCS: An automatic fruit counting system in modern apple orchard using abatement of abnormal fruit detection. Comput. Electron. Agric. 211, 108036. http://dx.doi.org/10.1016/j.compag.2023.108036, URL: https://www.sciencedirect.com/science/article/pii/S0168169923004246.

Xie, Y., Wang, H., Lu, Y., 2024. Cascaded-scoring tracklet matching for multi-object tracking. In: Liu, Q., Wang, H., Ma, Z., Zheng, W., Zha, H., Chen, X., Wang, L., Ji, R. (Eds.), Pattern Recognition and Computer Vision. Springer Nature, Singapore, pp. 161–173. http://dx.doi.org/10.1007/978-981-99-8549-4_14.

Xu, Z., Yang, W., Zhang, W., Tan, X., Huang, H., Huang, L., 2022. Segment as points for efficient and effective online multi-object tracking and segmentation. IEEE Trans. Pattern Anal. Mach. Intell. 44 (10), 6424–6437. http://dx.doi.org/10.1109/TPAMI.2021.3087898, URL: https://ieeexplore.ieee.org/document/9449985.

Yuan, Z., Gong, J., Guo, B., Wang, C., Liao, N., Song, J., Wu, Q., 2024. Small object detection in UAV remote sensing images based on intra-group multi-scale fusion attention and adaptive weighted feature fusion mechanism. Remote. Sens. 16 (22), 4265. http://dx.doi.org/10.3390/rs16224265, URL: https://www.mdpi.com/2072-4292/16/22/4265, Publisher: Multidisciplinary Digital Publishing Institute.

Zabawa, L., Kicherer, A., Klingbeil, L., Töpfer, R., Roscher, R., Kuhlmann, H., 2022. Image-based analysis of yield parameters in viticulture. Biosyst. Eng. 218, 94–109. http://dx.doi.org/10.1016/j.biosystemseng.2022.04.009, URL: https://www.sciencedirect.com/science/article/pii/S1537511022000861.

Zhai, Y., Zhang, L., Hu, X., Yang, F., Huang, Y., 2025. A dynamic Kalman filtering method for multi-object fruit tracking and counting in complex orchards. Sensors 25 (13), 4138. http://dx.doi.org/10.3390/s25134138, URL: https://www.mdpi.com/1424-8220/25/13/4138, Publisher: Multidisciplinary Digital Publishing Institute.

Zhang, X., Lu, S., Karkee, M., Zhang, Q., 2020. Full stages of wine grape canopy and clusters. Washington State University, http://dx.doi.org/10.7273/000001846. Washington State University Research Exchange.

Zhang, Y., Sun, P., Jiang, Y., Yu, D., Weng, F., Yuan, Z., Luo, P., Liu, W., Wang, X., 2022. ByteTrack: Multi-object tracking by associating every detection box. In: Avidan, S., Brostow, G., Cissé, M., Farinella, G.M., Hassner, T. (Eds.), In: Computer Vision – ECCV 2022, vol. 13682, Springer Nature Switzerland, Cham, pp. 1–21. http://dx.doi.org/10.1007/978-3-031-20047-2_1, URL: https://link.springer.com/10.1007/978-3-031-20047-2_1, Series Title: Lecture Notes in Computer Science.

Zhu, J.-Y., Park, T., Isola, P., Efros, A.A., 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE International Conference on Computer Vision. ICCV, pp. 2242–2251. http://dx.doi.org/10.1109/ICCV.2017.244, URL: https://openaccess.thecvf.com/content_iccv_2017/html/Zhu_Unpaired_Image-To-Image_Translation_ICCV_2017_paper.html.

Zou, Z., Hao, J., Shu, L., 2022. Rethinking bipartite graph matching in realtime multi-object tracking. In: 2022 Asia Conference on Algorithms, Computing and Machine Learning. CACML, pp. 713–718. http://dx.doi.org/10.1109/CACML55074.2022.00124, URL: https://ieeexplore.ieee.org/abstract/document/9852565?casa_token=R1I15Yh97xUAAAAA:lT_xi0_xAZJBhmTmLsm4VMbwzV4RrkgkhmewwX-HM8eCUEP89Tb8M1IBbJyMrsV-PFpPnhIN.