Multi-Object Tracking in Agricultural Applications using a Vision Transformer for Spatial Association

Byron Hernandez[a], Henry Medeiros[a]

[a]*Department of Agricultural and Biological Engineering, University of Florida, Gainesville, 32611-0570, Florida, U.S.A.*

Abstract

This paper introduces a Multi-Object Tracking (MOT) framework for agricultural applications that estimates global positions in pixel coordinates using the local feature matching transformer – LoFTR. We design an efficient tracker that augments the capabilities of a state-of-the-art tracking algorithm by incorporating a novel association strategy based on spatial information of targets leaving and returning the camera field of view. We evaluate our framework using the publicly available LettuceMOT benchmark dataset and an adapted version of the AppleMOTS benchmark dataset that we denominate AppleMOT. Our experimental results demonstrate that our method outperforms cutting-edge algorithms for robotic plant tracking in the LettuceMOT dataset. The evaluation metrics show average improvements of up to 25% compared to the best publicly available results, demonstrating the benefits of our spatial association approach. For the AppleMOT dataset, we obtained bounding-box-based MOT evaluation metrics comparable to the segmentation-based (MOTS) counterparts presented in the original AppleMOTS paper. These findings highlight the effectiveness and potential of our approach in addressing the unique challenges posed by agricultural environments.

1. Introduction

The increasing global population [1] and decreasing availability of agricultural workers [2] highlight the importance of addressing the critical issue of food production. To meet escalating demands for food while ensuring improved quality and efficiency throughout supply chains, the agricultural industry has embraced the transformative potential of Industry 4.0 [3]. Specifically, integrating robotics and Artificial Intelligence (AI) within this framework is revolutionizing agricultural processes. Such processes include detecting fruits, plants, flowers, and weeds, as well as classifying related diseases [4, 5]. Plant phenotyping [6, 7, 8], irrigation [9], and harvesting [10] are also related areas of current research of great significance in agricultural research and production management [11].

In this evolving landscape, computer vision has emerged as a critical component in the sensing and perception mechanisms of robotic systems that perform agricultural tasks. Visual perception plays a vital role in precision agriculture and related applications involving human-robot interaction [12, 13] as it enables the detection, segmentation, and tracking of plants, flowers, fruits, and humans [14]. Multiple object tracking (MOT) is a remarkably complex but essential task for advanced autonomous systems.

MOT is an active research area in computer vision with complex challenges and broad applicability across various sectors, including urban surveillance, autonomous driving, and agriculture. Recent methods such as Simple Online and Real-Time Tracking (SORT) [15] and DeepSORT [16] integrate advanced data association techniques with features obtained using deep learning strategies. These approaches have shown outstanding

performance in MOT problems in urban and human-centered scenarios. Improvements in motion prediction mechanisms, such as those proposed in CenterTrack [17], which focuses on predicting object displacements, and in ByteTrack [18], which aims to minimize identity switches, further contributed to significant recent advances in MOT techniques applied to urban and human-centered scenarios [19, 20]. Recently, segmentation-based approaches, such as TrackRCNN [21] and PointTrack [22], introduced the ability to precisely delineate objects of interest and keep track of their identity over time. Monitoring objects at the level of individual pixels enables a more accurate description of the physical characteristics of objects of interest, such as their shapes and sizes [23]. It also increases the robustness of the tracker against temporary occlusions, which are very common in agricultural research and production environments [24]. Unfortunately, the training data requirements of current segmentation and tracking techniques still prevent their broad applicability to a variety of agricultural scenarios [25].

Applying MOT techniques to agricultural settings introduces challenges not typically encountered in urban environments or other human-centered applications. Agricultural settings are characterized by high object homogeneity since crops and plants exhibit limited distinctive visual features, making tracking individual objects particularly challenging [26]. The dynamic nature of these environments, compounded by factors such as varying illumination conditions, weather changes, and the inherent clutter of natural scenes, further complicates the reliable detection and tracking of relevant objects in agricultural contexts [27].

Accurately tracking fruits and plants in the agricultural sector is crucial for automating chemical spraying and yield estimation [28]. These applications are highly dependent on the precision of tracking systems, as they directly influence the efficiency and effectiveness of resource use in farming operations. Fruit counting is fundamental for yield estimation, which informs farmers in decisions about crop management and market supply chains [11]. Reliable MOT systems can automate the counting process, ensuring that each fruit is accurately accounted for, which is vital for economic planning and reducing waste [26]. For automated spraying, accurate tracking ensures that each plant receives the correct treatment without unnecessary applications of pesticides and nutrients. The application of adequate amounts of chemicals optimizes productivity, reduces runoff, and mitigates the environmental impact of agricultural operations [9]. Efficient automated systems ensure that each plant or area is sprayed or treated exactly once, maximizing coverage and minimizing waste [11]. These operations require the precise localization of each fruit and plant and consistent identification over time, even when plants are occluded or temporarily move out of the field of view (FOV) of the camera.

The availability of relevant datasets for agricultural applications is limited. In the 2020 survey by Lu and Young [29], only 34 datasets were identified for image classification, object detection, and segmentation in agricultural scenarios. Joshi et al. [30] noted that the majority of the agricultural datasets in their 2023 study were sourced from Lu and Young's work [29]. Furthermore, only two publicly available datasets are explicitly designed for MOT in agricultural scenarios: AppleMOTS [26] and LettuceMOT [27]. These datasets provide annotated trajectory identifiers for apples and lettuces, respectively.

The LettuceMOT dataset is particularly relevant as it captures diverse scenarios related to autonomous robotic navigation in agricultural fields. These scenarios include dynamic obstacle avoidance and actions such as leaving and returning to the working area to recharge batteries or refill chemicals for spraying. In addition to the LettuceMOT dataset, in [27], the authors present a benchmark evaluation of several tracking algorithms for lettuce tracking, including cutting-edge techniques such as SORT [15] and ByteTrack [18].

The AppleMOTS dataset addresses the notable gap in datasets for homogeneous agricultural objects, where similar appearance and environmental conditions complicate tracking tasks [26]. This dataset, which contains manually annotated segmentation masks of apples captured using an

uncrewed aerial vehicle (UAV) and wearable sensor platforms, is the first resource for developing and testing MOTS algorithms in agricultural scenarios. In addition to encouraging research in this relevant field, the dataset highlights the increased difficulty posed by the homogeneity and dense clustering typical of orchard settings.

To address the specific challenges exhibited by datasets such as LettuceMOT, the LettuceTrack model [31] uses the spatial relationships among objects to monitor their trajectory. This approach exploits the stationary nature of plants in scenarios featuring linear robot movement, guiding tracking by the central axis of the orchard's layout. This approach results in a significant improvement in tracking accuracy. However, it faces difficulties capturing the spatial relations among different plants when the robot does not follow a straight-line trajectory. This limitation hinders the correct association and longer-term re-association of plants or fruits.

Given the complexities highlighted above, there is a critical need to develop MOT systems specifically tailored to handle the unpredictable elements of agricultural environments. The objective of this work was to devise an MOT method capable of keeping track of the identities of multiple agricultural objects of interest, such as plants and fruits, even when those go out of and later return to the camera's field of view, without losing generality in the broader MOT context.

Therefore, we propose a novel tracking method for agricultural settings. We design a tracking-by-detection algorithm based on the bounding box regression technique for tracking introduced in [32], which we denominate FloraTracktor. We also incorporate into our tracker a novel spatial association module, providing a flexible and efficient tracking framework. The proposed approach builds upon the Local Feature Matching Transformer (LoFTR) introduced in [33] to perform object association and longer term re-association. In contrast to the extended Tracktor++ [32], which relies on separate modules for camera alignment, motion updates, and re-identification (reID), our method adopts LoFTR to build a unified module that seamlessly handles alignment, motion, and re-association. As a result, our improved algorithm eliminates the need for tracking annotations, outperforming Tracktor, which does not require tracking annotations either, and Tracktor++, which relies on temporally consistent ID labels for training. This paper presents the following key contributions:

- We introduce a robust tracking-by-detection framework for effective plant tracking.
- The framework incorporates a spatial association module that leverages spatial relationships among static objects to estimate the robot's motion.
- A comprehensive experimental evaluation establishes a new benchmark performance level for the LettuceMOT dataset and the AppleMOT dataset. AppleMOT is an adaptation of AppleMOTS for the task of bounding-box-based MOT. The performance gains shown by the proposed framework highlight the potential for substantial advancements in plant tracking accuracy and precision.
- Our framework is publicly available in the following repository: ag-tracking

The remainder of this paper is organized as follows. Section 2 provides a detailed description of our tracking and association algorithms. Section 3 first describes the experimental setup based on the protocols proposed in [27, 31] and a set of benchmark metrics for the AppleMOT dataset before presenting the performance of our approach in comparison with existing methods. Finally, Section 4 concludes the paper.

2. Materials and Methods

Figure 1 illustrates our tracking framework. It comprises two main building blocks: a tracking-by-detection algorithm and a spatial association module. The tracking algorithm draws inspiration from the paradigm introduced in [32]. Our tracker handles occlusions, false positives, and false negatives by

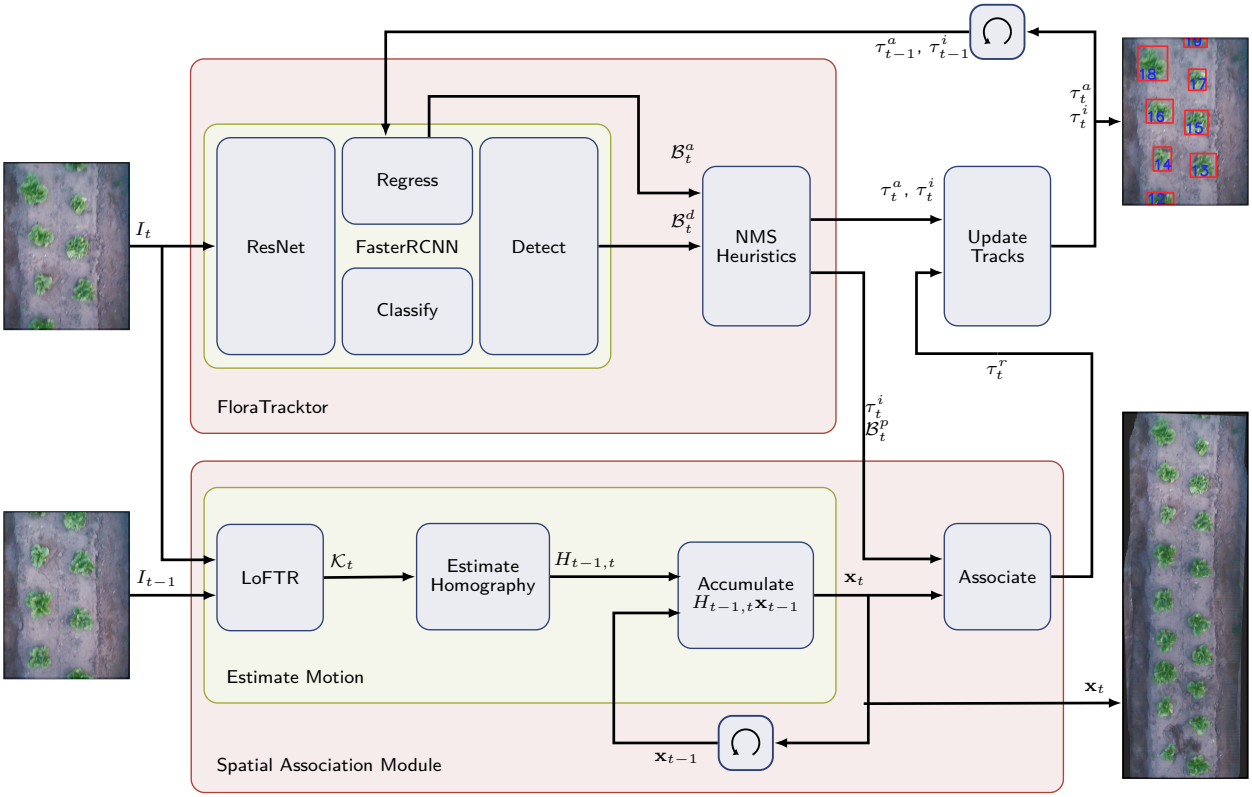

Figure 1: Our framework performs real-time video tracking using input frames I_t and I_{t-1}. The FloraTracktor module updates active tracks, denoted as τ^a_t. The Spatial Association Module carries out image matching between consecutive frames, estimates a global transformation $H_{t-1,t}$ used to update the global coordinates $\mathbf{x}_t$, and finally uses the estimated global position $\mathbf{x}_t$ to associate potential new tracks in $\mathcal{B}^p_t$ with inactive tracks in τ^i_t.

comparing retained and suppressed bounding boxes in a single non-maximum suppression (NMS) step. The second component is a novel association module tailored for agricultural applications that leverages spatial relationships among plants. The association module provides a global position estimate of each object of interest, allowing for longer-term re-association even when objects go out of and back into the camera's FOV. Our tracking framework is denominated FloraTracktor+. Algorithm 1 presents the global pipeline of the approach. Please refer to Appendix A for a comprehensive description of our notation. The following sections describe the two main components of our method in details.

2.1. FloraTracktor

As described in [32], Tracktor uses a detector from the region-based convolutional neural network family (R-CNN [34]). Specifically, it uses the FasterRCNN model [35] as its base detector. It employs the region of interest (ROI) regression head [36] to predict the bounding boxes of active tracks τ^a_t given the latent features of the input image at time t and the bounding boxes of the active tracks in the previous frame τ^a_{t-1}. It then uses the set of regressed bounding boxes $\mathcal{B}^a_t = \{\mathbf{b}^{a_1}_t, \mathbf{b}^{a_2}_t, \cdots\}$ and the set of new detections $\mathcal{B}^d_t = \left\{\mathbf{b}^{d_1}_t, \mathbf{b}^{d_2}_t, \cdots\right\}$ to distinguish between false positives, occluded objects, and potential new tracks. Each bounding box

Algorithm 1 FloraTracktor+: Iterative tracking updates integrating FloraTracktor with the Spatial Association Module.

Input: Set of input images $\mathcal{I} = \{I_t\}_{t=1}^{N}$
Output: Set of plant tracks $\mathcal{T} = \{\tau_t^a\}_{t=1}^{N}$

$I_{t-1} \leftarrow I_0$, $\mathbf{x}_{t-1} \leftarrow [0,0,1]^\top$, $\tau_{t-1}^a \leftarrow \emptyset$, $\tau_{t-1}^i \leftarrow \emptyset$
for each $I_t \in \mathcal{I}$ **do**
    $\tau_t^a,\ \tau_t^i,\ \mathcal{B}_t^p \leftarrow$ FloraTracktor($I_t,\ \tau_{t-1}^a,\ \tau_{t-1}^i$)
    $\mathbf{x}_t \leftarrow$ EstimateMotion($I_t,\ I_{t-1},\ \mathbf{x}_{t-1}$)
    $\tau_t^r \leftarrow$ Associate($\tau_t^i,\ \mathcal{B}_t^p,\ \mathbf{x}_t$)
    $\tau_t^a,\ \tau_t^i,\ \tau_t \leftarrow \left(\tau_t^a \cup \tau_t^r\right),\ \left(\tau_t^i \setminus \tau_t^r\right),\ \left(\tau_t^a \cup \tau_t^i\right)$
    $I_{t-1},\ \mathbf{x}_{t-1},\ \tau_{t-1}^i \leftarrow I_t,\ \mathbf{x}_t,\ \tau_t^i$
end for
$\mathcal{T} = \bigcup_{t=1}^{N} \tau_t^a$

at frame t is represented by the vector $\mathbf{b}_t^{l_j} = \left[x_t^{l_j}, y_t^{l_j}, w_t^{l_j}, h_t^{l_j}, s_t^{l_j}\right]^\top$, where $\mathbf{c}_t^{l_j} = (x_t^{l_j}, y_t^{l_j})$ denotes the coordinates of its centroid, $w_t^{l_j}$ its width, $h_t^{l_j}$ its height, and $s_t^{l_j}$ its confidence score. We define each track in τ_t^l as a vector $\boldsymbol{\tau}_t^{l_j} = [\mathrm{ID}^{l_j}, \mathbf{b}_{\text{last}}^{l_j\top}, \mathbf{x}^{l_j}]^\top$, where ID^{l_j} is the track ID, $\mathbf{b}_{\text{last}}^{l_j}$ is the bounding box of the track in the previous frame it was observed, and $\mathbf{x}^{l_j}$ is its estimated global position. If the regressed bounding boxes have confidence scores higher than s_{active}, their corresponding tracks stay alive. Otherwise, they become part of the set of inactive tracks τ_t^i.

Like Tracktor, we take the image I_t as input to perform detection and regress the bounding boxes of the active tracks τ_{t-1}^a in the previous frame. Tracktor performs three NMS steps: i) among the raw detections in $\mathcal{B}_t^d$; ii) among the regressed bounding boxes of the active tracks in $\mathcal{B}_t^a$ to account for occlusions; and iii) between each active track bounding box in $\mathcal{B}_t^a$ and all the new detections in $\mathcal{B}_t^d$ to discard detections already covered by existing tracks. Instead of performing NMS separately on $\mathcal{B}_t^d$ and $\mathcal{B}_t^a$ to account for possible occlusions, we perform a single NMS step on their union using a unique intersection-over-union (IoU) threshold λ_{nms}. After the NMS step, we use the retained bounding boxes to determine whether they belong to an active track, to a track that is currently occluding and thus temporarily deactivating another track, or if they are new detections that may correspond to a new object and hence require creating a new track. We denote this set of potentially new bounding boxes as $\mathcal{B}_t^p = \{\mathbf{b}_t^{p_1}, \mathbf{b}_t^{p_2}, \cdots\}$. To activate an inactive track in τ_t^i or create a new one, a new bounding box must have a confidence score greater than s_{new}.

In agricultural applications, all the objects in the field must be consistently tracked for precise spraying and accurate counting while allowing for the robot's trajectory to be flexible and enabling dynamic planning. Therefore, our framework stores all the deactivated tracks to allow for longer-term re-association. Since appearance features are not sufficiently discriminative for the re-identification of individual plants, we retain only the estimated global locations of the target objects, as discussed in Section 2.2. In other words, our algorithm relies exclusively on the bounding boxes provided by an object detector to keep track of the global positions of all the objects of interest captured by the data acquisition platform.

Finally, our method creates new tracks considering the size of the bounding box and its distance from the image edge. The size is generalized to the diameter of a circular approximation of the bounding box, which is given by

$$\phi\left(\mathbf{b}_t^{l_j}\right) = 2 \cdot \sqrt{\frac{w_t^{l_j} \cdot h_t^{l_j}}{\pi}}. \tag{1}$$

The new tracks centroid $\mathbf{c}_t^{p_j}$ must be at most $f_d \cdot \phi\left(\mathbf{b}_t^{l_j}\right)$ away from the image edge to be considered valid, where f_d is a dataset-specific threshold. Our tracker is denoted as FloraTracktor($\cdot$) in Algorithm 1.

2.2. Spatial Association Module

Our association module is based on the LoFTR [33] image matching model, which generates keypoint correspondences between two partially overlapping images. Given as input the pair of frames (I_{t-1}, I_t), LoFTR generates a set of C point correspondences $\mathcal{K}_t = \left\{(\mathbf{k}_t^c, \mathbf{k}_{t-1}^c)\right\}_{c=1}^{C}$ where $\mathbf{k}_t^c = [x_t^c, y_t^c]^\top$ are the

coordinates of the c-th keypoint on frame t. This process is referred to as LoFTR$(\cdot)$ in Algorithm 2.

The set of correspondences is then used to estimate a homogeneous transformation $H_{t-1,t} \in \mathbb{R}^{3\times3}$ between frames I_{t-1} and I_t, which defines the relationship between points in subsequent frames. We estimate $H_{t-1,t}$ using RANSAC [37] to handle possible outliers in the set of correspondences $\mathcal{K}_t$. The n-point algorithm [38] is used at each RANSAC iteration. These two steps are performed by the function EstimateHomography$(\cdot)$ in Algorithm 2. As Step 4 of the algorithm indicates, this transformation is then used to update the camera position $\mathbf{x}_t$ at time t.

Algorithm 2 Homography-based motion estimation.

1: **function** EstimateMotion($I_t,\ I_{t-1},\ \mathbf{x}_{t-1}$)
2: $\quad\mathcal{K}_t \leftarrow$ LoFTR($I_{t-1},\ I_t$)
3: $\quad H_{t-1,t} \leftarrow$ EstimateHomography($\mathcal{K}_t$)
4: $\quad\mathbf{x}_t \leftarrow \mathbf{x}_{t-1} + H_{t-1,t} \cdot \mathbf{x}_{t-1}$
5: $\quad$**return** $\mathbf{x}_t$
6: **end function**

Our implementation of Algorithm 1 executes the FloraTracktor$(\cdot)$ and EstimateMotion$(\cdot)$ functions in parallel. Upon the conclusion of both processes, we carry out a distance-based association step by invoking the function Associate$(\cdot)$, which is summarized in Algorithm 3 and described in detail below.

The function Associate$(\cdot)$ first obtains the set of inactive bounding boxes $\mathcal{B}_t^i$ from the set of inactive tracks τ_t^i, using the GetBoundingBoxes$(\cdot)$ function. For each bounding box $\mathbf{b}_t^j \in \mathcal{B}_t^p \cup \mathcal{B}_t^i$, we translate its centroid to the global coordinate system using the global camera position $\mathbf{x}_t$ so that the global coordinates of the centroids at time t are given by $\mathbf{x}_t^{l_j} = \mathbf{c}_t^{l_j} + \mathbf{x}_t$. The algorithm builds a distance matrix $M_t \in \mathbb{R}^{|\mathcal{B}_t^p|\times|\mathcal{B}_t^i|}$ containing the Euclidean distances between the global coordinates of the bounding boxes in $\mathcal{B}_t^p$ and $\mathcal{B}_t^i$. That is, let $x_t^{p_k}$ and $x_t^{i_k}$ be the global coordinates of the bounding boxes corresponding to the predicted and inactive tracks, the elements of M_t are given by $M_{p_k i_k} = ||\mathbf{x}_t^{p_k} - \mathbf{x}_t^{i_k}||$, $p_k = 1, \ldots, |\mathcal{B}_t^p|$, and $i_k = 1, \ldots, |\mathcal{B}_t^i|$.

Algorithm 3 Spatial association step.

1: **function** Associate($\tau_t^i,\ \mathcal{B}_t^p,\ \mathbf{x}_t$)
2: $\quad\mathcal{B}_t^i \leftarrow$ GetBoundingBoxes(τ_t^i)
3: $\quad$**for** $\mathbf{b}_t^j \in \mathcal{B}_t^p \cup \mathcal{B}_t^i$ **do**
4: $\quad\quad\mathbf{c}_t^j \leftarrow$ Centroid($\mathbf{b}_t^j$)
5: $\quad\quad\mathbf{c}_t^j \leftarrow \mathbf{c}_t^j + \mathbf{x}_t$
6: $\quad$**end for**
7: $\quad M_t \leftarrow$ Dist($\mathcal{B}_t^p, \mathcal{B}_t^i$)
8: $\quad\mathcal{A}_t \leftarrow$ Hungarian(M_t)
9: $\quad\tau_t^r \leftarrow \emptyset$
10: $\quad$**for** $(\mathbf{b}_t^{p_i},\ \mathbf{b}_t^{i_j}) \in \mathcal{A}_t$ **do**
11: $\quad\quad$**if** $M_{ij} < d^p$ **then**
12: $\quad\quad\quad\tau_t^r \leftarrow \tau_t^r \cup \left\{\boldsymbol{\tau}_t^{i_j}\right\}$
13: $\quad\quad$**end if**
14: $\quad$**end for**
15: $\quad$**return** τ_t^r
16: **end function**

We employ the Hungarian algorithm Hungarian$(\cdot)$ to find the optimal associations $\mathcal{A} = \left\{(\mathbf{b}_t^{p_j}, \mathbf{b}_t^{i_j})\right\}_{j=1}^{\min(|\mathcal{B}_t^p|,\,|\mathcal{B}_t^i|)}$ between inactive and potential new tracks. The track corresponding to $\mathbf{b}_t^{i_j}$ becomes active only if its distance to an existing track is below the threshold d^p, defined by

$$d^p = \frac{2}{|\mathcal{B}_t^p \cup \mathcal{B}_t^i|} \sum_{\forall \mathbf{b}^{l_j} \in \mathcal{B}_t^p \cup \mathcal{B}_t^i} \phi(\mathbf{b}^{l_j}), \tag{2}$$

which represents the average diameter of the circular bounding box approximation. The algorithm returns the set of re-associated tracks τ_t^r.

3. Experiments and Results

This section shows the performance of our framework with experiments conducted on the LettuceMOT and AppleMOT datasets. We begin with a brief introduction to the datasets and the metrics used for evaluation. In our first experiment, we present our main results on the LettuceMOT dataset. We evaluate the performance of our framework through comparisons with the

baseline provided by the LettuceMOT paper [27] and the benchmark established by the LettuceTrack paper [31]. Then, we establish benchmark performance MOT metrics for the AppleMOT dataset using our framework and compare it to the performance of the ByteTrack algorithm [18], which is currently one of the best-performing general MOT methods. We also examine the impact of the modifications introduced to Tracktor by FloraTracktor and FloraTracktor+. Finally, a sensitivity analysis is conducted on the parameters of our tracking framework: s_{new}, s_{active}, and λ_{nms}.

3.1. Datasets

We use the publicly available LettuceMOT and AppleMOTS datasets to evaluate our method. To use the AppleMOTS dataset, we derive a bounding-box-based version denominated AppleMOT. The construction of AppleMOT and the main properties of each dataset are described below.

3.1.1. The LettuceMOT dataset

This dataset consists of eight video sequences of a lettuce patch captured by an RGB camera mounted on a mobile robotic platform [27]. Each sequence includes annotated bounding boxes and frame-to-frame consistent identifiers for each lettuce plant. The different sequences represent typical situations in robotic-aided agricultural tasks, including various weather conditions, different plant growth stages, and various platform motion patterns. The eight sequences comprise 5,466 frames with a resolution of 810×1080 pixels and 707 unique object instances with 42,735 annotations.

The dataset contains four sequences captured in a forward motion of the robot, denoted as `straight1-straight4`. Two additional sequences represent obstacle avoidance by moving forward and backward in a lettuce row and are called `B&F1-B&F2` sequences. Finally, the two remaining sequences represent chemical refilling or battery charging situations, where the robot goes outside the crop row and re-enters at a different point. These sequences are referred to as the `O&I1-O&I2` sequences.

3.1.2. The AppleMOTS dataset

The AppleMOTS dataset comprises 12 video sequences captured using three distinct systems: i) a Matrice 210 RTK V2 UAV, ii) a Parrot Anafi UAV, and iii) a custom wearable sensor [26]. Each sequence includes annotated instance segmentation masks and identifiers for individual apples that remain consistent over the video frames. These sequences represent diverse scenarios encountered in computer vision-based agricultural tasks, encompassing varying illumination conditions and stages of fruit maturity. The 12 sequences comprise 1,673 frames with a resolution of 1296×972 and 2304 unique object instances with 86,000 annotated masks.

Since our method performs tracking based on bounding boxes, we use the mask annotations provided by the dataset to obtain the corresponding bounding boxes. The MOT-formatted bounding boxes, along with the original RGB images, comprise the AppleMOT dataset.

3.2. Evaluation Metrics

Our evaluation is based on the well-established and widely recognized CLEAR-MOT metrics [39] and the Higher Order Tracking Accuracy (HOTA) metric introduced in [40]. Appendix B includes a detailed description of the metrics used in this work.

MOTA (Multi-Object Tracking Accuracy) measures the overall tracking accuracy by computing the alignment of predicted and actual detections while maintaining consistent object identities over time. It penalizes identity switches but not identity transfers.

IDP (ID Precision) measures the precision of identity assignments, penalizing incorrect positive identity associations.

IDR (ID Recall) measures the recall of identity assignments, penalizing missed identity associations.

IDF1 (ID F1 Score) combines the precision and recall of identity assignments, offering a balanced measure of identity consistency over time.

AssPr (Association Precision) measures the precision of trajectory associations across frames by penalizing incorrect associations over time.

AssRe (Association Recall) measures the recall of trajectory associations across frames by penalizing missed associations over time.

AssA (Association Accuracy) measures the overall accuracy of trajectory associations, penalizing incorrect and missed associations over time, thus accounting for both ID switches and transfers.

DetA (Detection Accuracy) measures the accuracy of the detection process by comparing the predicted detections to actual detections and penalizing both false positives and negatives.

HOTA (Higher Order Tracking Accuracy) measures tracking performance by balancing detection accuracy (DetA) and association accuracy (AssA). It reflects both the accuracy of detecting objects and the precision of maintaining their trajectories over time. HOTA is calculated as $HOTA_\alpha = \sqrt{DetA_\alpha \cdot AssA_\alpha}$, providing a holistic view of tracking performance. Refer to [40] for more details.

All metrics described are scored on a scale where higher values indicate better performance (↑) with a maximum possible score of 1, or 100%. Except for MOTA, which can take negative values under critically poor tracking conditions, the metrics have a lower bound of 0. All the results reported in the following sections were obtained using the TrackEval tool [41].

3.3. Lettuce Tracking Performance Evaluation

We compare the performance of our proposed approach with baseline results from the LettuceMOT dataset paper [27] and the results reported in the LettuceTrack paper [31]. We follow the performance evaluation procedures described in the respective references to ensure a fair comparison. To compare with the LettuceMOT method, we train our detector using the sequences `straight1` and `straight3` and test it on the remaining sequences. For the LettuceTrack comparison, we train our detector using the sequences `straight3` and `straight4` and test it on the remaining sequences. The results below are obtained using the following default parameter values: $f_d = 1$, $s_{\text{new}} = 0.5$, $s_{\text{active}} = 0.5$, and $\lambda_{\text{nms}} = 0.2$.

Table 1 presents a performance comparison between the proposed method and the baseline results presented in [27]. Specifically, GIAOTracker [42] is the best approach in the LettuceMOT paper, where it is referred to as Bytetrack [18] + NSA Kalman Filter. FloraTracktor+ is our base tracker combined with our spatial association algorithm. The results demonstrate that FloraTracktor+ consistently surpasses the performance of the baseline method. This improvement is especially remarkable in association, as observed in the Association Accuracy (AssA), with an improvement of nearly 50% in the B&F sequences, 12% in the straight sequences, and 5% in the O&I sequences. The association directly and positively influences ID consistency, as quantified by the IDF1 metric. It is also worth highlighting the significant enhancement observed in the HOTA metric, which considers both temporal association and the maintenance of ID consistency over time. The HOTA metric shows an average improvement of 35% in the B&F sequences, 10% in the straight sequences, and 15% in the O&I sequences. Despite the general improvement, the O&I sequences still face difficulties since parts of the videos include frames without any visible object of interest.

Table 2 compares our FloraTracktor+ method with the baseline results from the LettuceTrack paper. In this scenario, SORT [15] and LettuceTrack [31] are the existing leading methods. The LettuceTrack paper omits assessments of the O&I sequences since the approach is tailored specifically for scenarios where the robot's movement is linear, anchoring tracking to the central axis of the orchard's layout [31]. FloraTracktor+ not only addresses this shortcoming but also significantly enhances association performance by more than 25%, leading

Table 1: Results obtained following the test methodology described in the LettuceMOT paper [27]: train on sequences `straight1` and `straight3` and test on the remaining sequences. Values not reported in [27] are represented by a dash (-). The ↑ means that higher values represent better performance.

Dataset	Method	MOTA↑	HOTA↑	DetA↑	AssA↑	AssRe↑	AssPr↑	IDF1↑	IDR↑	IDP↑
straight2	GIAOTracker [42]	90.20	87.24	86.31	88.15	-	-	94.72	-	-
	FloraTracktor+ (ours)	**98.30**	**98.21**	**99.36**	**98.04**	**98.41**	**99.58**	**98.50**	**98.56**	**98.44**
straight4	GIAOTracker [42]	86.43	84.41	84.01	84.83	-	-	92.722	-	-
	FloraTracktor+ (ours)	**98.09**	**97.10**	**99.17**	**98.04**	**98.70**	**99.30**	**98.51**	**98.74**	**98.44**
B&F1	GIAOTracker [42]	91.93	68.66	89.79	52.50	-	-	59.66	-	-
	FloraTracktor+ (ours)	**98.12**	**98.24**	**99.12**	**98.15**	**98.51**	**99.75**	**98.56**	**98.45**	**98.67**
B&F2	GIAOTracker [42]	86.51	62.73	84.02	46.86	-	-	52.07	-	-
	FloraTracktor+ (ours)	**98.08**	**98.94**	**99.12**	**97.84**	**98.51**	**99.30**	**98.43**	**98.54**	**98.41**
O&I1	GIAOTracker [42]	89.91	65.61	85.81	50.20	-	-	58.73	-	-
	FloraTracktor+ (ours)	**97.34**	**74.11**	**98.89**	**55.54**	**55.93**	**98.99**	**61.30**	**61.46**	**61.13**
O&I2	GIAOTracker [42]	51.30	52.90	61.92	45.23	-	-	46.25	-	-
	FloraTracktor+ (ours)	**95.38**	**72.61**	**96.36**	**54.71**	**55.46**	**98.39**	**58.06**	**59.03**	**57.11**

to a corresponding increase in IDF1 score of more than 15% and an overall HOTA score improvement of more than 25%. O&I results are omitted from Table 2 because they are virtually indistinguishable from those shown in Table 1.

3.3.1. Computation Time

FloraTracktor+ shows real-time performance, with an average execution time of 82.7 ± 22 milliseconds per frame, which corresponds to a rate of approximately 12 frames per second (fps). These results were obtained on a workstation with an NVIDIA GeForce 3090 GPU and an 11th Gen Intel Core i7-11700KF @ 3.60GHz CPU. The algorithm was implemented in Python without optimizations to reduce execution time. Optimizations, such as using tensorRT [43, 44] for model inference or employing model pruning and quantization as in [45, 46], could substantially reduce the computation time of the proposed framework, but they are beyond the scope of this study.

An essential feature of using PlantTracktor+ for ground plant tracking is that the spatial association module uses image matching for localization. LoFTR can find enough keypoint correspondences with 50% overlap between images without performance degradation [33]. Therefore, the relative displacement of an object should be accurately estimated if it is visible in approximately two frames in a row. This allows PlantTracktor+ to retain the object for effective longer-term re-association. For the LettuceMOT straight sequences, an object is visible for 40 to 50 frames from the first time it becomes fully visible to when it starts leaving the field of view. Hence, our method performs equally well even if the video frames are down-sampled at a 20:1 ratio. This allows for a twenty-fold increase in robot motion speed under the same data acquisition conditions.

3.4. Apple Tracking Performance Evaluation

The AppleMOT dataset allows us to evaluate the generalization capability of our approach. Since the AppleMOT dataset contains sequences with vanishing points in the images, we use a threshold $f_d = \max(I_W, I_H)$ in Eq. 1, where (I_W, I_H) represent the image dimensions. Following the AppleMOTS paper [26], we present the aggregated results over the test sequences `0006` to `0012`. Table 3 presents the performance comparison between our framework and the state-of-the-art tracker ByteTrack [18].

Our approach shows better tracking accuracy performance than ByteTrack as measured by the MOTA and HOTA metrics. The higher association accuracy and recall are largely due to the significantly lower detection accuracy obtained by ByteTrack. The dramatic reduction in true positive detections artificially inflates the association recall

Table 2: Results obtained following the test methodology described in the LettuceTrack paper [31]: train on sequences `straight3` and `straight4` and test on the remaining sequences. Values not reported in [31] are represented by a dash (-). The LettuceTrack paper does not provide results for the O&I sequences since it cannot handle these scenarios. The ↑ means that higher values represent better performance.

Dataset	Method	MOTA↑	HOTA↑	DetA↑	AssA↑	AssRe↑	AssPr↑	IDF1↑	IDR↑	IDP↑
straight1	SORT [15]	-	80.01	79.66	80.34	84.33	91.31	94.01	90.36	97.97
	LettuceTrack [31]	-	77.59	79.41	75.81	76.86	95.73	86.05	77.34	96.98
	FloraTracktor+ (ours)	**98.54**	**98.52**	**99.56**	**98.49**	**98.84**	**99.64**	**98.72**	**98.79**	**98.64**
straight2	SORT [15]	-	78.08	77.73	78.46	82.99	88.88	94.12	90.96	97.50
	LettuceTrack [31]	-	71.62	71.26	71.99	72.32	98.97	84.07	72.55	**99.95**
	FloraTracktor+ (ours)	**98.30**	**98.21**	**99.36**	**98.04**	**98.41**	**99.58**	**98.50**	**98.56**	**98.44**
B&F1	SORT [15]	-	58.31	78.57	43.30	44.39	91.72	54.28	51.63	57.22
	LettuceTrack [31]	-	76.81	79.69	74.03	75.42	94.50	85.30	76.76	95.97
	FloraTracktor+ (ours)	**98.12**	**98.24**	**99.12**	**98.15**	**98.51**	**99.75**	**98.56**	**98.45**	**98.67**
B&F2	SORT [15]	-	52.81	71.89	38.87	39.98	88.30	48.36	45.19	52.01
	LettuceTrack [31]	-	70.32	72.24	68.45	69.53	95.55	81.86	70.88	96.87
	FloraTracktor+ (ours)	**98.08**	**98.94**	**99.12**	**97.84**	**98.51**	**99.30**	**98.43**	**98.54**	**98.41**

Table 3: Summary of results following the test methodology proposed in the AppleMOTS paper [26]: train on sequences `0001` to `0005` and test on the remaining sequences. The ↑ means the higher the metric, the better performance.

Dataset	Method	MOTA↑	HOTA↑	DetA↑	AssA↑	AssRe↑	AssPr↑	IDF1↑	IDR↑	IDP↑
testing	ByteTrack [18]	32.99	38.21	28.86	**50.96**	**53.97**	74.97	**45.21**	31.49	**80.12**
	FloraTracktor (ours)	**48.55**	**45.45**	**60.08**	34.81	36.11	**86.66**	42.39	**38.91**	46.56

and, consequently, the accuracy. The comparable IDF1 obtained by both methods further supports this finding. The results obtained by our method are comparable with the Multi-Object Tracking and Segmentation Accuracy (MOTSA) reported in the AppleMOTS paper [26]. Note that the authors of [26] define image regions with limited visibility as "ignore regions." They report results for several tracking algorithms with and without excluding these regions at evaluation, but ground truth masks for these areas are not publicly available. The best method reported in [26] is PointTrack [22], which obtains a MOTSA of 52.9% when it filters the "ignore regions" and 46% when it does not. Our approach outperforms the equivalent MOTSA without filtering the "ignore regions" by 2.5%.

3.5. Evaluation of Improvements over Tracktor

We use the LettuceMOT dataset for this experiment since it includes sequences with ground plants and different scenarios of objects re-entering the camera's FOV, which are explicitly addressed by our method as described in Section 2. Table 4 presents a performance evaluation based on the overall MOTA, HOTA, and IDF1 metrics. Specifically, Tracktor and Tracktor++ are the original approaches described in [32]. FloraTracktor is our modified version of Tracktor, and FloraTracktor+ includes our spatial association module defined in Section 2.2.

Table 4 shows that FloraTracktor enhances performance by approximately 2–4%, with consistent standard deviations compared to Tracktor. Although

Table 4: Aggregated performance metrics for the proposed approach on the LettuceMOT dataset compared to the baseline trackers Tracktor and Tracktor++. The values are presented in the format $\mu \pm 3\sigma$ where μ is the mean value, and σ is the standard deviation. The ↑ means the higher the metric, the better performance.

Method	MOTA↑	HOTA↑	IDF1↑
Tracktor [32]	93.94 ± 0.92	81.71 ± 12.00	74.19 ± 18.88
Tracktor++ [32]	97.01 ± 0.68	74.38 ± 6.44	67.03 ± 12.68
FloraTracktor	95.29 ± 0.91	83.81 ± 12.27	78.03 ± 17.70
FloraTracktor+	$\mathbf{97.74 \pm 0.96}$	$\mathbf{91.98 \pm 10.77}$	$\mathbf{88.82 \pm 16.85}$

Tracktor++ achieves a modest increase in the MOTA metric due to its motion estimation module, it experiences a significant setback caused by false re-identifications that impair overall association, as reflected in the IDF1 metric. Consequently, this leads to a 7% decline in the HOTA metric in our experiments. Notably, our association algorithm in FloraTracktor+ not only achieves a MOTA score comparable to that of Tracktor++ but also leads to a 10% increase in IDF1 and an 8% HOTA improvement. Our method outperforms Tracktor with a 10% higher HOTA score and a 14% better IDF1 score.

3.6. Parameter Sensitivity Analysis

This section assesses the impact of parameter variations on the tracking performance of the proposed framework, mainly focusing on the MOTA, HOTA, and IDF1 metrics. As shown in Figure 2, our method exhibits consistent performance across various parameter settings. This stability persists even when parameter values fluctuate between 0.2 and 0.8, underscoring the model's reliability without meticulous parameter tuning.

In particular, the system's resilience to changes in the NMS threshold underscores the efficacy of the improved tracking-by-detection approach. This approach is less dependent on NMS for distinguishing between true and false positives than the standard Tracktor model. Additionally, the relatively low impact of the detection threshold, s_{new}, highlights the framework's capacity to handle slight detection inaccuracies effectively. The consistent performance across regression threshold values further validates our method's enhanced precision in bounding box prediction, a critical aspect of the temporal accuracy required in the spatial association module for agricultural monitoring.

3.7. The Spatial Association Module in Action

Figures 3-5 visually represent the estimated global positions of the lettuce plants. This representation demonstrates the accurate spatial relationships estimated by the proposed framework on various sequences of the LettuceMOT dataset, emphasizing the effectiveness of the spatial association module.

This qualitative analysis indicates that sequences categorized under O&I pose significant challenges, notably when the camera's FOV lacks target objects for extended periods. These gaps often lead to an accumulation of positional errors, which manifest as distortions in the reconstructed panorama. While these distortions primarily appear as rendering artifacts and do not directly affect the core performance metrics, they highlight potential issues with error accumulation in the estimated global position.

The quantitative results presented in Table 1 support these observations, indicating that despite the visual anomalies in the O&I sequence, the spatial association module significantly enhances tracking performance. This enhancement is particularly relevant in agricultural scenarios where robotic MOT involves nonlinear motion. However, distortions also point to limitations in our methodology, especially concerning error propagation in prolonged periods of complete absence of objects of interest. Our qualitative assessments further suggest that if our system has visibility of at least one trackable object, it can maintain accurate localization.

The implications of error buildup are relevant, especially for applications that rely on high cumulative positional accuracy. While our results validate the effectiveness of our approach, they also highlight areas requiring further refinement. Specifically, enhancing the robustness of the spatial association algorithm to handle extended periods of absolute object absence is a relevant area for ongoing and future research.

4. Conclusions and Future Work

We introduce a novel tracking algorithm designed specifically for mobile robots in agricultural environments, where the homogeneity of fruits and plants, coupled with nonlinear robot motion paths, presents significant challenges to conventional tracking frameworks. Our experiments show that our algorithm effectively addresses these challenges, demonstrating robustness and effectiveness, particularly in ground-plant tracking scenarios. A central feature of our approach is the innovative

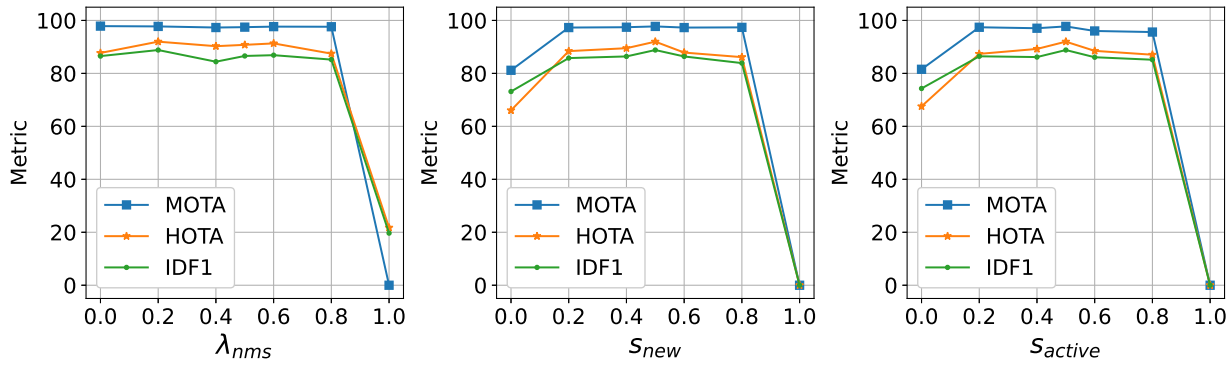

Figure 2: Performance of our approach for different values of s_{new}, s_{active}, and λ_{nms}.

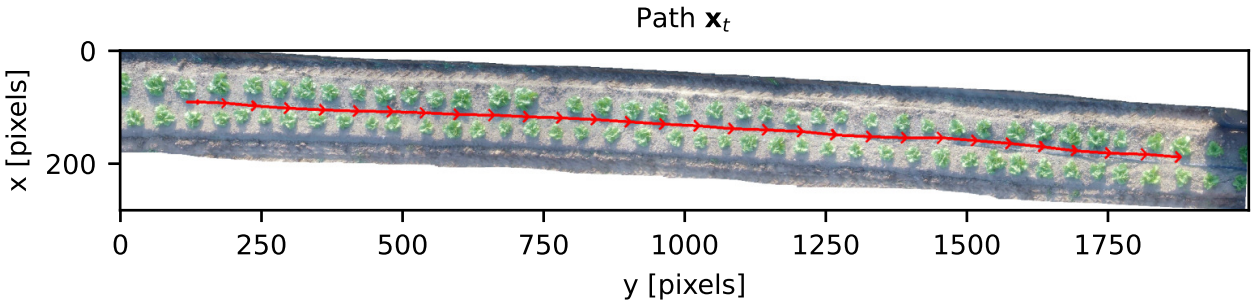

Figure 3: Global position estimates for the robot (camera) in sequence straight4 over 300 frames. The estimated trajectory on the global coordinate system (in pixels) is illustrated with a red path with arrows representing the direction of the speed vector.

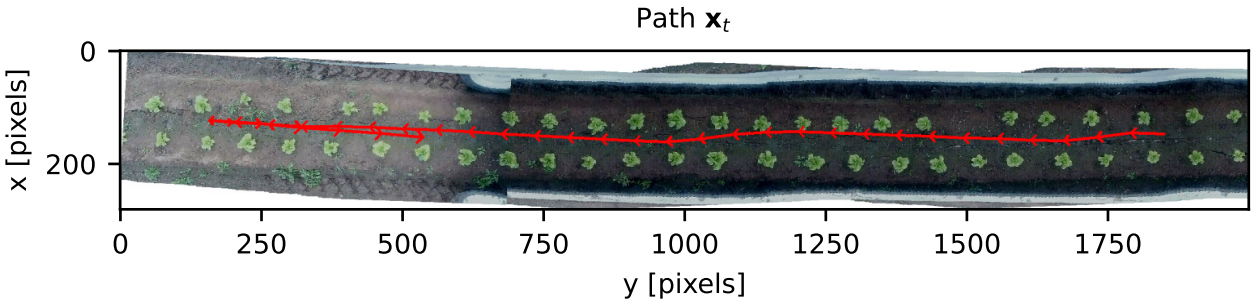

Figure 4: Global position estimates for the robot (camera) in sequence B&F1 over 360 frames. The estimated trajectory on the global coordinate system (in pixels) is illustrated with a red path with arrows representing the direction of the speed vector.

spatial association method, which leads to significant performance improvements, particularly when at least one object of interest remains within the camera's FOV.

Remarkably, our approach stands out as it does not require tracking annotations, laying the foundation for a potential semi-automated tracking labeling methodology. Furthermore, our tracker requires

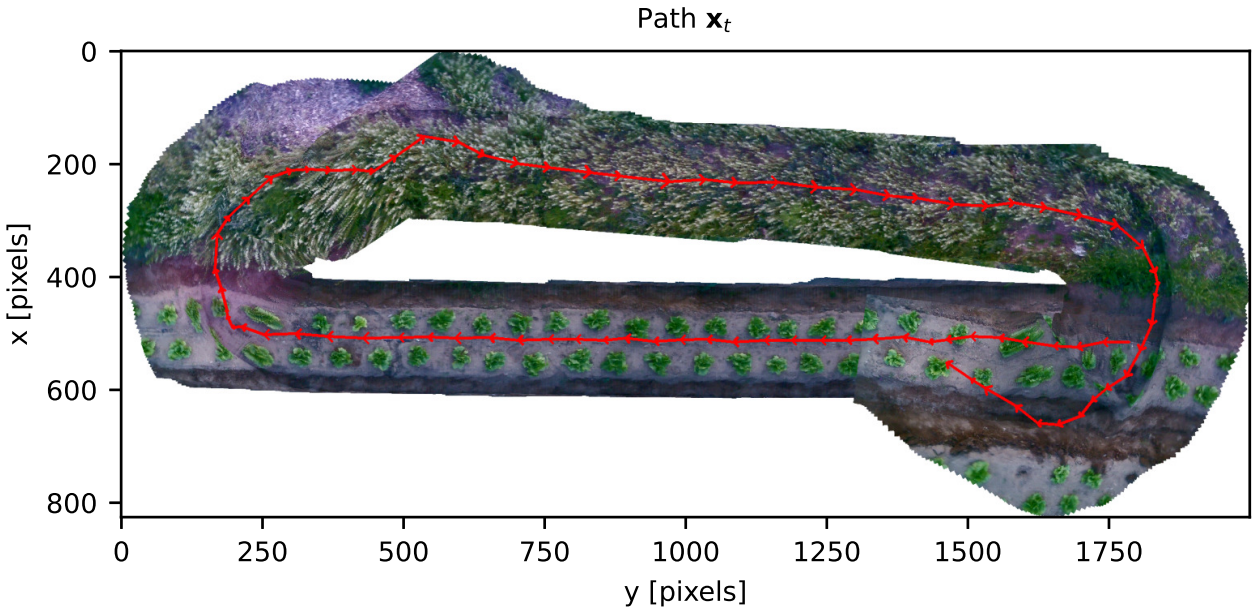

Figure 5: Global position estimates for the robot (camera) in sequence O&I2 over 850 frames. The estimated trajectory on the global coordinate system (in pixels) is illustrated with a red path with arrows representing the direction of the speed vector.

few parameter adjustments, offering a user-friendly interface without compromising performance. The performance of our algorithm remains relatively stable across large ranges of these parameters, underscoring its versatility and positioning it as a valuable asset in agricultural contexts.

Our work advances plant tracking technology and establishes a baseline performance for the modified AppleMOTS dataset and benchmark performance for the LettuceMOT dataset. Through a comprehensive set of experiments, results, and discussions, we have demonstrated the FloraTracktor+ framework's superior performance, surpassing both the LettuceMOT and LettuceTrack benchmarks. Furthermore, our sensitivity analysis has underscored the model's stability across diverse parameter settings, reducing the need for intricate parameter tuning.

The visual representation of the global position estimates presented in Section 3.7 illustrate the effectiveness of our spatial association module, even when faced with the challenges posed by scenarios of prolonged periods of absolute object absence. The proposed spatial location estimation module could seamlessly incorporate auxiliary data sources, such as GPS and IMU sensors, typically available in agricultural robots. This integration would facilitate overcoming one of the critical limitations of maintaining tracking accuracy in scenarios where the camera temporarily loses sight of the objects of interest. We intend to explore a sensor fusion strategy in future developments to enhance the robustness of the overall framework. Past studies suggest that the performance of such integrated systems often matches or surpasses that of the most effective individual sensor [47, 48, 49]. This potential improvement underscores the fundamental need to incorporate multiple data sources to ensure consistent and reliable tracking performance in diverse agricultural environments.

Acknowledgments

This research was partially funded by the National Science Foundation, Grant #2224591 and a University of Florida LIFT/AI seed grant.

References

[1] D. Dorling, World population prospects at the UN: our numbers are not our problem?, in: The Struggle for Social Sustainability, Policy Press, 2021, pp. 129–154.

[2] D. Benoît, S. Contzen, R. Nettle, M. T. Sraïri, The multiple influences on the future of work in agriculture: global perspectives, Frontiers in Sustainable Food Systems (2022).

[3] Y. Liu, X. Ma, L. Shu, G. P. Hancke, A. M. Abu-Mahfouz, From industry 4.0 to agriculture 4.0: Current status, enabling technologies, and research challenges, IEEE Transactions on Industrial Informatics (2020).

[4] M. Biswas, M. Ray, I. Jahan, S. Khan, S. Ahmad Saad, P. Bharman, Deep learning in agriculture: a review, Asian Journal of Research in Computer Science (2022).

[5] T. Saranya, C. Deisy, S. Sridevi, K. S. M. Anbananthen, A comparative study of deep learning and internet of things for precision agriculture, Engineering Applications of Artificial Intelligence (2023).

[6] M. H. Saleem, J. Potgieter, K. M. Arif, Automation in agriculture by machine and deep learning techniques: A review of recent developments, Precision Agriculture (2021).

[7] A. Tabb, H. Medeiros, A robotic vision system to measure tree traits, in: IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2017, pp. 6005–6012.

[8] P. A. Dias, A. Tabb, H. Medeiros, Apple flower detection using deep convolutional networks, Computers in Industry (2018).

[9] M. K. Saggi, S. Jain, A survey towards decision support system on smart irrigation scheduling using machine learning approaches, Archives of computational methods in engineering (2022).

[10] B. Darwin, P. Dharmaraj, S. Prince, D. E. Popescu, D. J. Hemanth, Recognition of bloom/yield in crop images using deep learning models for smart agriculture: A review, Agronomy (2021).

[11] I. Attri, L. K. Awasthi, T. P. Sharma, P. Rathee, A review of deep learning techniques used in agriculture, Ecological Informatics (2023).

[12] Z. Pezzementi, T. Tabor, P. Hu, J. K. Chang, D. Ramanan, C. Wellington, B. P. Wisely Babu, H. Herman, Comparing apples and oranges: Off-road pedestrian detection on the national robotics engineering center agricultural person-detection dataset, Journal of Field Robotics (2018).

[13] M. F. Kragh, P. Christiansen, M. S. Laursen, M. Larsen, K. A. Steen, O. Green, H. Karstoft, R. N. Jørgensen, FieldSAFE: dataset for obstacle detection in agriculture, Sensors (2017).

[14] D. I. Patrício, R. Rieder, Computer vision and artificial intelligence in precision agriculture for grain crops: A systematic review, Computers and Electronics in Agriculture (2018).

[15] A. Bewley, Z. Ge, L. Ott, F. Ramos, B. Upcroft, Simple online and realtime tracking, in: 2016 IEEE International Conference on Image Processing (ICIP), 2016, pp. 3464–3468.

[16] N. Wojke, A. Bewley, D. Paulus, Simple online and realtime tracking with a deep association metric, in: 2017 IEEE International Conference on Image Processing (ICIP), 2017, pp. 3645–3649.

[17] X. Zhou, V. Koltun, P. Krähenbühl, Tracking objects as points, European Conference on Computer Vision (2020).

[18] Y. Zhang, P. Sun, Y. Jiang, D. Yu, F. Weng, Z. Yuan, P. Luo, W. Liu, X. Wang, Bytetrack: Multi-object tracking by associating every detection box, in: European Conference on Computer Vision, 2022, pp. 1–21.

[19] P. Dendorfer, A. Osep, A. Milan, K. Schindler, D. Cremers, I. Reid, S. Roth, L. Leal-Taixé, MOTChallenge: A benchmark for single-camera multiple target tracking, International Journal of Computer Vision (2021).

[20] A. Geiger, P. Lenz, C. Stiller, R. Urtasun, Vision meets robotics: The KITTI dataset, The International Journal of Robotics Research (2023).

[21] P. Voigtlaender, M. Krause, A. Osep, J. Luiten, B. B. G. Sekar, A. Geiger, B. Leibe, MOTS: multi-object tracking and segmentation, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019, pp. 7942–7951.

[22] Z. Xu, W. Zhang, X. Tan, W. Yang, H. Huang, S. Wen, E. Ding, L. Huang, Segment as points for efficient online multi-object tracking and segmentation, in: European Conference on Computer Vision, 2020, pp. 264–281.

[23] M. Hussain, L. He, J. Schupp, D. Lyons, P. Heinemann, Green fruit segmentation and orientation estimation for robotic green fruit thinning of apples, Computers and Electronics in Agriculture (2023).

[24] W. Zhang, J. Wang, Y. Liu, K. Chen, H. Li, Y. Duan, W. Wu, Y. Shi, W. Guo, Deep-learning-based in-field citrus fruit detection and tracking, Horticulture Research (2022).

[25] R. Yao, G. Lin, S. Xia, J. Zhao, Y. Zhou, Video object segmentation and tracking: A survey, ACM Transactions on Intelligent Systems and Technology (TIST) (2020).

[26] S. de Jong, H. Baja, K. Tamminga, J. Valente, AppleMOTS: Detection, segmentation and tracking of homogeneous objects using MOTS, IEEE Robotics and Automation Letters (2022).

[27] N. Hu, S. Wang, X. Wang, Y. Cai, D. Su, P. Nyamsuren, Y. Qiao, Y. Jiang, B. Hai, H. Wei, LettuceMOT: A dataset of lettuce detection and tracking with re-identification of re-occurred plants for agricultural robots, Frontiers in Plant Science (2022).

[28] X. Wu, S. Aravecchia, P. Lottes, C. Stachniss, C. Pradalier, Robotic weed control using automated weed and crop classification, Journal of Field Robotics (2020).

[29] Y. Lu, S. Young, A survey of public datasets for computer vision tasks in precision agriculture, Computers and Electronics in Agriculture (2020).

[30] A. Joshi, D. Guevara, M. Earles, Standardizing and centralizing datasets for efficient training of agricultural deep learning models, Plant Phenomics (2023).

[31] N. Hu, D. Su, S. Wang, P. Nyamsuren, Y. Qiao, Y. Jiang, Y. Cai, LettuceTrack: Detection and tracking of lettuce for robotic precision spray in agriculture, Frontiers in Plant Science (2022).

[32] P. Bergmann, T. Meinhardt, L. Leal-Taixe, Tracking without bells and whistles, in: IEEE/CVF International Conference on Computer Vision, 2019, pp. 941–951.

[33] J. Sun, Z. Shen, Y. Wang, H. Bao, X. Zhou, LoFTR: Detector-free local feature matching with transformers, in: IEEE/CVF conference on computer vision and pattern recognition, 2021, pp. 8922–8931.

[34] R. Girshick, J. Donahue, T. Darrell, J. Malik, Rich feature hierarchies for accurate object detection and semantic segmentation, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2014, pp. 580–587.

[35] S. Ren, K. He, R. Girshick, J. Sun, Faster R-CNN: Towards real-time object detection with region proposal networks, Advances in neural information processing systems (2015).

[36] R. Girshick, Fast R-CNN, in: Proceedings of the IEEE international conference on computer vision, 2015, pp. 1440–1448.

[37] M. A. Fischler, R. C. Bolles, Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography, Communications of the ACM (1981).

[38] R. I. Hartley, In defense of the eight-point algorithm, IEEE Transactions on pattern analysis and machine intelligence (1997).

[39] K. Bernardin, R. Stiefelhagen, Evaluating multiple object tracking performance: the CLEAR MOT metrics, EURASIP Journal on Image and Video Processing (2008).

[40] J. Luiten, A. Osep, P. Dendorfer, P. Torr, A. Geiger, L. Leal-Taixé, B. Leibe, HOTA: A higher order metric for evaluating multi-object tracking, International Journal of Computer Vision (2020).

[41] A. H. Jonathon Luiten, TrackEval, `https://github.com/JonathonLuiten/TrackEval` (2020).

[42] Y. Du, J. Wan, Y. Zhao, B. Zhang, Z. Tong, J. Dong, GIAOTracker: A comprehensive framework for MCMOT with global information and optimizing strategies in VisDrone 2021, in: IEEE/CVF International Conference on Computer Vision, 2021, pp. 2809–2819.

[43] J. Xu, B. Wang, J. Li, C. Hu, J. Pan, Deep learning application based on embedded GPU, in: 2017 First International Conference on Electronics Instrumentation & Information Systems (EIIS), IEEE, 2017, pp. 1–4.

[44] P. Davoodi, C. Gwon, G. Lai, T. Morris, TensorRT inference with Tensorflow, in: GPU Technology Conference, 2019, p. 1.

[45] J. Terven, D.-M. Córdova-Esparza, J.-A. Romero-González, A comprehensive review of YOLO architectures in computer vision: From YOLOv1 to YOLOv8 and YOLO-NAS, Machine Learning and Knowledge Extraction (2023).

[46] G. Jocher, C. A., J. Qiu, YOLO by Ultralytics (2024).
URL `https://github.com/ultralytics/ultralytics`

[47] A. B. Torres, A. R. da Rocha, T. L. C. da Silva, J. N. de Souza, R. S. Gondim, Multilevel data fusion for the internet of things in smart agriculture, Computers and Electronics in Agriculture (2020).

[48] V. Barrile, S. Simonetti, R. Citroni, A. Fotia, G. Bilotta, Experimenting agriculture 4.0 with sensors: A data fusion approach between remote sensing, UAVs and self-driving tractors, Sensors (2022).

[49] W. Huang, Y. Wang, J. Xia, X. Jin, H. Zhu, B. Glamuzina, W. Yu, X. Zhang, Non-destructive classification of sturgeon stress using cross-modal data fusion and multi-input deep learning models, Computers and Electronics in agriculture (2024).

[50] T. Li, M. Sun, Q. He, G. Zhang, G. Shi, X. Ding, S. Lin, Tomato recognition and location algorithm based on improved YOLOv5, Computers and Electronics in Agriculture (2023).

Appendix A. Paper Notation

Nomenclature

a	Superscript indicating membership in the *active* set
d	Superscript indicating membership in the *detections* set (raw)
i	Superscript indicating membership in the *inactive* set
p	Superscript indicating membership in the *potentially new* set
r	Superscript indicating membership in the *re-associated* set
t	Subscript indicating time t
I_H	Frame height
I_W	Frame width
I_t	Frame at time t
$\mathcal{K}_t$	Set of keypoints correspondences between frame t and $t-1$, $\mathcal{K}_t = \left\{(k_t^c, k_{t-1}^c)\right\}_{c=1}^{C}$, $k_t^c \in \mathbb{R}^2$
$\mathbf{x}_t$	Global camera position at time t. The origin is $\mathbf{x}_0 = [0, 0]$, $\mathbf{x}_t \in \mathbb{R}^2$
$H_{t-1,t}$	Transformation for camera coordinates from time $t-1$ to time t
$\mathbf{b}_t^{l_j}$	j-th bounding box of set $\mathcal{B}_t^l$, where $l \in \{a, d, i, p\}$, comprised of its centroid coordinates $x_t^{l_j}$, $y_t^{l_j}$, width $w_t^{l_j}$, height $h_t^{l_j}$, and confidence $s_t^{l_j}$, $\mathbf{b}_t^{l_j} = \left[x_t^{l_j}, y_t^{l_j}, w_t^{l_j}, h_t^{l_j}, s_t^{l_j}\right]^\top$, $\mathbf{b}_t^{l_j} \in \mathbb{R}^5$
$\mathbf{c}_t^{l_j}$	Centroid of the l_j-th bounding box, $\mathbf{c}_t^j = [x,\ y]^\top$, $\mathbf{c}_t^l \in \mathbb{R}^2$
r^{l_j}	Radius of the circle enclosed by the bounding box $\mathbf{b}_t^{l_j}$, $r^{l_j} = (w_t^{l_j}/2 + h_t^{l_j}/2)/2 = (w_t^{l_j} + h_t^{l_j})/4$
τ_t^l	Tracks in set $l \in \{k, i, p, a\}$ at time t. Each track in τ_t^l is defined as $\boldsymbol{\tau}_t^{l_j} = [\mathrm{ID}^{l_j}, \mathbf{b}_{\text{last}}^{l_j\top}, \mathbf{x}^{l_j\top}, r^{l_j}]^\top$ containing the track ID, its bounding box in the last frame it was observed, its estimated global position, and the radius of the circle enclosed by the bounding box
$\mathcal{T}$	Set of all the tracks, active or inactive. $\mathcal{T} = \{\tau_t^a\}_{t=1}^{N}$
$\mathcal{B}_t^l$	Set of bounding boxes corresponding to the set $l \in \{a, d, i, p\}$ at time t, $\mathcal{B}_t^l = \left\{\mathbf{b}_t^{l_1}, \mathbf{b}_t^{l_2}, \cdots\right\}$
s_{new}	confidence threshold to initialize a new track (from a new detection)
s_{active}	confidence threshold to keep a track alive (from a regressed box)
λ_{nms}	global non-maxima suppression threshold
d^p	distance threshold for association defined as the average diameter assuming a circular approximation of the shape. The *potentially new* tracks re-activate an inactive track if the distance between them is below d^p.

Appendix B. Multiple Object Tracking Metrics

Table B.1: Summary of the MOT evaluation metrics used in this work. In the table, m_t, fp_t, and mme_t are the number of misses, false positives, and mismatches. $IDTP$, $IDFP$, and $IDFN$ are the identity-based true positives, false positives, and false negatives. For additional details, refer to [39, 40].

Metric	**Name**	**Description**	**Equation**
MOTA	Multi-Object Tracking Accuracy	Overall tracking accuracy, computed by aligning predicted and actual detections while maintaining consistent object identities over time. It penalizes identity switches but not identity transfers.	$1-\frac{\sum_t (m_t + fp_t + mme_t)}{\sum_t g_t}$
IDP	ID Precision	Precision of identity assignments, penalizing incorrect associations due to extra predictions.	$\frac{\lvert IDTP\rvert}{\lvert IDTP\rvert + \lvert IDFP\rvert}$
IDR	ID Recall	Recall of identity assignments, penalizing missed identity associations.	$\frac{\lvert IDTP\rvert}{\lvert IDTP\rvert + \lvert IDFN\rvert}$
IDF1	F1 score for ID associations	Geometric average between IDP and IDR.	$\frac{\lvert IDTP\rvert}{\lvert IDTP\rvert + 0.5\lvert IDFP\rvert + 0.5\lvert IDFN\rvert}$
$A(c)$	A score	Association score for true positive detection c. $TPA(c)$, $FPA(c)$, and $FNA(c)$ are the sets of true positive, false positive, and false negative detections with the same id as c.	$\frac{\lvert TPA(c)\rvert}{\lvert TPA(c)\rvert + \lvert FPA(c)\rvert + \lvert FNA(c)\rvert}$
AssPr	Association Precision	Precision of trajectory associations across frames by penalizing incorrect associations over time.	$\frac{1}{\lvert TP\rvert}\sum_{c\in(TP)} \frac{\lvert TPA(c)\rvert}{\lvert TPA(c)\rvert + \lvert FPA(c)\rvert}$
AssRe	Association Recall	Recall of trajectory associations across frames by penalizing missed associations over time.	$\frac{1}{\lvert TP\rvert}\sum_{c\in(TP)} \frac{\lvert TPA(c)\rvert}{\lvert TPA(c)\rvert + \lvert FNA(c)\rvert}$
AssA	Association Accuracy	Overall accuracy of trajectory associations, which penalizes incorrect and missed associations over time, thus accounting for both ID switches and transfers.	$\frac{1}{\lvert TP\rvert}\sum_{c\in(TP)} A(c)$
DetA	Detection Accuracy	Accuracy of similarity between the predicted and actual detections. Identities are not relevant for this metric.	$\frac{\lvert TP\rvert}{\lvert TP\rvert + \lvert FP\rvert + \lvert FN\rvert}$
HOTA	Higher Order Tracking Accuracy	Double Jaccard index accounting for detection accuracy and association Accuracy.	$\sqrt{\frac{\sum_{c\in(TP)} A(c)}{\lvert TP\rvert + \lvert FP\rvert + \lvert FN\rvert}}$