The Growing Strawberries Dataset: Tracking Multiple Objects with Biological Development over an Extended Period

Junhan Wen[1] Camiel R. Verschoor[2] Chengming Feng[1] Irina-Mona Epure[2]
Thomas Abeel[1] Mathijs de Weerdt[1]

[1]Delft University of Technology, [2]Birds.ai B.V.
{junhan.wen,c.feng-1,t.abeel,m.m.deweerdt}@tudelft.nl, {camiel,irina}@birds.ai

Abstract

Multiple Object Tracking (MOT) is a rapidly developing research field that targets precise and reliable tracking of objects. Unfortunately, most available MOT datasets typically contain short video clips only, disregarding the indispensable requirement for adequately capturing substantial long-term variations in real-world scenarios. Long-term MOT poses unique challenges due to changes in both the objects and the environment, which remain relatively unexplored. To fill the gap, we propose a time-lapse image dataset inspired by the growth monitoring of strawberries, dubbed The Growing Strawberries Dataset (GSD). The data was captured hourly by six cameras, covering a span of 16 months in 2021 and 2022. During this time, it encompassed a total of 24 plants in two separate greenhouses. The changes in appearance, weight, and position during the ripening process, along with variations in the illumination during data collection, distinguish the task from previous MOT research. These practical issues resulted in a drastic performance downgrade in the track identification and association tasks of state-of-the-art MOT algorithms. We believe The Growing Strawberries will provide a platform for evaluating such long-term MOT tasks and inspire future research. The dataset is available at https://doi.org/10.4121/e3b31ece-cc88-4638-be10-8ccdd4c5f2f7.v1.

1. Introduction

Multiple Object Tracking (MOT) is an exciting Computer Vision topic with wide applications in autonomous driving [25, 37], traffic monitoring [30, 34], video surveillance [32, 41], etc. While these studies mainly focused on video clips of a few minutes or even shorter [14, 23, 43, 67], consistent tracking over a longer period also has significant implications in real-world contexts. The supervision of cultivation and livestock [20, 24, 54, 66], the progression assessment of lesions and wounds [3, 10, 11, 29, 58], and the microscopic scrutiny of cells [2, 42] serve as intriguing illustrations of this pragmatic scenario. However, there is a lack of research on MOT algorithms applied for long-term purposes, particularly when the intrinsic properties of objects are also simultaneously developing. Furthermore, using a lower capture frequency over extended periods [11, 54, 66] leads to a substantial information loss, thereby heightening the challenges in accurate object tracking. Therefore, there is a pronounced need for a realistic dataset to bridge the gap between current MOT algorithms and their effective application over prolonged periods, so as to facilitate the advancement of effective methods.

Figure 1. An example subsequence of image segments from *GSD*, depicting the growth over five days. We can notice dramatic appearance changes and gradual enlargement during the development. In addition, even though the segments are selected to minimize lightness variations, slight differences in segment brightness may still be discernible due to the shifting angles of sunlight.

The tracking of biological development processes exemplifies a prominent long-term MOT challenge within this particular context [38, 39, 65, 68]. For instance, accurate growth monitoring of fruits and vegetables over time is a key ingredient to successful horticulture. Recent studies have demonstrated that images are feasible non-destructive tools to evaluate the status and quality of fruits [19, 26, 70]. Keeping track of the growth helps in planning harvest schedules, so as to achieve the peak quality and nutritional value of crops. To follow the growth of individual fruits through visual observations, automated image processing is required. We chose strawberries for our research because their 3-7 day life cycle allows for tracking noticeable appearance changes while maintaining a reasonable frame rate. In addition, the natural growth and horticulture activities also introduce object movements along frames.

In this paper, we introduce the first in-the-wild biological development monitoring dataset, *The Growing Strawberries* Dataset (*GSD*). The videos of *GSD* consist of time-lapse images of strawberry cultivation in six spots at two different greenhouses during the growing season in 2021 and 2022 respectively. The longitudinal observations of strawberries over their growth are supportive for ripeness assessment, yield prediction, and harvest planning for efficient supply [9, 50]. Unlike the trajectory tracking of common moving objects, *GSD* involves *long-term* tracking of developing objects under a low frame rate, which introduces the two following unique challenges to the MOT task.

Appearance changes result from the biological growth of strawberries and include changes in color, shape, and size, as depicted in Fig. 1. These are common properties when a biological object is developing over time, yet limited MOT research has taken these issues into concern. Unlike pedestrians or vehicles that remain visually consistent throughout short videos, strawberries undergo continuous changes in appearance during long-term tracking. Additionally, the visual appearances of strawberries are more similar to each other than those of the traffic participants, which are more colorful and varied. The in-frame discrimination and across-frame association result in challenges for the appearance descriptors, particularly when also confronting dynamic lighting situations and overnight intervals.

Irregular movements can be caused by horticulture operations or other human activities. They exhibit occasional co-occurrence with the strawberries' incremental movements from natural weight increase. For example, the natural increase in fruit weight or deliberate repositioning by horticulturalists can lead to changes in fruit positions. Human intervention can introduce unexpected occurrences like sudden object movements or camera view occlusions. Additionally, harvested fruits may permanently vanish from sight. Since the data is captured hourly, movements could lead to abrupt changes, e.g. position jumps or switches, which make many location changes of *GSD* objects non-linear and irregular. This characteristic from practice calls for research of discontinuous or interrupted videos, which has not been thoroughly investigated, whilst the joint effect with the appearance change still calls for more effective MOT solutions.

The main contributions of our work are: 1) We established *GSD*, a long-term MOT dataset that used six cameras to track the growth of 12 plants of strawberries in 2021 and 2022 in two different greenhouses. 2) We quantitatively compared *GSD* with one popular MOT dataset, *MOT20*, and proposed a unique MOT scenario: the temporal tracking of biologically developing objects in a sparse and long-term data collection. 3) We benchmarked the performance of five MOT algorithms to prove the challenges brought by our proposed scenario. 4) We visualized the importance of *GSD* from a realistic perspective. In all, our results evidence the limitations of state-of-the-art MOT algorithms for such a long-term MOT task, which highlights the emergence and necessity of proposing *GSD*.

2. Related Work

In this section, we briefly review popular object-tracking and temporal datasets that promote algorithm development and their limitations on scenarios, in order to highlight the uniqueness and importance of the *GSD*. Secondly, we summarize the concepts of state-of-the-art MOT algorithms and explain our method for evaluating the *GSD*.

2.1. Image Datasets for Multiple-Object Tracking

Datasets for MOT predominantly focus on trajectory tracking. Many of the recent tasks of the *MOT Challenge* [44] are motivated by surveillance and autonomous driving. Thus, they mostly focus on the tracking of pedestrians, vehicles, passengers, etc. [18, 23, 43]. For instance, *MOT20* [15] is a widely-used and representative MOT dataset and is extensively utilized by various algorithms as a benchmark to assess their performance. The majority of the sequences are short videos with 10-30 frames per second and lasting for a few minutes [14]. New challenges mainly originate from a higher amount and density of objects in emerging datasets [15, 56, 60]. However, there are limited changes in the characteristics of research scenarios. For instance, popular research objects such as pedestrians or vehicles are often characterized by regular or predictable movement patterns. As a result, a greater diversity of datasets is required to facilitate the generalization of MOT in broader domains [14, 67].

The majority of long-term temporal image datasets are used for substantial-scale change detection, e.g. the progress monitoring of construction, deforestation, urbanization, or animal migrations [17, 45, 47, 57, 64]. One of the shared goals is to track the temporal changes of large and (mostly-)static objects or of a comprehensive overview of objects. Therefore, the main concern in these studies is the pattern differences across images. On the other hand, these datasets have limited potential to motivate the development of MOT algorithms due to the restricted spatial movements of objects.

2.2. Image Datasets for Plant Science

Image datasets are vital for plant science. Sequential images are a practical data type to accomplish non-destructive tests and continuous growth monitoring. The majority of plant science research involving non-destructive testing of images is carried out within controlled and calibrated laboratory settings [40, 48, 70]. However, for fruits that do not ripen after harvest, it becomes impractical to rely on lab data for recording status updates during their growth. Existing in-field datasets primarily focus on one-shot fruit detection and lack information on the ripening progress due to limited object appearance changes over a short period [21, 33, 49, 69].

Moreover, hyper-spectrum images (HSI) play a valuable role in plant studies by developing numerical indicators and training machine-learning models [21, 28, 70]. Yet, integrating these images into agriculture practices is resource-intensive, given the already costly nature of HSI data collection. Therefore, we advocate for a more practical solution: an integrated temporal dataset merging images in the visual and near-infrared spectrum. The scarcity of non-visual images further emphasizes the need for such a comprehensive dataset.

2.3. Algorithms for Multiple-Object Tracking

Online MOT algorithms aim to perform real-time tracking of multiple objects in video sequences by continuously updating object identities and associations. Tracking-by-detection is the most widely-used strategy in achieving online MOT [1, 16, 59]. The strategy enhances the algorithms' adaptability and robustness, enabling them to easily accommodate and perform well in diverse scenarios. In addition, it has less reliance on high FPS of data collection than strategies building end-to-end detector-trackers such as [4], which exhibits a higher potential for successful adaptation and utilization in long-term MOT problems. Offline MOT solvers are also powerful tools as they utilize batches of frames [8, 52, 61]. Since the computation effort grows tremendously on larger datasets [1], it is out of the scope of the context of our dataset. Thus, online MOT algorithms are more applicable in *GSD*.

Algorithms following the tracking-by-detection strategy consist of two stages: i) applying object detection models and ii) associating bbox across frames. Research towards better (near-)real-time performance mainly focuses on enforcing the associating algorithm or a better interconnection between the two stages [59]. Generally, the association step concerns two criteria [56]: **i) The trajectory and motion of objects.** Many MOT algorithms are developed based on the Simple online and Real-time Tracking (SORT) algorithm, in which a Kalman filter framework is applied to analyze the velocity vectors [6, 12, 71]. The utilization of inertia measurement is a widely recognized approach for expeditiously handling the MOT task. Nevertheless, researchers argue that trajectories of spatially close objects are difficult to be distinguished [61]. **ii) The appearance of objects.** Deep learning techniques are usually applied to encode the appearance information of targets [13, 59, 61, 62]. Field-specific object properties are often integrated to enhance association performance [11, 52]. Particularly, when the frames are discontinuous or when the objects are occluded, appearance features are crucial in re-identifying and associating the tracklets to achieve consistent global tracking [55, 72–75]. Nevertheless, the sparsity of the image collection for *GSD* indicates a longer interval between frames, which exacerbates the existing complexity of the task.

[1] An example on *GSD* is demonstrated in the supplementary materials.

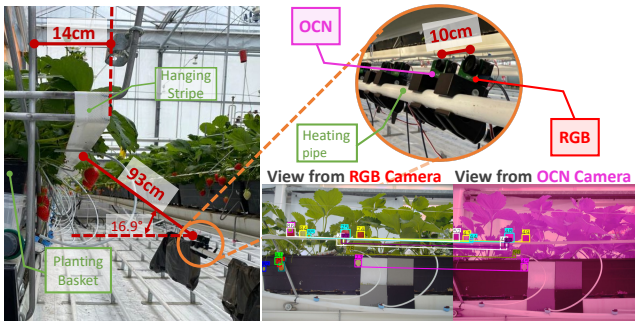

Figure 2. Detailed setup in the greenhouse in 2021. The left photo shows the positions of the white stripes, the planting baskets, and the heating pipe, which were all placed in parallel. The distance from the edge of the white stripes to the camera lens was 93cm. The RGB camera was placed 10 cm to the left of the OCN camera of each pair, as shown on the top right. The elevation angle of both cameras is 16.9°. Sample images from *RGB-1* and *OCN-1* are shown in the bottom right. Identical strawberries are color-coded. The setup is similar in 2022 with slightly varied dimensions.

3. The *Growing Strawberries* Dataset

We aimed to create a dataset about prolonged object tracking in a real-world setting for the purpose of long-term MOT. The growth of strawberries is a good example of a natural biological development process. Appearance changes and irregular movements happen during this ripening process. Such dynamics reveal special characteristics that are also shared among all kinds of agricultural crops.

To this end, we used six cameras (three RGB + three OCN [2]) to track the growth of 12 *Favori* plants over 30 weeks in 2021 and 32 weeks in 2022, in two greenhouses with different cultivation setups in The Netherlands. The cameras were paired in three sets, denoted as *RGB/OCN-1/2/3*. They captured time-lapse images in the greenhouse, such that videos of the entire ripening process were archived. We provide human-annotated bboxes for every strawberry, at all growth stages, and identities for corresponding trajectories.

3.1. Data Collection Setup

Since the ripening lasts around 7-14 days, we used hourly images for growth monitoring, such that a complete track of the plant is ensured with circa 100 observations. The strawberries were cultivated in planting baskets that hung from the ceiling. A heating pipe was hung beneath each planting basket. The cameras were attached to the heating pipe on the opposite side of the strawberry plant. Fig. 2 illustrates the detailed setup of the cameras in the greenhouse.

Both cameras faced the plants from parallel perspectives, where the OCN images were taken with a large view overlap with the RGB ones to provide hyper-spectral information.

[2] The channels are: Orange/615nm, Cyan/490nm, Near-Infrared/808nm.

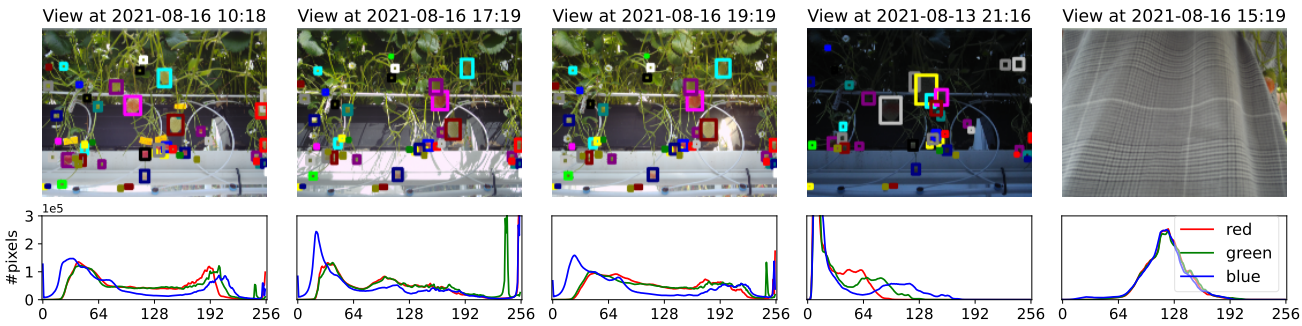

Figure 3. The upper row lists five image samples taken by *RGB-3*. The capture time is indicated in the title. The 1st - 4th images depict the normal changing pattern of sunlight during the day. We use the 4th image from a different date because the dawn and dusk were not captured every day. The 5th image shows how the view might be blocked due to human activities. The plot beneath each photo is the corresponding color spectrum from R/G/B channels respectively. The x-axes indicate the color value (encoded as 0-255). The y-axes are the power of the color spectrum with a shared amplitude. The color-coded rectangles illustrate the ground-truth (GT) bbox and trajectory annotations.

On average, 28 strawberry fruits from 4 plants were monitored by one RGB camera. We index all three RGB cameras as *RGB-1,2,3*. Fig. 3 shows the annotations of an example image sequence taken by camera RGB-3.

3.2. Ground-Truth Annotation

The trajectory annotations of the strawberries consist of bboxes with track identifiers (track IDs). Flowers of strawberries and paper tags for identifying fruits with further measurement results were annotated into different categories and were excluded from the benchmark experiments. Hereafter, we use the word "strawberry" referring to only the fruits.

The annotation was accomplished by drawing and marking bbox and track IDs. To remain consistent in labels, the first round of annotation was performed by a single person. Subsequently, two separate reviewers performed a manual check on the annotations to ensure accuracy and to mitigate potential labeling errors or personal biases. In this way, we guarantee accurate annotations. For an example, please see Fig. 3 and further in the supplementary materials.

All the images are 4000×3000 pixels. Due to the continuous data collection spanning the entire day, the illumination conditions exhibited significant and periodic variations. We therefore set up a brightness threshold and defined a subset specifically for the following benchmarking experiments.

Day images. The RGB images that were taken under normal lighting conditions are the majority share of the growth tracking task. Examples are as depicted in the left three photos in Fig. 3. We call this subset the "day images". Quantitatively, they were defined as the images with luminance (Y) [3] higher than 50. As is illustrated by the first three columns in Fig. 3, when the zenith angle of the sun changes during the day, the color spectrum of the photo shifts. This is a practical challenge brought by the in-the-wild data collection, which also aggravates the variation of object appearances.

[3]Luma, calculated according to ITU-R BT.601 standard [7].

Table 1. Statistical overview of the RGB images of *GSD*. The 2nd column lists the duration of data collection. The 3rd and 4th columns note the amounts of all images and the images used in the benchmarking studies respectively. The last two columns present the total number of bboxes and trajectories. An overview of the OCN images is presented in the supplementary materials.

Camera	Period	Total img	Anno. img	Total bbox	Total track
RGB-1	Apr 23 - Nov 9, 2021	4786	2823	67957	492
RGB-2	Apr 23 - Nov 9, 2021	4785	2638	64434	392
RGB-3	Jun 29 - Nov 9, 2021	3181	1761	70641	431
RGB-1	Feb 22 - Oct 3, 2022	5128	3369	93439	540
RGB-2	Feb 22 - Oct 3, 2022	4699	3062	117291	872
RGB-3	Feb 22 - Oct 3, 2022	5156	3330	109946	754

Remainder images. The annotations are available for all frames until most strawberries became invisible when the view became very dark or when the camera was occluded by human activities (e.g. the 5th photo Fig. 3). We defined the subset "darker images" as the photos that were taken under insufficient daylight (i.e. image brightness $\leq$ 50) but the strawberries were still visible to be annotated, for example, the 4th photo in Fig. 3. Nevertheless, without additional brightness normalization, darker images degraded the performance of the detection models. Considering that the number of darker images was limited (at most once during dawn and/or dusk), we excluded them in the benchmarking experiments to keep a fair performance comparison.

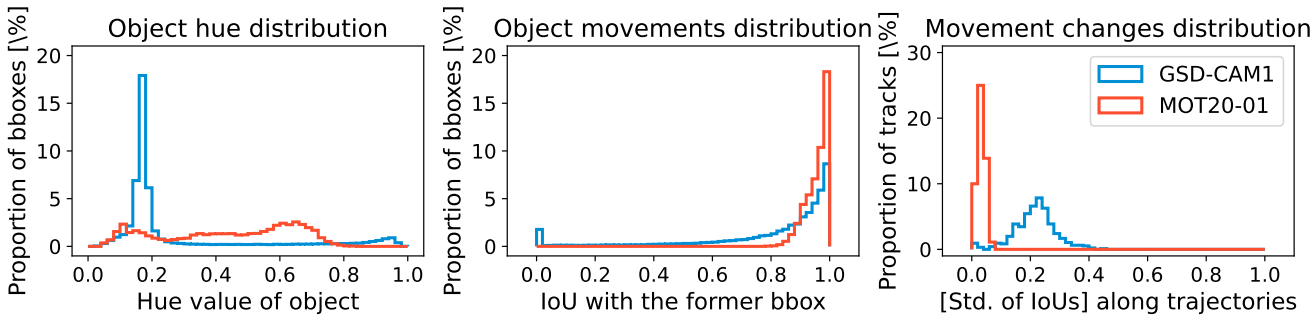

Figure 4. Quantitative comparisons of *GSD-2021-RGB-1* and *MOT20-01*, using the GT annotations. The 1st spectrum shows the distribution of object colors, posed by the Hue value averaged from the center 50% area of the bboxes. The 2nd plot illustrates the distribution of overall object movements, using the IoU as a metric. The 3rd plot presents the standard deviations of the bbox IoU of each trajectory.

Trajectory annotations. Overall, the trajectories of strawberries have an average length of 152 bboxes, yet it ranges from 2 to 600+ bboxes. The extra-long tracks resulted from slower growths under cool temperatures. In fact, there is still a notable proportion of tracks that last less than 20 segments, which are mostly incompatible with the natural growth cycle of strawberries. Two major reasons for these short tracks are: i) the back-layer ones only started to be visible after re-position practices from humans because the strawberries grew in dense clusters; ii) the growths were only partially monitored because the size increases of strawberries might squeeze the others out or into the frames.

3.3. Data Characterization

Compared to pedestrian-focused datasets such as the *MOT20*, *GSD* objects usually are more similar looking to one another, whilst they have more evident appearance changes over frames. In addition, larger and more irregular movements are observed in *GSD* trajectories.

Fig. 4 presents comparisons of the color and movement distribution of the sequence *GSD-2021-RGB-1* (shorten as *RGB-1* in the following text) and *MOT20-01*. The 1st subplot shows the hue value, calculated from the HSV color space [35], of all bboxes. Here, *RGB-1* shows a higher degree of monotonicity among the observations compared to *MOT20-01*, which also indicates larger challenges to the feature extractors. Nevertheless, for the same *GSD* object, the color keeps changing due to its biological development over the time span, together with the illumination condition. An example is shown in Fig. 5.

We measured the object movements by the Intersection of Union (IoU) of observations in adjacent frames because a large proportion of MOT algorithms consider a sequential matching of objects by including more and more frames in analysis. Followingly, larger movements are indicated from the left of the x-axis in the 2nd subplot. As Fig. 4 shows, the movements of *GSD* objects are more spread out, whilst the *MOT20* objects exhibit slower movements, holding a minimum IoU of 0.8. Moreover, there are a few bboxes that have minimal intersections with its previous observation, which introduces extraordinary challenges to the inertia measurement and the association algorithms. We also calculated the standard deviation (std.) of the IoU of each trajectory. The value indicates the irregularity of how each object moves. As the 3rd subplot shows, such irregularity in *RGB-1* is higher in magnitude.

4. Benchmark Studies

Since *GSD* has a large number of high-resolution images, we primarily restricted our attention to lighter, online solvers. In addition, we applied GMTracker [27] on a small subset to exemplify the performance with an offline solver [4].

We assessed the performance of the four MOT algorithms to demonstrate the challenges presented by *GSD*: i) ByteTrack [71] that performs an Intersection of Union (IoU) analysis after applying the Kalman filter as SORT does; ii) Observation-Centric SORT (OC-SORT) [12] that is enhanced against noised and non-linear movements; iii) DeepSORT [62] that introduces appearance descriptions to identify objects before applying the matching by movements; iv) StrongSORT [16] that improves the movement measurement and its balance with the appearance features. On top of the original settings, we altered the appearance-cost parameter (λ) of StrongSORT to introduce different emphases for appearance and motion information in the association stage.

Since all the algorithms share the tracking-by-detection strategy, we present our experiments from three aspects: the overall MOT performance of the four algorithms (and variations), detection-stage impact, and tracking-stage influence. Drawing upon the results, we explore the potential implications stemming from the distinctive characteristics of the *GSD*, which we contend represent challenges within biological development tracking applications.

[4] Our justification for using the subset is provided in the appendix.

4.1. Application of MOT algorithms on *GSD*

By dividing the subsets by cameras, we first trained three YOLOX-x models with a "leave-one-camera-out" cross-validation strategy. We employed the detections on the test set for the MOT performance evaluation. We conducted all experiments using the daytime subset of *GSD-2021* to ensure that darkness-related distractions were avoided, thus enabling a more equitable comparison. We reduced the IoU threshold to 0.1 in the association stages, due to the different object movement patterns as indicated in Fig. 4. We independently developed autoencoders to serve as the appearance descriptors for DeepSORT and StrongSORT. Detailed parameter settings and searching are noted in the supplementary.

We evaluated the overall performance by the widely-known MOT criteria: the Higher Order Tracking Accuracy (HOTA) [36] and the Multi-Object Tracking Accuracy (MOTA) [5]. The performance of track identification is described by accuracy (AssA), recall (AssRe), precision (AssPr), and the balanced criterion IDF1 [51]. We counted the number of identity switches (IDS) and the interruptions of trajectories (Fragmentation/FM) and divided the values by the amount of ground-truth (GT) tracks to compare with other datasets, e.g. *MOT20* or *MOT17*. They are noted as "IDS/Tr" and "FM/Tr" respectively.

4.2. Assessing Comprehensive MOT Performance

The performance metrics are summarized in Tab. 2. In general, the algorithms exhibited inferior performances on *GSD-2021* compared to their achievements on *MOT20*. Notably, compared to more comprehensive metrics such as HOTA and MOTA, all the criteria related to the evaluation of bbox association and trajectory identifications, e.g. IDF1 and AssA, indicate intense performance drops from their original benchmarks. The performance downgrade came with exaggerated frequencies of ID switches and trajectory interruptions. The numbers suggest that the *GSD* tracks have a relatively higher discontinuity as per the MOT algorithms, which could be caused by the increasing changes during the prolonged data collection. The results further evidence that *GSD* introduces a more challenging task than *MOT20* for the state-of-the-art MOT methods.

As shown in Tab. 2, ByteTrack performed the best in terms of HOTA, and OC-SORT was better in limiting the switching of track IDs. When adjusting the parameter λ in StrongSORT to increase the emphasis on motion over appearance matching, notable improvements in overall performance were observed. Hence, associating bounding boxes based on inertia measurements is proved to be relatively more applicable in this case. Nevertheless, we also notice that, whilst shifting the focus to object movements lessened the IDS/Tr, it also led to higher FM/Tr. It indicates that the current appearance-based methods need to be improved to handle data collected at such a sparse frequency.

Upon a dedicated processing time of 112 hours, GMTracker associated the first 750 frames of *RGB-1*. Notably, apart from the training process that already required substantial time and computational memory resources, it devoted over 2 hours to processing some of the frames, with a maximum time of 7498 seconds for a single frame. As evident in Tab. 2, the end-to-end network's performance matched the other benchmarks, yet was achieved by significantly more intensive use of resources [31, 46]. Hence, we remain our focus on the lighter solvers in subsequent discussions.

4.3. Analyzing Detection Performance and Impact

To verify the attainable optimal solution of the object-detection stage, we evaluated two state-of-the-art object detection methods on *GSD*, the anchor-based detector Faster R-CNN and the anchor-free detector YOLOX-x, following the "leave-one-camera-out" strategy. The Average Precision (AP) obtained by both models is noted in Tab. 3.

Due to limitations from the volume and properties of the training data, the detection performances were not so competitive as the private models that were specifically trained for the pedestrian-tracking challenges [16]. However, under a single-category setting, both detectors' performances aligned with the published detections of the *MOT20* testing set [15] and their respective model developers' benchmarks [22, 63]. Although these performances are not directly comparable due to the differences in the validating datasets, we argue that the difficulty level of the object detection task on *GSD* is not significantly higher than other datasets. Therefore, the main challenge brought by *GSD* lies in the association stage, which is also the main task of MOT.

Moreover, for a fair comparison of algorithm performances on *GSD*, we also utilized the metrics obtained from the public *MOT20* detection sets (provided on the MOT20 website [31, 46] As shown in Tab. 2, the MOTA scores achieved using the public *MOT20* detections are even lower than the results obtained on *GSD*. This divergence can be attributed to the limited accuracy of the public detection set. Nevertheless, even when emphasizing track identification metrics like HOTA and IDF1, substantial differences persist. Additionally, the algorithms' IDS/Tr and FM/Tr on *GSD* are still significantly higher compared to those on *MOT20*.

4.4. Decoupling Association from Prior Stages

To compare the specific accuracy of track association regardless of the detection performance, we benchmarked StrongSORT on GT bbox from *GSD-2021* and *MOT20*. For validation, we used *RGB-1* and *MOT20-01* as examples. As shown in Tab. 4, both MOTA were boosted due to the perfect-detection assumption. However, the improvements in HOTA and IDF1 on *RGB-1* experiment were not so significant as those in the *MOT20-01* experiment. Furthermore, noticeable gaps in performance are observed in IDS/Tr and FM/Tr.

Table 2. Performance metrics of the four original and two tailored MOT algorithms on the daytime subset of *GSD-2021* (*the GMTracker was only applied on the first 750 frames of *GSD-2021-RGB-1*). The results are compared with the performance metrics of the same algorithms implemented on the MOT20 test set, using the results with private detections in [12] and [16] and the results with public detections on the MOT20 challenge website. (**The performance of GMTracker was compared with its results on the *MOT17* test set, using the metrics claimed by [27]). The differences are indicated by red and teal texts that are noted at the top right of each value, representing performance degrades and improves, respectively. 'Pvt' and 'Pub' indicate whether the gap is with benchmarks using the private or public detections (and encoders if applicable). If one value is shown, it is compared with only the metrics claimed in the paper, obtained from private detections (and encoders). In terms of StrongSORT, λ is the default weight on the appearance cost, and λ' indicates an altered value.

MOT Algorithm	HOTA	MOTA	IDF1	AssA	AssRe	AssPr	IDS/Tr	FM/Tr
ByteTrack	39.8 Pvt:-21.5 Pub:-16.6	70.7 Pvt:-7.1 Pub:+3.7	39.4 Pvt:-35.8 Pub:-30.8	25.6	29.3	70.0	5.2 Pvt:+4.2 Pub:+4.7	5.4 Pvt:+4.2 Pub:+4.0
OC-SORT	39.7 Pvt:-22.4 Pub:-14.6	68.5 Pvt:-7.0 Pub:+8.6	39.5 Pvt:-36.4 Pub:-27.5	25.9	29.4	72.5	4.5 Pvt:+3.8 Pub:+4.1	5.3 Pvt:+4.4 Pub:+3.4
DeepSORT	34.5 -22.6	49.3 -22.5	33.0 -36.6	22.3	26.4	62.3	8.4 +7.3	5.4
StrongSORT(λ=0.98)	36.1 -25.4	49.3 -22.9	34.0 -41.9	23.9	27.7	64.7	8.8 +7.9	5.1
StrongSORT(λ'=0.5)	38.5	59.9	35.8	25.4	27.9	76.8	6.2	5.8
StrongSORT(λ'=0.02)	38.6	60.4	35.8	25.5	27.9	77.7	6.0	5.9
GMTracker*	37.7	60.2 +4.0**	31.7 -32.1**	22.2	23.2	85.0	20.3 +19.6**	3.8

Table 3. The first three rows show the AP of the detections of *GSD* and the public *MOT20* detections. All values are averaged over the three test sets split by the "leave-one-camera-out" strategy. The latter two rows present the original mAP benchmark for comparison.

Model-Dataset Configuration	AP
YOLOX-x on *GSD*	55.7
Faster R-CNN on *GSD*	55.8
Faster R-CNN on *MOT20* [15]	57.6
Faster R-CNN on COCO [63]	40.2 (mAP)
YOLOX-x on COCO [22]	59.2 (mAP)

The influence of the parameter λ follows a similar pattern as previously described – the emphasis on motion or appearance results in a trade-off between IDS/Tr and FM/Tr. Referring to the data characterization, the higher similarity in appearances among the *GSD* objects and the dynamic variation of them may contribute to the downgraded IDS/Tr performance. Considering that the data was collected over prolonged periods, the incorporation of appearance features is expected to assist in consolidating the fragmented tracklets, e.g. after human activity or overnight. Hence, it is advisable to tailor the utilization of appearance matching in MOT algorithms for scenarios involving sparse frame rates.

4.5. Evaluating Results from one Downstream Application: Growth Curve of Strawberries

One contribution of *GSD* is its provision of valuable information for agriculture practices, enabling precise anticipation of crop growth. Since the natural ripening pattern of strawberries is growing from green to red, we utilized the A* channel from the CIELAB color space [53], which essentially represents the levels of green or magenta. In Fig. 5, the blue curve demonstrates a sample A* variation of the object across frames. Marking associated observations with colored dots and un-associated ones with empty dots, the depicted process is fragmented into five segments by four tracklets suggested by ByteTrack (due to the best HOTA in Tab. 2), involving two IDS in tracklet #21 and #40. Notably, during the crucial period when the strawberry underwent the transition from green to red, which is a crucial factor in determining the timing of harvest, ByteTrack was unable to provide a thorough description of this transformation.

To evaluate the significance of performance deficiency from the perspective of realistic, downstream applications, i.e. tracking the biological development of objects, we set up thresholds to define the "cherry-picked tracks" that record relatively comprehensive monitoring of growth patterns. We chose tracks based on more significant variations of the object's transition from green to red, determined by the changes in the A* channel values in the CIELAB color space, or simply select the tracks with longer lengths. These tracks were considered "more important" ones as they provide more complete information about the growing progress of the crop. We implemented incremental thresholds to perform stricter filtering of their importance.

Fig. 6 discusses the specific performance of ByteTrack, the relatively more capable solution for *GSD*, on the different filtered subsets of *RGB-1*. As is depicted, the recall of track association declined as the track became more comprehensive about the biological development cycle. Simultaneously, there were increases in IDS/Tr and FM/Tr. The track length played a more significant role in the deterioration of performance under this particular scenario.

Table 4. Association-stage performance comparison of StrongSORT, with variations on the appearance-cost parameter λ, applied on the ground-truth detections. We use *G* and *M* to represent *GSD-RGB-1* and *MOT20-01* respectively in this table. In all experiments, the ground-truth locations of the bboxs were used, such that the algorithm performance was not influenced by the detection accuracy.

	HOTA		MOTA		IDF1		IDS/Tr		FM/Tr	
Algorithm	G	M	G	M	G	M	G	M	G	M
StrongSORT(λ=0.98)	51.5	98.6	83.5	99.5	43.5	98.6	6.3	0.0	3.7	0.0
StrongSORT(λ'=0.5)	51.4	99.3	83.6	99.5	42.5	99.4	5.9	0.0	4.6	0.0
StrongSORT(λ'=0.02)	52.2	99.2	85.1	99.5	42.7	99.4	5.2	0.0	4.2	0.0

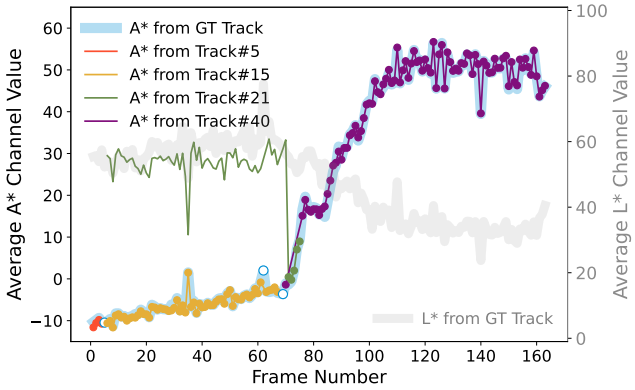

Figure 5. The color change of an example strawberry under the GT trajectory and the ByteTrack results. The x-axis indicates the sequence of frames. The y-axes are for the average A* values (scales on the left) and L* values (scales on the right) of the observations. The blue and gray translucent strokes illustrate the value of the GT annotations. The lines with filled dots are identified observations by ByteTrack, which are color-coded to indicate each track ID. If the object in one frame is not associated with any of the tracks, we put an empty dot on the A* curve from GT.

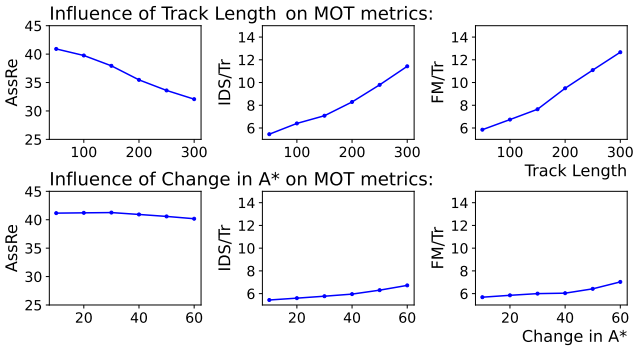

Figure 6. MOT performance change by selection criteria of trajectory subsets, demonstrated by recall (1st column), ID switching (2nd col.), and times of fragments (3rd col.) of tracklets. The first row illustrates the impact on the performance metrics when the tracks were filtered by different minimum lengths. Experiments for the second row selected the tracks by the differences of the average A* value of the last three and the first three bboxs.

Viewing from an application-oriented standpoint, the growth-tracking task also targets monitoring pivotal stages when fruits are ripening swiftly. Therefore, it is argued that there is potential for advancing state-of-the-art MOT algorithms, particularly in accurately identifying and associating objects within similar biological development processes.

5. Conclusion

With this paper, we propose a fully-annotated dataset that tracks the growth of in total of 3528 strawberries over 30 weeks in 2021 and 32 weeks in 2022 in two different greenhouses: *The Growing Strawberries* Dataset (*GSD*). It reveals a unique Multiple-Object-Tracking (MOT) challenge – following biologically developing instances over a prolonged period. In *GSD*, progressive appearance change and irregular movements are captured from the longitudinal observations of cultivation practices. For example, human interference with the sparse frame rate introduced drastically non-linear movement, which is challenging for many algorithms.

We benchmarked the performance of four online MOT algorithms on *GSD*. The obtained result metrics highlight the need for advancing MOT methods, particularly in associating the bounding-box association for long-term MOT tasks. The tracking continuity was affected by both appearance changes and diverse object motions, which also presented a trade-off when fine-tuning StrongSORT. Furthermore, an offline algorithm demonstrated the computational effort required to handle a large dataset such as *GSD*, yet achieving similar metrics. In summary, the results call for algorithms that could improve track associations while utilizing the features properly and efficiently.

Essentially, biological development is the principal property that makes the *GSD* challenge unique, but it can also provide insights for other long-term MOT tasks. For instance, monitoring other processes with incremental changes, such as cellular growth and corrosion expansion, etc. The information provided by more than the visual spectrum is also supportive of plant science [9, 50]. The *GSD* challenge highlights the need for reliable methods to handle in-the-wild data imperfections. The inevitable real-world challenges point out potential future research for robust data utilization.

References

[1] Mykhaylo Andriluka, Stefan Roth, and Bernt Schiele. People-tracking-by-detection and people-detection-by-tracking. In *2008 IEEE Conference on computer vision and pattern recognition*, pages 1–8. IEEE, 2008. 3

[2] Samreen Anjum and Danna Gurari. CTMC: Cell Tracking with Mitosis Detection Dataset Challenge. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, volume 2020-June, pages 4228–4237. IEEE, 6 2020. 1

[3] Rina Bao, Noor M Al-Shakarji, Filiz Bunyak, and Kannappan Palaniappan. Dmnet: Dual-stream marker guided deep network for dense cell segmentation and lineage tracking. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3361–3370, 2021. 1

[4] Philipp Bergmann, Tim Meinhardt, and Laura Leal-Taixe. Tracking Without Bells and Whistles. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 941–951. IEEE, 10 2019. 3

[5] Keni Bernardin and Rainer Stiefelhagen. Evaluating Multiple Object Tracking Performance: The CLEAR MOT Metrics. *EURASIP Journal on Image and Video Processing*, 2008:1–10, 2008. 6

[6] Alex Bewley, Zongyuan Ge, Lionel Ott, Fabio Ramos, and Ben Upcroft. Simple online and realtime tracking. In *2016 IEEE International Conference on Image Processing (ICIP)*, volume 2016-August, pages 3464–3468. IEEE, 9 2016. 3

[7] Sergey Bezryadin, Pavel Bourov, and Dmitry Ilinih. Brightness Calculation in Digital Image Processing. *International Symposium on Technologies for Digital Photo Fulfillment*, 1(1):10–15, 1 2007. 4

[8] Guillem Brasó and Laura Leal-Taixé. Learning a neural solver for multiple object tracking. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6246–6256, 2020. 3

[9] J.K. Brecht, F.E. Loaza, M.C.N. Nunes, J.P. Emond, I. Uysal, F. Badia, J. Wells, and J. Saenz. Reducing strawberry waste and losses in the postharvest supply chain via intelligent distribution management. *Acta Horticulturae*, 1120(1120):253–260, 7 2016. 2, 8

[10] Filiz Bunyak, Kannappan Palaniappan, Sumit Kumar Nath, TL Baskin, and Gang Dong. Quantitative cell motility for in vitro wound healing using level set-based active contour tracking. In *3rd IEEE International Symposium on Biomedical Imaging: Nano to Macro, 2006.*, pages 1040–1043. IEEE, 2006. 1

[11] Jinzheng Cai, Youbao Tang, Ke Yan, Adam P. Harrison, Jing Xiao, Gigin Lin, and Le Lu. Deep Lesion Tracker: Monitoring Lesions in 4D Longitudinal Imaging Studies. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15154–15164. IEEE, 6 2021. 1, 3

[12] Jinkun Cao, Jiangmiao Pang, Xinshuo Weng, Rawal Khirodkar, and Kris Kitani. Observation-centric sort: Rethinking sort for robust multi-object tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9686–9696, 2023. 3, 5, 7

[13] Long Chen, Haizhou Ai, Chong Shang, Zijie Zhuang, and Bo Bai. Online multi-object tracking with convolutional neural networks. In *2017 IEEE International Conference on Image Processing (ICIP)*, volume 2017-September, pages 645–649. IEEE, 9 2017. 3

[14] Achal Dave, Tarasha Khurana, Pavel Tokmakov, Cordelia Schmid, and Deva Ramanan. TAO: A Large-Scale Benchmark for Tracking Any Object. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 12350 LNCS, pages 436–454. Springer Science and Business Media Deutschland GmbH, 5 2020. 1, 2

[15] Patrick Dendorfer, Hamid Rezatofighi, Anton Milan, Javen Shi, Daniel Cremers, Ian Reid, Stefan Roth, Konrad Schindler, and Laura Leal-Taixé. Mot20: A benchmark for multi object tracking in crowded scenes. *arXiv preprint arXiv:2003.09003*, 2020. 2, 6, 7

[16] Yunhao Du, Zhicheng Zhao, Yang Song, Yanyun Zhao, Fei Su, Tao Gong, and Hongying Meng. Strongsort: Make deepsort great again. *IEEE Transactions on Multimedia*, 2023. 3, 5, 6, 7

[17] Biyanka Ekanayake, Johnny Kwok-Wai Wong, Alireza Ahmadian Fard Fini, and Peter Smith. Computer vision-based interior construction progress monitoring: A literature review and future research directions. *Automation in Construction*, 127:103705, 7 2021. 2

[18] Matteo Fabbri, Guillem Braso, Gianluca Maugeri, Orcun Cetintas, Riccardo Gasparini, Aljosa Osep, Simone Calderara, Laura Leal-Taixe, and Rita Cucchiara. MOTSynth: How Can Synthetic Data Help Pedestrian Detection and Tracking? In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10829–10839. IEEE, 10 2021. 2

[19] Takuya Fujinaga, Shinsuke Yasukawa, and Kazuo Ishii. Tomato Growth State Map for the Automation of Monitoring and Harvesting. *Journal of Robotics and Mechatronics*, 32(6):1279–1291, 12 2020. 1

[20] Joaquin Gabaldon, Ding Zhang, Lisa Lauderdale, Lance Miller, Matthew Johnson-Roberson, Kira Barton, and K Alex Shorter. Computer-vision object tracking for monitoring bottlenose dolphin habitat use and kinematics. *PloS one*, 17(2):e0254323, 2022. 1

[21] Zongmei Gao, Yuanyuan Shao, Guantao Xuan, Yongxian Wang, Yi Liu, and Xiang Han. Real-time hyperspectral imaging for the in-field estimation of strawberry ripeness with deep learning. *Artificial Intelligence in Agriculture*, 4:31–38, 1 2020. 2, 3

[22] Zheng Ge, Songtao Liu, Feng Wang, Zeming Li, and Jian Sun. Yolox: Exceeding yolo series in 2021. *arXiv preprint arXiv:2107.08430*, 2021. 6, 7

[23] A. Geiger, P. Lenz, and R. Urtasun. Are we ready for autonomous driving? The KITTI vision benchmark suite. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3354–3361. IEEE, 6 2012. 1, 2

[24] Jonathon A Gibbs, Alexandra J Burgess, Michael P Pound, Tony P Pridmore, and Erik H Murchie. Recovering wind-induced plant motion in dense field environments via deep learning and multiple object tracking. *Plant physiology*, 181(1):28–42, 2019. 1

[25] Shuman Guo, Shichang Wang, Zhenzhong Yang, Lijun Wang, Huawei Zhang, Pengyan Guo, Yuguo Gao, and Junkai Guo. A review of deep learning-based visual multi-object tracking algorithms for autonomous driving. *Applied Sciences*, 12(21):10741, 2022. 1

[26] Esmael Hamuda, Brian Mc Ginley, Martin Glavin, and Edward Jones. Improved image processing-based crop detection using kalman filtering and the hungarian algorithm. *Computers and electronics in agriculture*, 148:37–44, 2018. 1

[27] Jiawei He, Zehao Huang, Naiyan Wang, and Zhaoxiang Zhang. Learnable graph matching: Incorporating graph partitioning with deep feature learning for multiple object tracking. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5295–5305, 2021. 5, 7

[28] Sha Huang, Lina Tang, Joseph P Hupy, Yang Wang, and Guofan Shao. A commentary review on the use of normalized difference vegetation index (ndvi) in the era of popular remote sensing. *Journal of Forestry Research*, 32(1):1–6, 2021. 3

[29] Andrew Hui, Yannie O Y Soo, Vincent C T Mok, Ka Sing Lawrence Wong, and Andrew CF Hui Patrick Kwan TW Leung Y Soo Vincent CT Mok Lawrence KS Wong. Diagnostic value and safety of long-term video-EEG monitoring. *Hong Kong Medical Journal*, 13(3), 2007. 1

[30] Diego M Jiménez-Bravo, Álvaro Lozano Murciego, André Sales Mendes, Héctor Sánchez San Blás, and Javier Bajo. Multi-object tracking in traffic environments: A systematic literature review. *Neurocomputing*, 2022. 1

[31] kalman_pub. Bytetrack: Multi-object tracking by associating every detection box, 2022. Last submitted on November 23, 2021. 6

[32] In Su Kim, Hong Seok Choi, Kwang Moo Yi, Jin Young Choi, and Seong G Kong. Intelligent visual surveillance—a survey. *International Journal of Control, Automation and Systems*, 8(5):926–939, 2010. 1

[33] Nikolas Lamb and Mooi Choo Chuah. A Strawberry Detection System Using Convolutional Neural Networks. In *2018 IEEE International Conference on Big Data (Big Data)*, pages 2515–2520. IEEE, 12 2018. 2

[34] Xin Li, Kejun Wang, Wei Wang, and Yang Li. A multiple object tracking method using kalman filter. In *The 2010 IEEE international conference on information and automation*, pages 1862–1866. IEEE, 2010. 1

[35] Martin Loesdau, Sébastien Chabrier, and Alban Gabillon. Hue and saturation in the rgb color space. In Abderrahim Elmoataz, Olivier Lezoray, Fathallah Nouboud, and Driss Mammass, editors, *Image and Signal Processing*, pages 203–212, Cham, 2014. Springer International Publishing. 5

[36] Jonathon Luiten, Aljosa Osep, Patrick Dendorfer, Philip Torr, Andreas Geiger, Laura Leal-Taixé, and Bastian Leibe. HOTA: A Higher Order Metric for Evaluating Multi-object Tracking. *International Journal of Computer Vision*, 129(2):548–578, 2 2021. 6

[37] Chenxu Luo, Xiaodong Yang, and Alan Yuille. Exploring simple 3d multi-object tracking for autonomous driving. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10488–10497, 2021. 1

[38] Wenhan Luo, Junliang Xing, Anton Milan, Xiaoqin Zhang, Wei Liu, and Tae-Kyun Kim. Multiple object tracking: A literature review. *Artificial Intelligence*, 293:103448, 2021. 1

[39] Wenhan Luo, Xiaowei Zhao, and Tae-Kyun Kim. Multiple object tracking: A review. *arXiv preprint arXiv:1409.7618*, 1:1, 2014. 1

[40] Oka Mahendra, Hilman F. Pardede, Rika Sustika, and R. Budiarianto Suryo Kusumo. Comparison of Features for Strawberry Grading Classification with Novel Dataset. In *2018 International Conference on Computer, Control, Informatics and its Applications (IC3INA)*, pages 7–12. IEEE, 11 2018. 2

[41] Akshay Mangawati, Mohammed Leesan, HV Ravish Aradhya, et al. Object tracking algorithms for video surveillance applications. In *2018 International Conference on Communication and Signal Processing (ICCSP)*, pages 0667–0671. IEEE, 2018. 1

[42] Erik Meijering, Oleh Dzyubachyk, Ihor Smal, and Wiggert A. van Cappellen. Tracking in cell and developmental biology. *Seminars in Cell & Developmental Biology*, 20(8):894–902, 10 2009. 1

[43] Anton Milan, Laura Leal-Taixé, Ian Reid, Stefan Roth, and Konrad Schindler. Mot16: A benchmark for multi-object tracking. *arXiv preprint arXiv:1603.00831*, 2016. 1, 2

[44] MOTChallenge. MOTChallenge - multiple object tracking benchmark, 2023. Accessed on August 22, 2023. 2

[45] Ladan Najafizadeh and Jon E. Froehlich. A Feasibility Study of Using Google Street View and Computer Vision to Track the Evolution of Urban Accessibility. In *Proceedings of the 20th International ACM SIGACCESS Conference on Computers and Accessibility*, pages 340–342, New York, NY, USA, 10 2018. ACM. 2

[46] OCSORTpublic. Observation-centric sort with public detections, 2022. Last submitted on March 29, 2022. 6

[47] Jonas Osterloff, Ingunn Nilssen, and Tim W. Nattkemper. A computer vision approach for monitoring the spatial and temporal shrimp distribution at the LoVe observatory. *Methods in Oceanography*, 15-16:114–128, 4 2016. 2

[48] Matti Pastell, Lemsalu. Madis, and Victor Bloch. Strawberry dataset for object detection, 2 2022. 2

[49] Isaac Pérez-Borrero, Diego Marín-Santos, Manuel E. Gegúndez-Arias, and Estefanía Cortés-Ancos. A fast and accurate deep learning method for strawberry instance segmentation. *Computers and Electronics in Agriculture*, 178:105736, 11 2020. 2

[50] M. Moshiur Rahman, M. Moniruzzaman, Munshi Rashid Ahmad, B.C. Sarker, and M. Khurshid Alam. Maturity stages affect the postharvest quality and shelf-life of fruits of strawberry genotypes growing in subtropical regions. *Journal of the Saudi Society of Agricultural Sciences*, 15(1):28–37, 1 2016. 2, 8

[51] Ergys Ristani, Francesco Solera, Roger Zou, Rita Cucchiara, and Carlo Tomasi. Performance Measures and a Data Set for Multi-target, Multi-camera Tracking. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 9914 LNCS, pages 17–35. Springer Verlag, 2016. 6

[52] Amir Roshan Zamir, Afshin Dehghan, and Mubarak Shah. Gmcp-tracker: Global multi-object tracking using generalized

minimum clique graphs. In *Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part II 12*, pages 343–356. Springer, 2012. 3

[53] Janos Schanda. Cie colorimetry. *Colorimetry: Understanding the CIE system*, pages 25–78, 2007. 7

[54] C.P. Schofield, J.A. Marchant, R.P. White, N. Brandl, and M. Wilson. Monitoring Pig Growth using a Prototype Imaging System. *Journal of Agricultural Engineering Research*, 72(3):205–210, 3 1999. 1

[55] Jamie Sherrah and Shaogang Gong. Tracking Discontinuous Motion Using Bayesian Inference. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 1843, pages 150–166. Springer Verlag, 2000. 3

[56] Ramana Sundararaman, Cedric De Almeida Braga, Eric Marchand, and Julien Pettre. Tracking Pedestrian Heads in Dense Crowd. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3864–3874. IEEE, 6 2021. 2, 3

[57] Adam Van Etten, Daniel Hogan, Jesus Martinez Manso, Jacob Shermeyer, Nicholas Weir, and Ryan Lewis. The Multi-Temporal Urban Development SpaceNet Dataset. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6394–6403. IEEE, 6 2021. 2

[58] Avantika Vardhan, Marcel Prastawa, Neda Sadeghi, Clement Vachet, Joseph Piven, and Guido Gerig. Joint Longitudinal Modeling of Brain Appearance in Multimodal MRI for the Characterization of Early Brain Developmental Processes. In *LNCS*, volume 8682, pages 49–63. Springer, Cham, 2015. 1

[59] Zhongdao Wang, Liang Zheng, Yixuan Liu, Yali Li, and Shengjin Wang. Towards Real-Time Multi-Object Tracking. In *Computer Vision – ECCV 2020*, pages 107–122, 2020. 3

[60] Mark Weber, Jun Xie, Maxwell Collins, Yukun Zhu, Paul Voigtlaender, Hartwig Adam, Bradley Green, Andreas Geiger, Bastian Leibe, Daniel Cremers, et al. Step: Segmenting and tracking every pixel. *arXiv preprint arXiv:2102.11859*, 2021. 2

[61] Longyin Wen, Zhen Lei, Ming-Ching Chang, Honggang Qi, and Siwei Lyu. Multi-Camera Multi-Target Tracking with Space-Time-View Hyper-graph. *International Journal of Computer Vision*, 122(2):313–333, 4 2017. 3

[62] Nicolai Wojke, Alex Bewley, and Dietrich Paulus. Simple online and realtime tracking with a deep association metric. In *2017 IEEE International Conference on Image Processing (ICIP)*, volume 2017-September, pages 3645–3649. IEEE, 9 2017. 3, 5

[63] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2, 2019. 6, 7

[64] Shuyuan Xu, Jun Wang, Xiangyu Wang, and Wenchi Shou. Computer Vision Techniques in Construction, Operation and Maintenance Phases of Civil Assets: A Critical Review. In *36th International Symposium on Automation and Robotics in Construction (ISARC 2019)*, 5 2019. 2

[65] Yingkun Xu, Xiaolong Zhou, Shengyong Chen, and Fenfen Li. Deep learning for multiple object tracking: a survey. *IET Computer Vision*, 13(4):355–368, 2019. 1

[66] Hao Yang, Fangle Chang, Yuhang Huang, Ming Xu, Yangfan Zhao, Longhua Ma, and Hongye Su. Multi-object tracking using deep sort and modified centernet in cotton seedling counting. *Computers and Electronics in Agriculture*, 202:107339, 2022. 1

[67] Linjie Yang, Yuchen Fan, and Ning Xu. Video Instance Segmentation. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5187–5196. IEEE, 10 2019. 1, 2

[68] Alper Yilmaz, Omar Javed, and Mubarak Shah. Object tracking: A survey. *Acm computing surveys (CSUR)*, 38(4):13–es, 2006. 1

[69] Yang Yu, Kailiang Zhang, Li Yang, and Dongxing Zhang. Fruit detection for strawberry harvesting robot in non-structural environment based on Mask-RCNN. *Computers and Electronics in Agriculture*, 163:104846, 8 2019. 2

[70] Chu zhang, Chentong Guo, Fei Liu, Wenwen Kong, Yong He, and Binggan Lou. Hyperspectral imaging analysis for ripeness evaluation of strawberry with support vector machine. *Journal of Food Engineering*, 179:11–18, 6 2016. 1, 2, 3

[71] Yifu Zhang, Peize Sun, Yi Jiang, Dongdong Yu, Fucheng Weng, Zehuan Yuan, Ping Luo, Wenyu Liu, and Xinggang Wang. ByteTrack: Multi-object Tracking by Associating Every Detection Box. In *European Conference on Computer Vision 2022*, pages 1–21. Springer, Cham, 2022. 3, 5

[72] Zikai Zhang, Bineng Zhong, Shengping Zhang, Zhenjun Tang, Xin Liu, and Zhaoxiang Zhang. Distractor-Aware Fast Tracking via Dynamic Convolutions and MOT Philosophy. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1024–1033. IEEE, 6 2021. 3

[73] Bin Zhou, Xuemei Duan, Dongjun Ye, Wei Wei, Marcin Woźniak, Dawid Połap, and Robertas Damaševičius. Multi-Level Features Extraction for Discontinuous Target Tracking in Remote Sensing Image Monitoring. *Sensors*, 19(22):4855, 11 2019. 3

[74] Zikun Zhou, Jianqiu Chen, Wenjie Pei, Kaige Mao, Hongpeng Wang, and Zhenyu He. Global Tracking via Ensemble of Local Trackers. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8751–8760. IEEE, 6 2022. 3

[75] Zhen Zhou, Yan Huang, Wei Wang, Liang Wang, and Tieniu Tan. See the Forest for the Trees: Joint Spatial and Temporal Recurrent Neural Networks for Video-Based Person Re-identification. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6776–6785. IEEE, 7 2017. 3