Smart Agricultural Technology 8 (2024) 100506

Contents lists available at ScienceDirect

Smart Agricultural Technology

journal homepage: www.journals.elsevier.com/smart-agricultural-technology

ASPEN study case: Real time *in situ* apples detection and characterization

Camilo Chiang [a,*], Alice Monney [b], Phillipe Monney [b], Danilo Christen [b]

[a] Agroscope, Digital Production, Tänikon 1 8356 Ettenhausen, Switzerland
[b] Agroscope, Institute for Plant Production Systems, Rue des Eterpys 18 1964 Conthey, Switzerland

ARTICLE INFO

Keywords:
Machine learning
YOLOv8
Apple
SLAM
Terrestrial laser scanner

ABSTRACT

Due to an increasing demand for food and pressures on our ecosystem, mechanisation and automation in agriculture has been proposed as one of the main solutions to the problems associated with overpopulation given today's life standards. To encourage the use of new technologies and bridge the gap between plant and computer science, here we validate an open-source pipeline capable of predicting real time *in situ* fruit characteristics, specifically in this case for apples. Using Agroscope's phenotyping tool (ASPEN), we achieve an average precision at intercept over union of 50 % of 0.75 when using YOLOv8 - m as the object detection algorithm, and with thanks to the use of multiple sensors, we find an average diameter error of 4.4 mm in the task of apple size determination. Our research demonstrates that although the pipeline tends to underestimate the actual fruit size, size estimation cannot only be used to determine the size of apples per scan, but also to track temporal apple size distribution in 4 different varieties. This research supports ASPEN in potentially replacing traditional field measurements, also suggesting that other traits could also be digitally measured for standard orchard phenotyping, either for scientific or agricultural output goals. Finally, we make publicly available a new dataset of more than 600 images (Agroscope_apple dataset) and a pre-trained model based on YOLOv8 and specifically trained for the in-situ apple detection task. By doing so, we hope to increase the accessibility and use of new technologies in the field of agriculture.

Introduction

Due to an ageing and growing population [1] in conjunction with the current anthropogenically induced climate crisis [2], our food system is challenged to meet future demand while reducing pressure on our ecosystems (*e.g.* [3]). These challenges have produced a recent surge in more efficient methods, and in particular for those based on remote sensing technologies and automation. Although the long-term effects of these will only be known in the future [4], some technologies already show positive effects in agriculture. For example, the selection of plant genotypes based on trait descriptions, phenotyping, and specific conditions that highlight plant traits, especially under stress, are one of the most important procedures used in agricultural research today [5]. While phenotyping is not a new technique, is still a labour-intensive task [5].

For the adequate selection of apple varieties in agricultural contexts, one of the most consumed fruits (*i.e.* [6]), a considerable amount of time is required in phenotyping due to its perennial cycle and its height, making it a difficult manual task. Considering the basic subtask of fruit counting, several authors proposed the use of images. Image analysis for fruit detection was initially explored using several approaches: shape based as demonstrated by Whittaker et al. [7] using Hough transformation algorithms [8], colour segmentation techniques as Bulanon and Kataoka [9] demonstrate with its limitations: *e.g.* green apples are significantly harder to identify compared with red apples (*e.g.* [10]), or texture based detection techniques, as leaves, fruits, and branches present different textures or combinations of these different techniques (*e.g.* [11]). Independently, these techniques have demonstrated limitations, especially in full tree images, as fruits are relatively small objects and all of the methods include highly manually fine-tuned parameters.

Since the introduction of algorithms-based on neural networks (NN), not only the robustness and generalization of the object detection task has increased, but there has also been a reduction in processing time. Liu et al. [12] demonstrates that when using a fully convolutional neural network (FCN) they could detect, localize, and track either apples or oranges along multiple images. A remarkable case is the usage of region-based convolutional neural networks (R-CNN), as many have demonstrated apple detection with high precision when using variations

* Corresponding author.
E-mail address: camilo.chiang@agroscope.admin.ch (C. Chiang).

https://doi.org/10.1016/j.atech.2024.100506
Received 11 April 2024; Received in revised form 12 June 2024; Accepted 11 July 2024
Available online 19 July 2024

of R-CNN. Gené-Mola et al., [13] demonstrates that when using Mask R-CNN in tilled apples orchard images, they could reach an average precision rate at interception over union 0.5 (AP@0.5) of 0.859 in a 288 images dataset. Moreover, when using full tree images, Häni et al. [14] reach an AP@0.5 of 0.775 and 0.763 when using either faster R-CNN or Mask R-CNN respectively, where in both cases, smaller objects of less than 32 pixels (px) were harder to detect with an AP of 0.296 in images of a resolution of 1290 × 720 px. Despite higher accuracy and robustness than previous methods before NN, none of the algorithms have the capacity to work in real time, or in other words, to process more than 30 images per second when working with RGB images bigger than a resolution of 400 × 400 px. With the introduction of the YOLO algorithm [15], exhibiting a simpler architecture than previous NN, near real time results were reported for first time in the object detection task in an ordinary desktop computer equipped with a graphical processing unit (GPU). Continuous improvements in this family of algorithms (*i.e.* [16] and [17]), allowed us to reach real time object detection with similar precision to that of R-CNN, allowing many researchers to develop variations of these algorithms tailored to different crops *i.e.* apples [18], banana [19], cherry [20], tomatoes [21], lemons [22], or mango [23] among others. In 2020, Kuznetsova et al., demonstrated that when using the algorithm YOLOv3 they reach a detection rate in apple images of 0.908 with a precision of 0.922 and an average speed of 19 ms (*circa* 52 frames per second, FPS) per image when using images of resolution 416 × 416 px.

Later in 2021, the same research group also demonstrated that when using YOLOv5 trained in the same dataset, they could reach a higher recall of 0.978 [24]. Yan et al., [18] shows that using either YOLOv5s or a modified version specifically optimized for a apple detection task, their AP was 0.817 and 0.867 respectively, where their specific network was slightly slower than compared with YOLOv5s (76 vs 66 FPS). To translate these results to the field in real time and in a reproducible way, several challenges need to be addressed, for example: working with higher amounts of data (*e.g.* [25]), optimization of the increasing computation power [26], and standardizing benchmark datasets are crucial for reliable comparisons and advances in the field. Many of the previously mentioned examples use either close-up images or private images datasets, this makes it harder to compare these results even though apple datasets for object detection are available (*e.g.* [14,27,28]).

To translate 2D object detections to *in situ* yield estimation, 3D localization must demonstrate better results than 2D detections (*i.e.* [12,13] and [29]) as it allows us to filter for the background and remove false negatives. To achieve 3D detections, multiple pipelines are available, each with its unique trade-offs. Here we highlight two specific cases: 1) motion-enabled 2D object detection and 3D sensors object detection. Motion-enabled 2D object detection allows for 3D estimations using information from 2D frames. Roy et al., [30] shows that thanks to a semi-supervised colour-based algorithm together with camera motion to merge the multiple views (homography), it was possible to count the number of apples and estimate the fruit yield with an overall accuracy of 0.91 to 0.94 precision rate across their different datasets. Similarly, Häni et al., [31] highlights that using a combination of semi supervised methods plus NN based methods for fruit detection (2D) and camera motion, they could reach yield accuracies in the range of 0.956 to 0.978, depending on the dataset. In a similar setup and results, using structure from motion (sfm) from both sides of a row of apple trees, Gené-Mola et al., [13] processed 11 trees with 582 images in a total of 345 min. Although a 2D analysis was carried out for localizing apples, 3D information was mainly used for filtering false negatives, and this improved the quality of the detections.

Regarding 3D sensors, frequently light detection and ranging (LiDAR) technology has been successfully used for yield estimation. While this technology has been already utilized for many years, especially in mining and forestry, its use in agriculture only has increased in the last couple of years mainly due to a reduction in the price of equipment [32]. For instance, using the different reflectance properties of fruits, an equivalent method to colour filtering in 2D, Gené-Mola et al., [33] identified up to 0.824 fraction of apples with a 0.104 false detection rate of apples present in LiDAR 3D based dataset with an average time of 9.6 s per tree when using their DBSCAN based algorithm in the LiDAR data. In *a posteri* work, they reach a higher identification rate of 0.87 when using LiDAR in combination with applications of force airflow [29]. Is important to consider that in both cases a mobile terrestrial laser scanner (MTLS) was used together with a real-time kinematics global navigation system (RTK-GNSS) in a post processing step. Comparable to the case of RGB images, although there are datasets available that could be used for apple detection in 3D point clouds using supervised learning (*i.e.* [34]) these have been rarely used mainly due to today's neural networks for 3D detection are currently in an emerging stage (*e.g.* [35]), have lower detection rates possibly due to lacks in training data, higher computational requirements than 2D methods, are highly dependent on data which can vary in quality from sensor to sensor. In addition to the variation due to the used scanning methodology [36] and the required usage of pre-defined point cloud size, which forces the point cloud to be subdivided and may introduce errors and slower processing times [37].

While previous studies reach high accuracies when using either 2D or 3D data, the reliance on a single sensor could lead to failures in the pipeline, *e.g.* if no overlap exists between two consecutive images, or if there is a lack of features due to illumination issues as an example, or the nature of a non-structured environment which may motion estimation between frames or scans impossible to convert. As an alternative, the fusion of multiple sensors has been proposed, where the usage of LiDAR in fusion with RGB cameras during yield estimation has been already successfully demonstrated in post processing (*i.e.* [38,39]) or recently in real time using simultaneous localization and mapping algorithms (SLAM, *e.g.* [40]). Once this bottleneck problem has been solved in terms of working in real time for object detection and 3D reconstruction, the phenotyping capabilities of such a pipeline need to be evaluated. In the present study we aim to present for first time in the apple crop, a pipeline that can run real time *in situ* phenotyping, specifically to measure apple size along the growing period. For this task, we use the Agroscope phenotyping tool (ASPEN). Unlike our previous introductory work to the tool [40] we highlight the outputs as a function of time. To increase reproductively of our results, we make publicly available all outputs of this research, including a new dataset labelled for apple detection (Agroscope_apple dataset) and a pre-trained model for object detection.

Materials and methods

For a real time *in situ* characterization of an apple orchard, an ASPEN unit was used. For a detailed description of ASPEN, the reader is referred to the online project's repository[1] and to our previous publication [40]. In brief, ASPEN is a 3D reconstruction tool engineered for agricultural scenarios and even more, it is specifically designed to be compatible with other tools used in the field of agricultural research (*i.e.* multispectral cameras), which rely on an embedded platform based in Jetson Xavier NX, using ROS in Ubuntu 18.04 LTS environment and robotic operating system (ROS, melodic) and multiple sensors, including a LiDAR sensor (Mid70, Livox), a RGBD camera (D415 Realsense, Intel) and an inertial measurement unit (IMU; BMI088, Bosh). These sensors are then used for the 3D reconstruction of the scanned environment using a SLAM algorithm and fruit localization, tracking, and characterization. In contrast to our previous work, FAST-LIVO [41] was selected as a SLAM algorithm as it was more robust than the previously used SLAM (R3Live; [42]) under a less structured environment than in our preceding research (data non shown). For the purposes of this publication, we here highlight the details of the ASPEN pipeline when

[1] www.github.com/camilochiang/ASPEN

using ASPEN in apple orchards, with special emphasis in object detection and object characterization.

Object detection and characterization

To train an object detection algorithm *in situ* orchard apple detection the publicly available dataset Minneapple was used [14] together with our own dataset (Agroscope_apple). The Minneapple dataset consists of

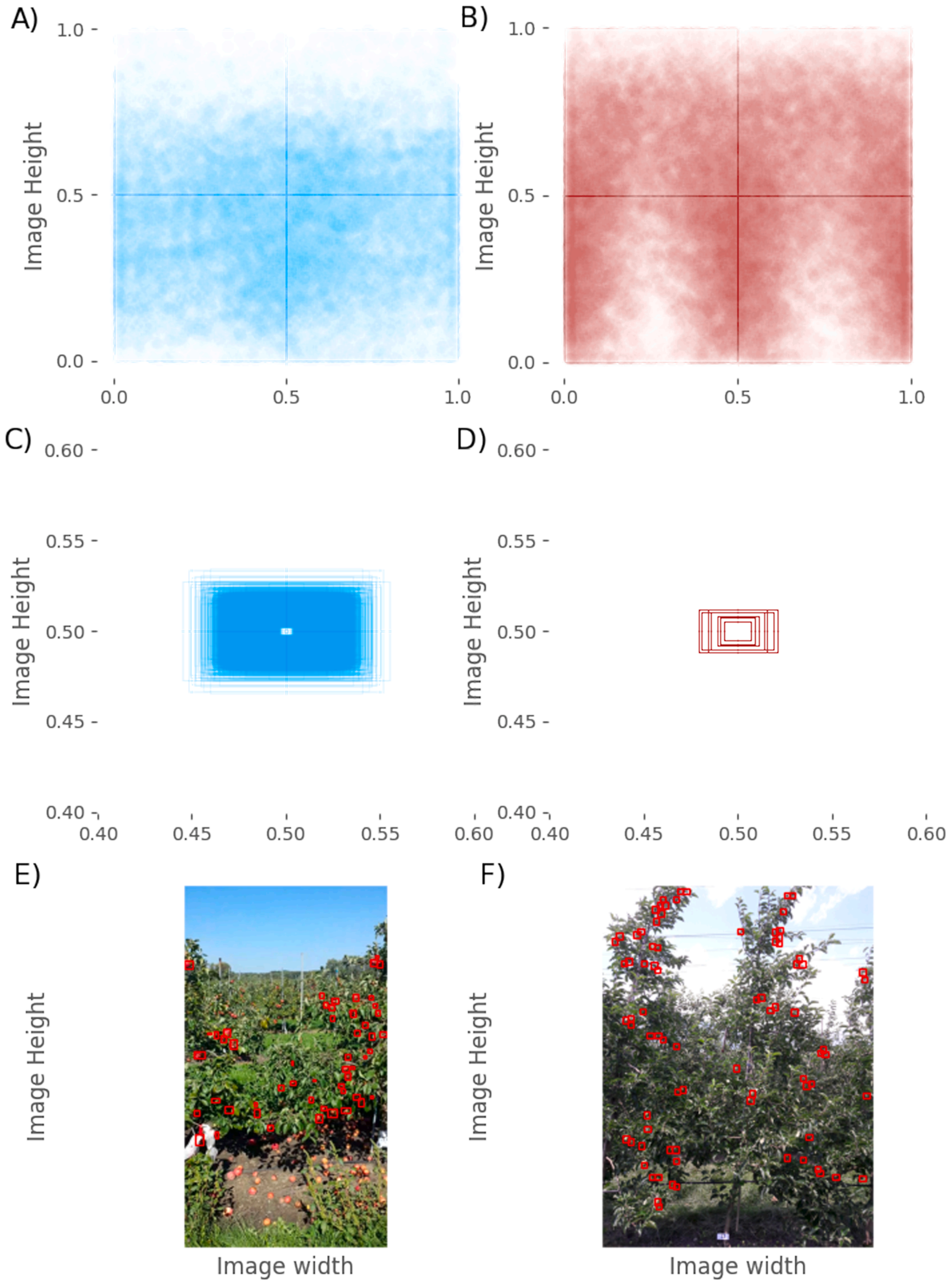


Fig. 1. Comparison of the Minapple dataset (A, C and E) against the newest and Agroscope_apple dataset (B, D and F). A and B (shown by the object distribution along the differently normalized frames; C and D) the difference in size of the multiple annotations along the image frames (normalized), and E and F, a random selection of images from each dataset to show the used criteria for annotation regarding either the Minapple dataset or the Agroscope_apple dataset respectively, where each red square correspond to an apple instance.

1000 high-resolution images of full trees in orchards with more than 41, 000 instances annotated for instance segmentation, which was transformed for the object detection task using a python script. Agroscope_apple is a new dataset of 623 images of apples trees with more than 55,000 annotations that were captured at the research site of Agroscope (Conthey, Switzerland) in previous years. All new images correspond to full tree images with at least one tree in each image with different degrees of illumination. The images were captured 2–3 m from the tree row, where each image contains one tree in the image center and multiple trees at either side or in the background. Trees were selected with the aim of increasing image variety, therefore apples from different sizes and colours are present in the dataset. More details are present in Fig. 1. The images were captured using a Samsung sm-a510F cell phone at two different resolutions: 2448 × 3264 px and 3096 × 4128 px. In contrast to the Minneapple dataset, we chose to annotate all apples present in any one picture. This new dataset was carefully annotated using Office PowerPoint (Microsoft® 2016) where a rectangular form was added for each instance. *A posteriori*, the coordinates of each instance were extracted using a python script into text files with the YOLO format annotation.

As the main goal of ASPEN is to work in real time in an embedded environment (*i.e.* Jetson Xavier NX) without the requirement of high volumes of data transference to other computers, YOLOv8 was selected for object detection, due its high inference speed and detection rate. YOLOv8 is a family of object detection architectures and models pretrained in the COCO dataset capable of running in real time, *i.e.* capable of processing more than 30 frames per second (FPS) at a resolution of 1280 × 1280 px [17]. As YOLOv8 contains several models of different complexities, all of these were trained in our dataset at two different resolutions (512 and 1024 px), with the aim of selecting the more efficient model and optimize the available computer resources. We trained all models using 300 epochs with a batch size of 5 images, using transfer learning from a pre-trained model based in the COCO dataset. The main difference between the different models (n, s, m, l and x) is the amount of feature extraction modules and convolution kernel position within the networks [17]. This creates a series of models with different complexities ranging from 8 to 268 giga float point operations per seconds (GFLOPs). Furthermore, to fully evaluate the effect of the addition of our new dataset, the model that performed better was also trained using only the Minneapple adapted dataset. The main components of the YOLOv8 are its backbone network, neck network and detect network. The backbone convolutional network takes care of aggregating detailed images and form features, meanwhile the neck network mainly generates feature pyramid networks (FPN) which are transmitted to the detect network. The detect network is then used in the application of anchor boxes and its associated probabilities, which given the right conditions the user sees as a “detection box”. For a detailed description of the network, the reader is referred to Terven and Cordova-Esparza [43] for a review of the algorithm.

To determine the size of the detected objects, these are first detected in a central portion of an image of resolution 1040 × 1040 px from the available RGBD frame with resolution 1720 × 1080 px. This is done to optimize the speed of the object detection algorithm and maximize image resolution, avoiding to resizing the full image frame. After each apple detection, these are tracked along the image frames using an object tracking algorithm (MOT): SORT [44]. Once a tracked instance passes a region of interest (ROI), which was determined by the closest point of the camera frame to the vegetation and the direction of the movement allowing for the longest possible tracking period, the real size of the instance is measured using the available closest in time depth frame from the RGBD sensor in a parallel process. For this, the boundary box of each detected object is used to crop the RGBD frame and calculate fruits diameters. The selection of the MOT algorithm and region of interest (ROI) size were selected base in previous experiments [40].

Data collection for 3D reconstruction and apple size determination

Four consecutive lines of four different apple varieties planted in 2008 with a density of 1923 trees per ha (4 m x 1.3 m) comprising different colorimetrics (Braeburn, Diwa, Golden and Fuji) were scanned with ASPEN in an orchard located at the research site of Agroscope, Conthey (Switzerland). Each line containing at least 8 trees of the same variety. From each variety, 3 trees were selected per row, and the diameter of 5 apples per tree was measured along the growing season weekly. To evaluate the capacity of ASPEN, these trees were scanned three times within a period of five weeks during July - September 2022, where the fruits not only change in size, but also in colour along the season. Following the recommendations of several studies, the orchard was scanned at approximately solar noon to maximize the quality of the RGB channels, and to avoid the negative influence of shadows in the different measurements (*i.e.* [45]). The rows were scanned on both sides by walking parallel to them at an approximate speed of 4 km h^{-1}. Special care was taken to visualize the complete tree in each image frame, keeping an approximate distance of at least 2.5 m from the trees and in an angle of 45° with respect to the tree line what allows bigger depth in the RGB frame and LiDAR sensor what facilitate the convergence of the SLAM algorithm and a longer tracking period of each fruit (see section above).

Data analysis

To analyse and replicate our results in different platforms, multiple consecutive robot operating system (ROS) bag files of 1 min were recorded during each measuring date using ASPEN. Due to the lack of a 3D segmentation algorithm capable of segmenting each individual tree within the full point cloud (beyond the scope of this paper), these files where then reviewed for time extraction of the approximate boundary limits between trees in a desktop computer (Lenovo ThinkPad P15, Intel core i9, GPU NVIDIA Quadro RTX 5000 Max-Q, 16VGb). Thanks to the timing and localization of the detections, and the manually boundary determinate, it was then possible to determine distribution of apple diameter measurements per tree.

Manual apple diameters time series were filtered out to account for out layers*: e.g.* apples that reduced in size with time. Once outliers were detected, the following measurements were then also removed from the time series. Time series were complemented using a linear model with a quadratic equation fitted using normalized data with respect to the first diameter measurement of all complete time series per variety (data non shown).

To compare the results between manual and digitally measured diameters distributions, a two-sample Kolmogorov-Smirnov test was performed within date and variety. Due to the significantly different number of samples, the digitally measured diameters were subsampled using a random sampler. This was only accepted after confirming no statistical difference between the sub sample population and the original population with the same previously mentioned test. To account for potential variability due to random sampling, the analysis was repeated 100 times to obtain a more robust estimate of statistical significance. P values where then corrected using Bonferroni correction for significance (P value $< 0.05/100$). Finally, to measure the average error of the ASPEN, a linear correlation was fitted between the average values measured digitally and manually.

All post statistical analysis were done using python 3.8 [46] and the package statsmodels (version 0.13.5, [47]).

Results

The addition of our 623 new images to the Minneapple dataset resulted in a total of 1623 images, an increase of 62 % over the original dataset for *in situ* apple object detection. The addition of 55,921 new instances more than doubles the previous number of instances. The distribution of the annotations from the Minneapple dataset, and the added images is present in Fig. 1A and B. In our case smaller annotations were used mainly due significantly bigger tress which increased the distance between the camera and the tree (Fig. 1C and D). A randomly selected image from each dataset can be seen in Fig. 1E and F, aiming to highlight the differences between datasets: Minneapple dataset does not consider fruits in the floor, meanwhile Agroscope_apple have every instance annotated. Similarly, deeper shadow levels are appreciated in F than compared with E, where we aimed to increase the diversity of the dataset.

When using YOLOv8 as an object detector within the merged dataset, there was a significant effect of the image resolutions ($p < 0.05$) used in the mean average precision at interception over unit of 0.5 (mAP@0.5). The complexity and size of the model plays an important role only at resolution 512 px (Fig. 2A). When considering the m model at resolution 1024px a mAP@0.5–0.95, which is calculated at multiple interception over unions from 0.5 to 0.95 in steps of 0.05, a value of 0.37 was obtained, and as demonstrated by Fig. 2B the bigger effect of the increase of resolution was an augment in the recall, indicating a reduction in the total number of false negatives. However, the addition of our new dataset has a negative effect when compared with only training YOLOv8-m in the Minneapple dataset reducing mAP@0.5 from 0.9 to 0.825 and mAP@0.5–0.95 from 0.4 to 0.37 in the trained data (Data not shown).

The average size manually measured along the growing season per variety can be observed in Fig. 3. Although differences were detected along the varieties, similar dynamics are observed where these were better described by a linear model with a quadratic function ($R^2 = 0.92$). On average, Fuji produce bigger apples meanwhile Diwa is the variety with the smallest apples at the end of the season, with an average difference of up to 9.9 mm between varieties.

When observing the apples size distribution for the 3 specific dates when the orchard was also digitally scanned (Fig. 4), it is possible to observe the apple size population dynamics shifting to bigger diameters with time, independent of the used methodology. When comparing the population distributions using the Kolmogorov-Smirnov test (Table 1), only in the latest date it was not possible to reject the null hypothesis of difference between populations for all varieties, indicating that independent of the method nor the variety, the distribution of the measurements did not differ at the end of the experiment.

The average of the diameter measurements for each variety and date correlate with a R^2 of 0.83 from the digital method when compared to the manual measurements (Fig. 5). From Fig. 5 it is possible to observe that the digital method tends to on average, underestimate the average real size by 4.4 mm.

Discussion

The results of the different YOLOv8 models tested are in line with results of previous authors, where object detection produces worse results at lower resolutions and this is mainly due to the small objects (*e.g.* fruits) having a tendency to disappear during the data preparation stage for the object detection algorithm due reductions in image resolution. When evaluating their apple dataset for object detection, Häni et al., [14], reached a mAP@0.5–0.95, of 0.438 when using faster RCNN, compared with 0.433 and 0.341 when using Mask RCNN and Tilled FRCNN. When considering mAP@0.5, a higher value of 0.775 was reached when using faster RCNN. Despite YOLOv8-m performing better than the previously tested algorithms by Häni et al., [14], the negative effect in mAP@0.5 due to the addition of our dataset could be explained by enrichment of the dataset by smaller objects (Fig. 2). Häni *et al.*, (2022) confirmed the main issue in their dataset was the small objects category with an mAP@0.5 of 0.297 compared with 0.871 for large objects. This suggest that even with a larger dataset YOLOv8 does not perform as well with small objects. Additional studies have obtained higher mAP@0.5 when using YOLO for fruits (*i.e.* [18,24,48,49]), nevertheless, its comparison with the current study is not possible due to either usage of close-up images or image tilling to improve object detections (*e.g.* see images in [50]). However, in our case thanks to using YOLOv8-m, we achieve an inference time of *circa* 1.8 ms when using a tensorRT FP16 model in a desktop computer. When considering the full image processing pipeline and visualization, this transfer to *circa* 130 FPS.

Previous authors have demonstrated even higher speeds when using previous versions of the same algorithm (*i.e.* [18,50]). Our findings in context with this literature indicate that processing timing is no longer the bottle neck in real time data analysis when video capture is undertaken at walking speed. For higher speeds, commercial platforms exist (*i.g.* [51,52]) yet in contrast to ASPEN, these heavily rely on GPS sensors for object localization. While new versions of the YOLO algorithm have potential for further improvement in our results (*i.e.* [53]), an exploration of their applicability to this specific problem is beyond the scope of this paper.

The fruit growth reported in Fig. 3 follows the expected growth of apples where growth is predominantly controlled via cell expansion properties [54] and in the current case crop load is specifically influenced by soil water potential (data non shown). At the end of the experiment all varieties had reached their growing potential what is visible in Fig. 3, and moreover, when observing the apple size population dynamics (Fig. 4), the ASPEN pipeline demonstrates the ability to track the apple growth and correlate positively with manual measurements (Fig. 5), even though the population distributions differ, particularly during the second sampling date (table 1). In a post-processing stage, Gené-Mola et al. [55] correlated image-based measurements of apple diameters transformed to 3D point clouds that correlated up to 0.91 with their respective hand measurements once heavily occlude apples were removed from the analysis. When considering all apples independent of their visibility, similar to our study, they reach a correlation of 0.57, lower than the presented pipeline. It is important to highlight the previously mentioned authors calculate errors per apple, with a mean absolute error of 3.7 mm in their best case, meanwhile in our study the mean error of 4.4 mm correspond to the average error as the measurements were not specifically compared per apple but rather per population distribution. Similar to our case, their methodology tends to overestimate apples diameters, and this could be partially explained by variations within fruit, the methodology used to calculate fruit diameter, and errors related to the usage technologies. In our case, the D415 camera has an approximate depth accuracy of 2 % at 2 m and this fits the average observed error (4 vs 4.4 mm).

Examining the population measurements specifically during the first and second dates, differences were found between the methods used (table 1). A conceivable partial explanation for this finding may lie in the selection of bigger apples for the manual measurements at the begging of the experiment, meanwhile the digital tool did not have this bias. Independent of this, for the second and third date apples partially occluded by other objects, *e.g.* leaves and other fruits, may have contributed to an increase in erroneous measurements in the lower tail of the distribution of the digital measurements (Fig. 4). Interestingly for the third date, this potential bias was not enough to have an impact in any of the different Kolmogorov-Smirnov tests. Still, this suggests that to improve the accuracy of ASPEN, new steps could be implemented.

Although less precise than previous works, thanks to the use of a RGBD camera and LiDAR scanner, the ASPEN pipeline can run in real time and here it demonstrates its capacities not only for object detection and localization, but also for object characterization. Moreover, thanks to the parallelization of the pipeline, object characterization does not increase processing times as size determination could be run faster than object detection. From a software perspective, an amodal segmentation mask; basically an estimator of the non-visible part of an object (*i.e.* [56]), may help to remove measurement errors due occlusion problems, one of the main problems related to size estimation as demonstrated by Gené-Mola et al. [55]. An amodal segmentation mask has already been

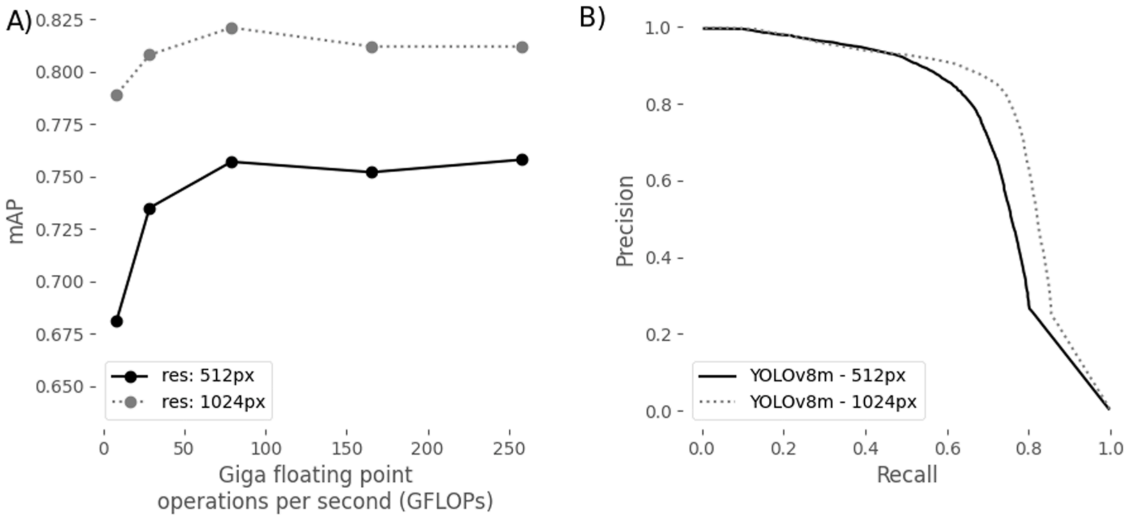


Fig. 2. Training results of YOLOv8 algorithm family. In A) the effect of the models' architecture in the mAP and B) precision recall curve of the medium size models, where in both figures two different resolutions were used (512 and 1024 px).

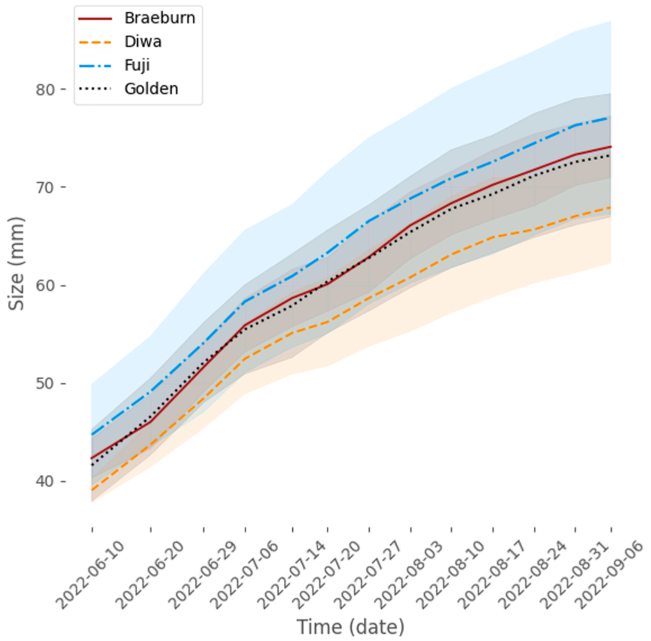


Fig. 3. Average size of 15 fruits for four different apple varieties (Braeburn, Diwa, Fuji and Golden) measured along the growing 2022 season at the Agroscope research site in Conthey, Switzerland.

demonstrated to positively effect size measurements of broccolis and apples ([57] and [58] respectively) this is due to reductions in error measurements after removing measurements from highly occluded fruits, and in the case of apples we are able to increase the correlation to real measurements from 0.8 to 0.91 compared with manually measured apples [58]. While modifying tree architecture to resemble narrow canopies and simpler planar crowns has shown promise for object detection tasks (*e.g.*, [59]), widespread implementation across all apple orchards may not be currently feasible.

For better yield estimations (data non-shown), it would been beneficial to utilize a RGBD camera with a highly dynamic range, as this could assist in improving object detections in shadow areas. Our previous research [40] showed this was not critical as a result of the presence of indirect light. Different orchards tend to be dominated by direct light and this creates difficulties in the acquisition of high contrasting images. To solve this problem previous authors, have chosen to capture images in the presence of artificial light, or work under controlled levels of light which require additional equipment (*i.e.* [45,51,52,60]).

Although our pipeline functions in real time, to fully automate the process two additional software improvements would be required: first a reliable SLAM algorithm, and second, a tree segmentation algorithm. The selected SLAM must work in non-structural environments, and this is not always the case for research when using FAST-LIVO ([41]; data non-shown). As current SLAM algorithms are mainly optimized to work in structured environments, we select FAST-LIVO as the best alternative to use in parallel with visual information for odometry. Regardless, the lack of clear features and repetitive visual patterns make it difficult for the algorithm to converge, especially under highly contrasting light conditions. A critical limitation in the field of SLAM is the scarcity of publicly available datasets that accurately represent non-structured environments coming from mobile terrestrial laser scanners, especially in the field of agriculture.

Equally important, a tree segmentation algorithm able to segment each tree within the point cloud would be beneficial. While many algorithms exist for tree segmentation based on LiDAR data, the most important are tailored for forestry usage (*i.e.* [61]). While in agricultural usage, Wellington et al., [62] demonstrates the capacity of ground surface and individual tree segmentation from a 2D LiDAR sensor when using a Markov random field (MRF) and a hidden semi-Markov model respectively, in a citrus orchard in an offline per row analysis. *A posteriori*, following the same methodology that relies on previous knowledge of the space between trees, Underwood et al., [63] reached a matching performance of 86.8 % with respect to a manually annotated dataset in the tree segmentation task in a parkour equivalent to 26 linear km in a commercial orchard. Westling et al., [64] inspired by the work of Vicari et al., [65] could go one step further and segment trees with crossing canopies without previous knowledge of the orchard required previously, taking on average 49 s per tree. Although the previous methods were efficient, these are functional in complex scenarios with undercrossing canopies and in real time. Even though 3D neural network segmentation has been improved in recent years, especially after the introduction of point net [36], it is not until recently that tailored algorithms have been developed specifically for orchards [66] or even leaves [67]. The capacities and speed of these are still to be evaluated.

Conclusion

In this paper, we validate an open-source pipeline that allows for real

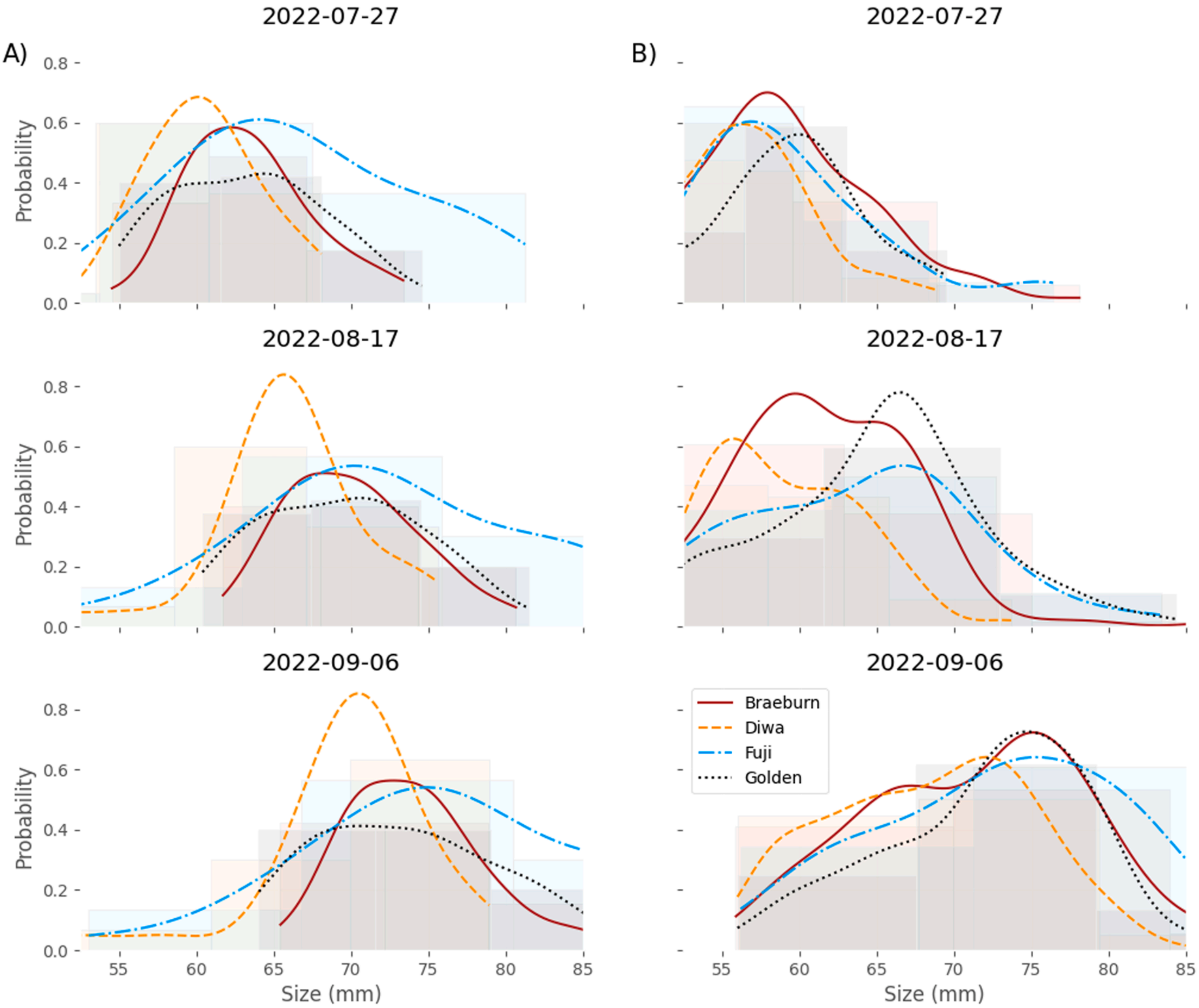
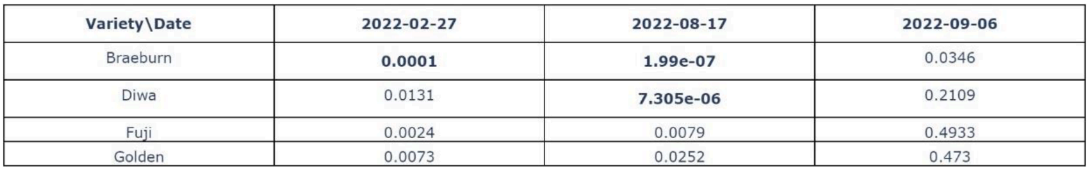


Fig. 4. Diameter measurement distributions for 4 different apple varieties scanned at 3 different dates using either A) manual measurements or B) digital measurements. For the manual measurements, 15 apples per variety were used meanwhile for the digital measurements a variable value was used from 29 to 300 apples.

Table 1
p values of the Kolmogorov-Smirnov test between manual and digital measurements per each date and variety, corrected by Bonferroni (p value threshold = 0.05/100).



Variety\Date	2022-02-27	2022-08-17	2022-09-06
Braeburn	**0.0001**	**1.99e-07**	0.0346
Diwa	0.0131	**7.305e-06**	0.2109
Fuji	0.0024	0.0079	0.4933
Golden	0.0073	0.0252	0.473

time *in situ* apple characterization. We demonstrate that ASPEN, detects, tracks and characterizes *in situ* apples independent of colour and/or variety of apples scanned. When using a RGBD camera, a correlation of 0.83 between average automated measurements and average manual measurements is achieved with a mean error of 4.4 mm.

Although in real time opoerations, improvements could be made with respect to hardware and software, leaving room for further research. To reduce the gap between computer sciences and agricultural researchers, we open source a new dataset of more than 600 images used in this research[2] and we hope this can be used as a benchmark together with previous datasets to evaluate future algorithms in detecting apples in full tree images. Similarly, we make the weights of our trained models available to accelerate the use of image-based technologies in agriculture.

[2] https://zenodo.org/doi/10.5281/zenodo.12623539

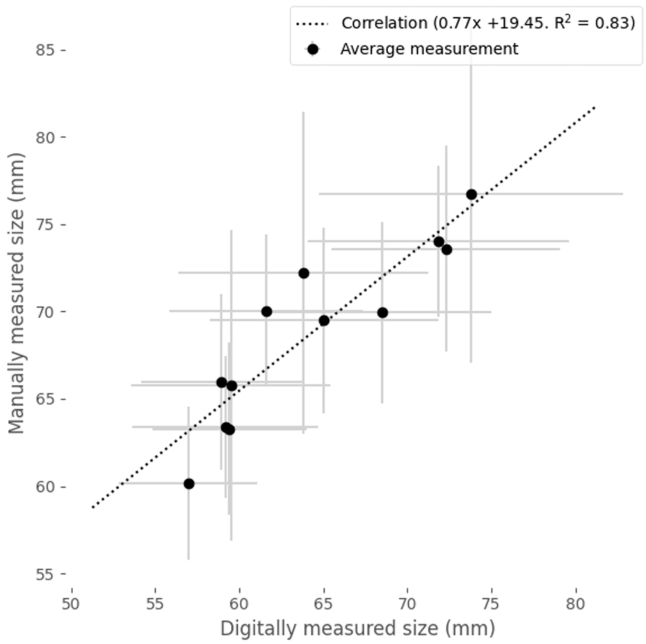


Fig. 5. Average diameter measurement distribution along 4 different apple varieties scanned at 3 different dates ($n = 12$), where the X axis correspond to the average digitally measured size, and Y axis correspond to the average manual measurements. For the manual measurements, 15 apples per variety were used meanwhile for the digital measurements a variable value was used from 29 to 300 apples. Grey lines correspond to either X or Y standard deviation.

Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work the author(s) used DeepL Write to facilitate the writing of the manuscript. After using this tool/service, the author(s) reviewed and edited the content as needed together with a native English speaker. The author(s) take(s) full responsibility for the content of the publication.

CRediT authorship contribution statement

Camilo Chiang: Writing – review & editing, Writing – original draft, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Alice Monney:** Data curation. **Phillipe Monney:** Investigation, Conceptualization. **Danilo Christen:** Writing – review & editing, Supervision, Project administration.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data Availability

Data will be made available on request.

References

[1] United Nations, Department of economic and social affairs, population division. 2022. World Population Prospects 2022: Data Sources. (UN DESA/POP/2022/DC/NO. 9).

[2] Raymond Pierrehumbert, There is no plan B for dealing with the climate crisis, Bullet. Atomic Scient. 75 (5) (2019), https://doi.org/10.1080/00963402.2019.1654255.

[3] A. Steiner, G. Aguilar, K. Bomba, J.P. Bonilla, A. Campbell, R. Echeverria, R. Gandhi, C. Hedegaard, D. Holdorf, N. Ishii, K. Quinn, B. Ruter, I. Sunga, P. Sukhdev, S. Verghese, J. Voegele, P. Winters, B. Campbell, D. Dinesh, S. Huyer, A. Jarvis, A.M. Loboguerrero Rodriguez, A. Millan, P. Thornton, L. Wollenberg, S Zebiak, Actions to transform food systems under climate change, CGIAR research program on climate change, agriculture and food security (CCAFS), Wageningen, The Netherlands, 2020.

[4] M. Kranzberg, Technology and history: "Kranzberg's laws", Technol. Cult. (1986) https://doi.org/10.2307/3105385.

[5] R. Pieruschka, U. Schurr, Plant phenotyping: past, present and future, Plant Phenomics. (2019), https://doi.org/10.34133/2019/7507131.

[6] C. O`Neil, T. Nicklas, V. Fulgoni, Consumption of apples is associated with a better diet quality and reduced risk of obesity in children: national Health and Nutrition Examination Survey (NHANES) 2003–2010, Nutr. J. (2015), https://doi.org/10.1186/s12937-015-0040-1.

[7] A.D. Whittaker, G.E. Miles, O.R. Mitchell, Fruit location in a partially occluded image, Trans. ASABe 30 (1987) 591–596, https://doi.org/10.13031/2013.30444.

[8] R.O. Duda, P.E. Hart, Use of the Hough transformation to detect lines and curves in pictures, Commun ACM (1972), https://doi.org/10.1145/361237.361242.

[9] D.M. Bulanon, T.A. Kataoka, Fruit detection system and an end effector for robotic harvesting of Fuji apples, Agricult. Eng. Internat. 12 (2010) 203–210.

[10] Y.S. Zhao, L. Gong, Y.X. Huang, C.L. Liu, A review of key techniques of vision-based control for harvesting robot, Comput. Electron. Agric. (2016), https://doi.org/10.1016/j.compag.2016.06.022.

[11] J. Zhao, J. Tow, J. Katupitiya, On-tree fruit recognition using texture properties and color data, in: Proceedings of the IEEE/RSJ international conference on intelligent robots and systems, 2005, https://doi.org/10.1109/IROS.2005.1545592.

[12] X. Liu, S. Chen, S. Adtiya, N. Sivakumar, S. Dcunha, C. Qu, C. Taylor, J. Das, V. Kumar, Robust fruit counting: combining deep learning, tracking and structure from motion, in: proceedings of the IEEE/RSJ International conference on intelligent robots and systems, 2018, https://doi.org/10.1109/IROS.2018.8594239.

[13] J. Gené-Mola, R. Sanz-Cortiellaa, J. Rosell-Poloa, J. Morros, J. Ruiz-Hidalgo, V. Vilaplana, E. Gregoria, Fruit detection and 3D location using instance segmentation neural networks and structure-from-motion photogrammetry, Comput. Electron. Agric. (2020), https://doi.org/10.1016/j.compag.2019.105165.

[14] Häni, N., Roy, P. and Isler, V. 2020. Minneapple: a benchmark dataset for apple detection and segmentation. Preprint. 10.48550/arXiv.1909.06441.

[15] Redmon J., Divvala S., Girshick R. and Farhadi A. 2015. You only look once: unified, real-time object detection. Preprint. 10.48550/arXiv.1506.02640.

[16] Bochkovskiy, A., Wang, C. and Liao, H. 2020. YOLOv4: optimal speed and accuracy of object detection. Preprint. 10.48550/arXiv.2004.10934.

[17] Jocher, G., Chaurasia, A. and Qiu, J. YOLO by ultralytics. 2023. https://github.com/ultralytics/ultralytics accessed 21st August 2023.

[18] B. Yan, P. Fan, X. Lei, Z. Liu, F. Yang, A real-time apple targets detection method for picking robot based on improved YOLOv5, Remote Sens. (2021), https://doi.org/10.3390/rs13091619.

[19] L. Fu, Z. Yang, F. Wu, X. Zou, J. Lin, Y. Cao, J. Duan, YOLO-banana: a lightweight neural network for rapid detection of banana bunches and stalks in the natural environment, Agronomy (2022), https://doi.org/10.3390/agronomy12020391.

[20] R. Gai, N. Chen, H. Yuan, A detection algorithm for cherry fruits based on the improved YOLO-v4 model, Neural Comput. Applicat. (2021), https://doi.org/10.1007/s00521-021-06029-z.

[21] G. Liu, J. Nouaze, P. Touko, J. Kim, YOLO-tomato: a robust algorithm for tomato detection based on YOLOv3, Sensors (2020), https://doi.org/10.3390/s20072145.

[22] G. Li, X. Huang, J. Ai, Z. Yi, W. Xie, Lemon-YOLO: an efficient object detection method for lemons in the natural environment, IET. Image Process. (2021), https://doi.org/10.1049/ipr2.12171.

[23] A. Koirala, K. Walsh, Z. Wang, C. McCarthy, Deep learning for real-time fruit detection and orchard fruit load estimation: benchmarking of 'MangoYOLO, Precis. Agric. (2019), https://doi.org/10.1007/s11119-019-09642-0.

[24] A. Kuznetsova, T. Maleva, V. Soloviev, YOLOv5 *versus* YOLOv3 for apple detection, in: A.G. Kravets, A.A. Bolshakov, M. Shcherbakov (Eds.), Cyber-Physical Systems: Modelling and Intelligent Control. Studies in Systems, Decision and Control, 2021, https://doi.org/10.1007/978-3-030-66077-2_28.

[25] M. Hilbert, P. López, The World's technological capacity to store, communicate, and compute information, Science (1979) (2011), https://doi.org/10.1126/science.1200970.

[26] C. Sun, A. Shrivastava, S. Singth, A. Gupta, Revisiting unreasonable effectiveness of data in deep learning era, in: IEEE International conference on computer vision (ICCV), 2017, https://doi.org/10.1109/ICCV.2017.97.

[27] S. Bargoti, J. Underwood, Deep fruit detection on orchards, in: 2017 IEEE International conference on robotics and automation (ICRA), 2016, https://doi.org/10.1109/ICRA.2017.7989417.

[28] J. Gené-Mola, R. Sanz-Cortiella, J. Rosell-Polo, J. Morros, J. Ruiz-Hidalgo, V. Vilaplana, E. Gregorio, Fuji-sfm dataset: a collection of annotated images and point clouds for fuji apple detection and location using structure-from-motion photogrammetry, Data Brief. (2020), https://doi.org/10.1016/j.dib.2020.105591.

[29] J. Gené-Mola, E. Gregorio, F. Aut, J. Guevara, J. Llorens, R. Sanz-Cortiella, A. Escolà, J. Rosell-Polo, Fruit detection, yield prediction and canopy geometric characterization using LIDAR with forced air flow, Comput. Electron. Agric. (2020), https://doi.org/10.1016/j.compag.2019.105121.

[30] P. Roy, A. Kislay, P. Plonski, J. Luby, V. Isler, Vision-based preharvest yield mapping for apple orchards, Comput. Electron. Agric. (2019), https://doi.org/10.1016/j.compag.2019.104897.

[31] N. Häni, P. Roy, V. Isler, A comparative study of fruit detection and counting methods for yield mapping in apple orchards, J. Field. Robot. (2019), https://doi.org/10.1002/rob.21902.

[32] U. Weiss, P Biber, Plant detection and mapping for agricultural robots using a 3D LiDAR sensor, Rob. Auton. Syst. 59 (2011), https://doi.org/10.1016/j.robot.2011.02.011.

[33] J. Gené-Mola, E. Gregorio, F. Aut, Guevara, R. Sanz-Cortiella, A. Escolà, J. Llorens, J. Morros, J. Ruiz-Hidalgo, V. Vilaplana, J Rosell-Polo, Fruit detection in an apple orchard using a mobile terrestrial laser scanner, Biosyst. Eng. (2019), https://doi.org/10.1016/j.biosystemseng.2019.08.017.

[34] J. Gene-Mola, E. Gregorio, F. Auat, J. Guevara, J. Llorens, R. Sanz-Cortiella, A. Escolà, J. Rosell-Polo, LFuji-air dataset: annotated 3D LiDAR point clouds of Fuji apple trees for fruit detection scanned under different forced air flow conditions, Data Brief. (2020), https://doi.org/10.1016/j.dib.2020.105248.

[35] Qi. C., Su, H., Mo, K. and Guibas, L. 2017. Pointnet: deep learning on point sets for 3D classification and segmentation. Preprint. 10.48550/arXiv.1612.00593.

[36] C. Qi, H. Su, M. Niessner, A. Dai, M. Yan, L. Guibas, Volumetric and multi-view cnns for objects classification on 3D data, in: IEEE /CVF computer vision and pattern recognition conference, 2016, https://doi.org/10.48550/arXiv.1604.03265.

[37] S. Hoque, M.D. Arafat, S. Xu, A. Maiti, Y. Wei, A comprehensive review on 3d object detection and 6d pose estimation with deep learning, IEEe Access. (2021), https://doi.org/10.1109/ACCESS.2021.311439.

[38] M. Stein, S. Bargoti, J. Underwood, Image based mango fruit detection, localization and yield estimation using multiple view geometry, Sensors (2016), https://doi.org/10.3390/s16111915.

[39] J. Underwood, C. Hung, B. Whelan, S. Sukkarieh, Mapping almond orchard canopy volume, flowers, fruit and yield using LiDAR and vision sensors, Comput. Electron. Agric. (2016), https://doi.org/10.1016/j.compag.2016.09.014.

[40] C. Chiang, D. Tran, C. Camps, ASPEN study case: real time *in situ* tomato detection and localization for yield estimation, Agricult. Res. Techn. Open Access J. (2024), https://doi.org/10.19080/ARTOAJ.2024.28.556406.

[41] Zheng, C., Zhu, Q., Xu, W., Guo, Q. and Zhang, F. 2022. FAST-LIVO: fast and tightly-coupled sparse-direct LiDAR-inertial-visual odometry. Preprint. 10.48550/arXiv.2203.00893.

[42] Li, J. and Zhang, F. 2021. R3LIVE: a Robust, Real-time, RGB-colored, LiDAR-Inertial-Visual tightly-coupled state Estimation and mapping package. Preprint. 10.48550/arXiv.2109.07982.

[43] Terven, J. and Cordova-Esparza, D. 2023. A Comprehensive Review of YOLO: from YOLOv1 and beyond. MACHINE LEARNING AND KNOWLEDGE EXTRAction. 10.3390/make5040083.

[44] A. Bewley, Z. Ge, L. Ott, F. Ramos, B Upcroft, Simple online and real time tracking, in: 2016 IEEE International conference on image processing (ICIP), 2016, https://doi.org/10.1109/ICIP.2016.7533003.

[45] J. Gené-Mola, J. Llorens, Rosell-Polo, E. Gregorio, J. Arno, F. Solanelles, J. Martinez-Casasnuevas, A Escolà, Asseing the performance of RGB-D sensors for 3D fruit crop canopy characterization under different operating and lighting conditions, Sensors (2020), https://doi.org/10.3390/s20247072.

[46] G. Van Rossum, F. Drake, Python3 Reference manual, Ca. CreateSpace, Scotts Valley, 2009.

[47] Seabold., S. and Perktold, J. 2010. Statsmodels: econometrics and statistical modeling with python. 9th Python in science conference.

[48] A. Kuznetsova, T. Maleva, V. Soloviev, Using YOLOv3 algorithm with pre- and post-processing for apple detection in fruit-harvesting robot, Agronomy (2020), https://doi.org/10.3390/agronomy10071016.

[49] Y. Egi, M. Hajyzadeh, E. Eyceyurt, Drone-computer communication based tomato generative organ counting model using YOLOv5 and deep-sort, Agriculture (2022), https://doi.org/10.3390/agriculture12091290.

[50] A. Borja Parico, T. Ahamed, Real time pear fruit detection and counting using YOLOv4 models and deep SORT, Sensors (2021), https://doi.org/10.3390/s21144803.

[51] A. Scalisi, L. McClymont, J. Underwood, P. Morton, S. Scheding, I. Goodwin, Reliability of a commercial platform for estimating flower cluster and fruit number, yield, tree geometry and light interception in apple trees under different rootstocks and row orientations, Comput. Electron. Agric. (2021), https://doi.org/10.1016/j.compag.2021.106519.

[52] A. Scalisi, L. Mcclymont, M. Peavey, P. Morton, S. Scheding, J. Underwood, I. Goodwin, Detecting, mapping and digitising canopy geometry, fruit number and peel colour in pear trees with different architecture, Sci. Hortic. (2024), https://doi.org/10.1016/j.scienta.2023.112737.

[53] Wang, C., Yeh, I. and Mark, H. 2024. YOLOv9: learning what you want to learn using programmable gradient information. Preprint. https://arxiv.org/abs/2402.13616.

[54] K. Jahed, P. Hirst, Fruit growth and development in apple: a molecular, genomics and epigenetics perspective, Front. Plant Sci. (2023), https://doi.org/10.3389/fpls.2023.1122397.

[55] J. Gené-Mola, R. Sanz-Cortiella, J. Rosell-Polo, A. Escolà, E. Gregorio, In-field apple size estimation using photogrammetry-derived 3D point clouds: comparison of 4 different methods considering fruit occlusions, Comput. Electron. Agric. (2021), https://doi.org/10.1016/j.compag.2021.106343.

[56] Follmann, P., König, R., Hàrtinger, P. and Klostermann, M. 2018. Learning to see the invisible: end-to-end trainable amodal instance segmentation. Preprint. 10.48550/arXiv.1804.08864.

[57] M. Blok, E. van Henten, F. van Evert, G. Kootstra, Image-based size estimation of broccoli heads under varying degrees of occlusion, Biosyst. Eng. (2021), https://doi.org/10.1016/j.biosystemseng.2021.06.001.

[58] J. Gene-Mola, M. Ferrer-Ferrer, E. Gregorio, P. Blok, J. Hemming, J. Morros, J. Rossel-Polo, V. Vilaplana, J. Ruiz-Hidalgo, Looking behind occlusions: a study on amodal segmentation for robust on-tree apple fruit size estimation, Comput. Electron. Agric. (2023), https://doi.org/10.1016/j.compag.2023.107854.

[59] G. Bortolotti, K. Bresilla, M. Piano, L. Grappadelli, L. Manfrini, 2D tree crops training system improve computer vision application in field: a case study, IEEE Internat. Workshop Metrol. Agricult. Forest. (MetroAgriFor) (2021), https://doi.org/10.1109/MetroAgriFor52389.2021.9628839.

[60] F. Esser, R. Rosu, A. Cornelissen, L. Klingbeil, H. Kuhlmann, S. Behnke, Field robot for high-throughput and high-resolution 3D plant phenotyping: towards efficient and sustainable crop production, IEEE Robot. Autom. Magaz. (2023), https://doi.org/10.1109/MRA.2023.3321402.

[61] M. Wielgosz, S. Puliti, P. Wilkes, R. Astrup, Point2tree (p2t) – framework for parameter tunning of semantic and instance segmentation used with mobile laser scanning data in coniferous forest, Remote Sens. 15 (2023), https://doi.org/10.3390/rs15153737.

[62] Wellington, C., Campoy, J., Khot, L. and Ehsani, R. 2012. Orchard tree modeling for advanced sprayer control and automatic tree inventory [access on 2023.08.31: https://www.cs.cmu.edu/~mbergerm/agrobotics2012/07Wellington.pdf].

[63] J. Underwood, G. Jagbrant, J. Nieto, S. Sukkarieh, Lidar-based tree recognition and platform localization in orchards, J. Field. Robot. (2015), https://doi.org/10.1002/rob.21607.

[64] F. Westling, J. Underwood, M. Bryson, Graph-based methods for analyzing orchard tree structure using noisy point cloud data, Comput. Electr. Eng. (2021), https://doi.org/10.1016/j.compag.2021.106270.

[65] M. Vicari, M. Disney, P. Wilkes, A. Burt, K. Calders, W. Woodgate, Leaf and wood classification framework for terrestrial lidar point clouds, Methods Ecol. Evol. (2019), https://doi.org/10.1111/2041-210X.13144.

[66] X. Bu, C. Liu, H. Liu, G. Yang, Y. Shen, J. Xu, DFSNet: a 3D point cloud segmentation network toward trees detection in an orchard Scene, Sensors (2024), https://doi.org/10.3390/s24072244.

[67] G. Roggiolani, F. Magistri, T. Guadagnino, J. Behley, C. Stachniss, Unsupervised pre-training for 3D leaf instance segmentation, IEEe Robot. Autom. Lett. (2023), https://doi.org/10.1109/LRA.2023.3320018.