Smart Agricultural Technology 11 (2025) 100964

Contents lists available at ScienceDirect

Smart Agricultural Technology

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

Apple phenotyping using deep learning and 3D depth analysis: An experimental study on fruitlet sizing during early development

Giorgio Checola [a,*], Damiano Moser [b], Paolo Sonego [a], Cristian Iob [b], Franco Micheli [b], Pietro Franceschi [a,*]

[a] Unit of Digital Agriculture, Research and Innovation Centre, Fondazione Edmund Mach, Via E. Mach 1, 38098 San Michele all'Adige, Italy
[b] Unit of Fruit Experimental Cultivation, Technological Transfer Centre, Fondazione Edmund Mach, Via E. Mach 1, 38098 San Michele all'Adige, Italy

ARTICLE INFO

Keywords:
Machine vision
Depth camera
Apple phenotyping
Size estimation
Precision agriculture
Fruitlets
Early development

ABSTRACT

Current research in apple-growing focuses on collecting extensive biometric data to better understand physiological processes, improve orchard productivity and predict yields. In this context, fruit thinning has emerged as a key horticultural practice to enhance fruit size and quality while preventing alternate bearing. Despite the growing role of plant imaging technologies in agronomic management, fruitlet sizing remains challenging, particularly in early phenological stages.

To address this challenge, we developed an RGB-D-based vision pipeline that combines YOLO models with depth information and relies on the statistical analysis of frame series to detect and cluster fruitlets into flower corymbs, providing both fruitlet counting and diameter estimates for each video acquisition. After obtaining an AP@0.5 and AP@[0.5:0.95] of respectively 0.894 and 0.77 in fruitlet detection, along with a precision of 0.881 and a recall of 0.846, our approach efficiently processed video frames, extracting the most reliable data for each labeled cluster. While the comparison of true positive estimates with calibrated caliper measurements showed a mean RMSE of 1.05 mm, challenges remain in achieving the correct fruitlet count, with a mean counting error of 0.63 fruitlets per video. Additionally, the proposed workflow retrieved the exact number of fruitlets as the ground truth in 56.4% of the videos, increasing to 75% when excluding those videos where the correct fruitlet count was never detected in any frame by the YOLO model.

Despite these limitations, our results are promising, proposing a potential data acquisition tool without compromising the reliability of traditional practices. This approach could pave the way for future applications, including the evaluation of plant growth regulator trials and the development of predictive models for yield and productivity optimization.

1. Introduction

In apple orchard management, fruitlet thinning is one of the key horticultural practices for achieving high-quality fruit production. This process, which can be performed manually, mechanically or through the application of chemical thinners, ensures an adequate crop load and promotes a proper return to bloom in the following year, thereby preventing alternate bearing [1]. Thinning involves removing excess fruitlets to enhance apple quality in terms of size, color, and sugar content. These fruitlets typically grow in clusters known as flower corymbs, each consisting of a central "king fruit" (KF) surrounded by 4-6 smaller lateral fruits. The growth rate between them defines the hierarchy, a crucial indicator for predicting fruit abscission, i.e., the "physiological drop" of young fruitlets during early development [2].

Monitoring fruit size has become essential for better understanding physiological processes, optimizing thinning practices during early developmental stages, and guiding coherent horticultural decisions through predictive models [3–5]. However, building such models requires extensive data, including fruit dimensions, environmental conditions, and the type and application rates of chemical thinners.

A major challenge in this process is the efficient and scalable collection of accurate fruit biometric data in orchard environments. Traditional manual methods are time consuming, labour-intensive and prone to errors, making large-scale data acquisition impractical. To

* Corresponding authors. +39 0461615571, +39 0461615556.
E-mail addresses: giorgio.checola@fmach.it (G. Checola), damiano.moser@fmach.it (D. Moser), paolo.sonego@fmach.it (P. Sonego), cristian.iob@fmach.it (C. Iob), franco.micheli@fmach.it (F. Micheli), pietro.franceschi@fmach.it (P. Franceschi).

https://doi.org/10.1016/j.atech.2025.100964
Received 3 March 2025; Received in revised form 9 April 2025; Accepted 18 April 2025
Available online 24 April 2025

overcome these limitations, we introduced an automated vision tool for analyzing flower corymbs, providing fruitlet size and count estimates during in-field surveys, even under challenging conditions such as partial occlusion from leaf cover, small fruit size, and variable lighting. While human involvement remains essential, integrating deep learning with advanced RGB-D cameras can significantly improve survey efficiency and accuracy. Depth cameras, in particular, play a crucial role by enabling real-world size estimation of detected objects through distance measurement, offering a faster and more objective mean to characterize the canopy.

Current research in apple detection and sizing has predominantly focused on the later developmental stages (pre-harvest), due to its stronger correlation with apple yield estimation [6–8]. However, recent studies [9] have proposed using machine vision systems to estimate the potential fruiting capacity of the branch, indirectly assessing the importance of thinning during the early growth stages.

The first application on green fruitlets was introduced two years earlier, with a YOLOv5 model for accurate apple fruitlet detection [10]. Recognizing the importance of flower thinning, a recent study presented a YOLO-based approach for flower detection and clustering in orchard environments [11]. Sapkota et al. [12] took an additional step by combining the state-of-the-art YOLOv8 segmentation model with 3D point cloud data to estimate fruitlet sizes. Meanwhile, Freeman and Kantor [13] proposed a robotic solution for sizing apple fruitlets through an autonomous viewpoint planner. However, these studies focused primarily on individual phenological stages. In contrast, our work offers a comprehensive analysis across the growing season, specifically between BBCH 72 and 74, by comparing traditional caliper measurements with camera-based estimates. A key aspect of this agronomic challenge is to observe how growth differs depending on the type of bud from which each corymb emerges. We demonstrate that a camera-based monitoring system can reliably replicate the results of manual data collection, thereby reducing labor demands and improving the scalability, objectivity, and frequency of field monitoring.

The main contributions of this paper are:

- a deep learning-based workflow for extracting key fruit data, including size and count, in complex orchard environments;
- a fully annotated dataset of apple fruitlets during early development, collected using depth camera recordings, including ground-truth measurements of the selected corymbs;
- an experimental study comparing apple growth across different types of buds, using both traditional caliper measurements and camera-based estimates.

2. Materials and methods

2.1. Data collection

The experimental study began with in-field data acquisition: thirty-five flower corymbs were carefully tagged from a commercial Fuji apple block (Aztec clone), grafted on M9 rootstock and trained using a tall spindle system.

The process involved recording individual videos of the corymbs using a standardized and reproducible approach: an average of 10-seconds video at a distance of around 30 cm, capturing target fruitlets from multiple orientations. We closely replicated the agronomist's procedure in real-world conditions that has to address possible issues such as the partial occlusion of the objects due to increased leaf cover, the small size of fruits, and poor lighting conditions — obstacles that make this task impractical for autonomous ground vehicles. Each video was recorded at a resolution of 640 × 480 pixels for both RGB and depth sensors, with a frame rate of 15 fps. These settings were chosen to balance computing efficiency — critical for data handling and model training — with data quality, ensuring sufficient detail in RGB frames for subsequent analysis, and enabling very close-range recording below the minimum operating distance (see Section 2.1.1 for more details).

Seven monitoring sessions were conducted between April and May, covering the period from fruit set to the stage when fruitlets slightly exceed 40 mm in diameter. Surveys took place on April 24, April 29, May 6, May 10, May 14, May 20, and May 29. During the first inspection, 30 corymbs were recorded, with the number increasing to 35 in the following surveys. In addition to the bud type, metadata such as the orientation of the vegetative wall and the presence of the king fruit was also recorded. Flower buds are borne on shoots or short spurs at terminal position, with some exception in 1-year shoots, which may carry flowers derived from lateral buds [14]. Based on current literature [15], we compared 3 types of flower buds, illustrated in Fig. 1A for clarity:

1. *apical*: located at the terminal position of one-year-old shoots;
2. *lateral*: axillary buds produced in the basal leaf axils on extension (one-year-old) shoots;
3. *spur*: found at the tips of shortened older shoots with a length less than 5 cm.

Table 1 provides a more detailed summary of the ground-truth data, listing the labeled corymbs categorized by their bud type (Apical, Lateral, Spur) and orientation (East, West). Manual caliper measurements of fruitlet diameters, required for model validation, were collected after each field survey.

2.1.1. Hardware settings

For close-range data acquisition, we used the Intel® RealSense™ Depth camera D435i, part of the D400 series of cameras, a lineup that integrates Intel's latest depth-sensing technologies and an inertial measurement unit (IMU) for additional motion tracking. Unlike typical RGB cameras, depth cameras create pixel-wise distance measurements through dual sensors, enabling accurate 3D mapping. The D435i supports a maximum RGB resolution of 1920 × 1080, and a depth output resolution up to 1280 × 720 pixels, with optimal depth results achieved at 848 × 480 due to the wider field of view (FOV) of 87° × 58°. Compared to its D415 counterpart, the D435i has a shorter minimum depth sensing distance of 28 cm at maximum resolution. Lowering the depth resolution to 640 × 480 further reduces this distance to 17.5 cm, as specified in the datasheet [16].

Raw depth data from such devices often contain gaps and missing pixels, which can degrade system performance. Using `pyrealsense2` [17], the python wrapper for Intel RealSense SDK 2.0, we applied pre-processing algorithms, including a spatial filter, which uses 1-D edge-preserving filter to enhance the smoothness of the reconstructed data [18]. This was combined with a heuristic hole-filling mode to correct minor artefacts with minimal performance impact. Other filters were discarded, after verifying the ineffectiveness of interpolation for the processed data.

2.1.2. Dataset

Excluding the corymbs that experienced total fruit abscission, the data acquisition process resulted in 234 video files in ".bag" format, and a total of 1054 fruitlet measurements. The bag files were generated using the Intel® RealSense™ SDK 2.0, which allows for synchronized recording of RGB and depth streams along with metadata.The measures were divided by date and bud type to analyze differences in growth behavior over time.

Fruitlet detection has posed challenges due to the rapid phenological changes fruitlets undergo as they grow, significantly affecting their size, shape, and color. Fig. 1B illustrates how quickly this transition occurs, highlighting the slower growth of lateral buds compared to apical ones. Initially (4 to 10 mm in size), flowers may be larger than the fruitlets, hiding the visual appearance of their bodies. Over time, fruitlets take on an elliptical shape, while flowers gradually wither and disappear. After about a month, the body starts shifting from a violet hue to green, acquiring a more rounded shape. Some fruitlets, however, may undergo

Fig. 1. (A) Positions of bud types (Apical, Lateral, and Spur) along a branch and shoot. (B) Example of fruitlet growth over time by bud type: first column shows corymb #7, second column corymb #5, and third column corymb #3.

Table 1
Classification of ground-truth data in the experimental study.

Bud type	Count	Orientation	Labels
Apical	10	East	1,4,7,10,13
		West	16, 19, 22, 25, 28
Lateral	10	East	2, 5, 8, 11, 14
		West	17, 20, 23, 26, 29
Spur	15	East	3, 6, 9, 12, 15
		West	18, 21, 24, 27, 30, 101, 102, 103, 104, 105

abscission, appearing similar to those in the early developmental stages.

As illustrated in Fig. 2A, the distribution of ground-truth data started revealing a bimodal pattern by the fourth survey. According to the plant phenological development, three weeks after petal fall, one or two fruits began growing more rapidly, while the others showed signs of wilting. This trend was particularly pronounced in apical and spur buds, whereas lateral buds often exhibited higher rates of fruitlet abscission, with fewer developing into mature apples — findings consistent with previous research [19,20]. The contrast between apical and lateral buds became even more evident during the last monitoring session. Corymbs from apical and spur buds included several fruitlets exceeding 30 mm in diameter, while lateral measurements rarely surpassed 15 mm, likely due to the abscission of the main fruitlets, a process influenced by both natural physiological thinning and the application of growth regulators. These observations are confirmed by Fig. 2B: at the beginning of the phenological stage, lateral annual corymbs included fewer fruitlets compared to the other two branches, demonstrating a greater tendency towards wilting and drop. By the end of the acquisition process, they had an average of just two fruitlets, compared to three in annual buds and four in spur buds.

2.2. Fruitlet detection and sizing

The pipeline of the proposed automated vision tool consists of three main blocks: detection, sizing and video frame analysis. Fig. 3 provides

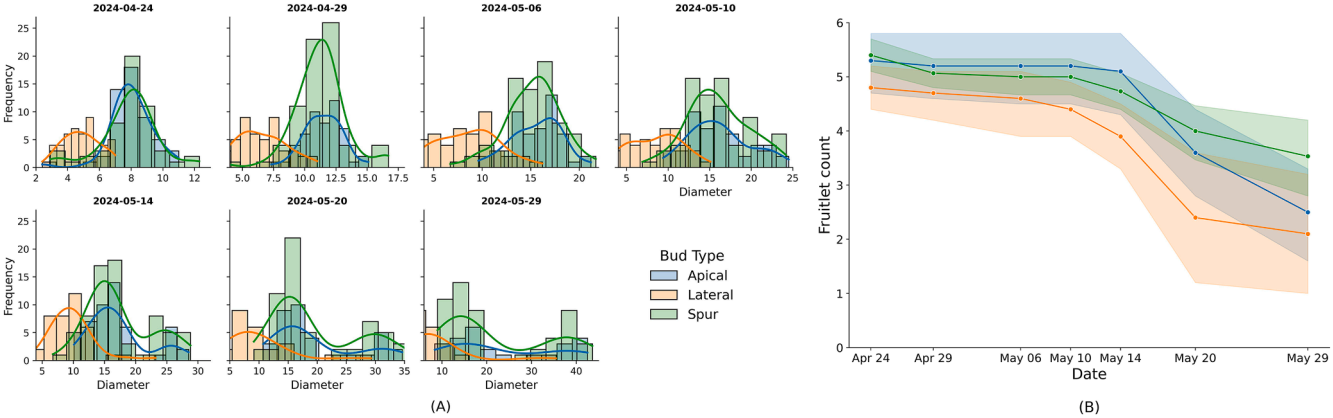

Fig. 2. Analysis of manual measurements. (A) Distribution of fruitlet diameters by bud type across survey dates. (B) Fruitlet drop trend over time by bud type.

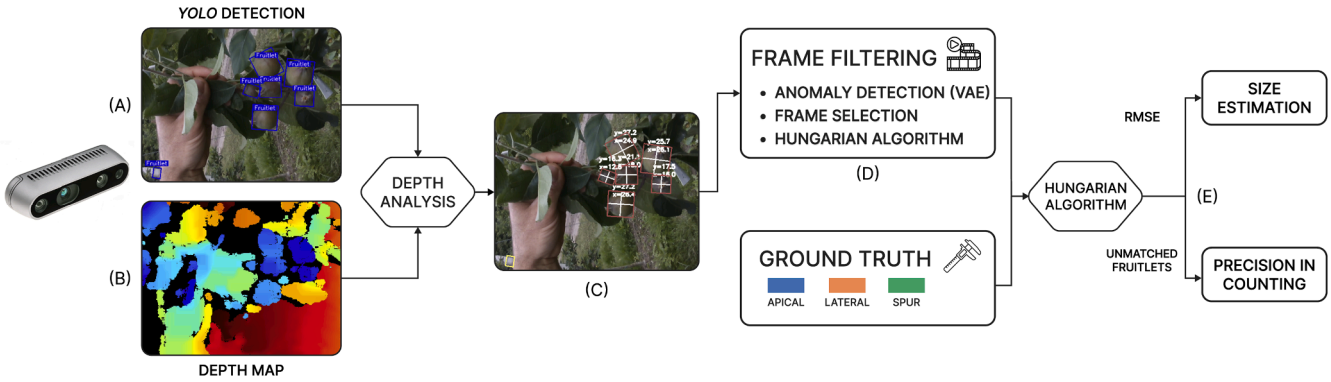

Fig. 3. Structure of the proposed workflow: (A) an RGB frame with detected fruitlets, (B) the corresponding depth map, and (C) the output image with clustering and size estimation performed. (D) The post-processing steps, including frame selection and filtering, and (E) the performance assessment of the estimated data.

an overview of the workflow, detailing the steps of frame filtering and the assessment of validation method. All described procedures can be fully reproduced on the same dataset by following the instructions in the Github repository (see Data availability). The first step involved training a deep learning model, specifically focusing on the bounding box detection method (Fig. 3A).

We started by extracting RGB images and depth frames from the bag files, followed by a spatial stream alignment. This process created a synthetic depth stream, mapping frames to the same RGB viewports (https://dev.intelrealsense.com/docs/rs-align). This alignment was crucial to ensure that each pixel in the two frames corresponded to the same spatial point in the scene, resulting in approximately 16000 RGB frames. After that, we applied stratified random sampling to ensure a balanced representation across videos: 481 images were selected for training, validating and testing the model.

Given the dynamic traits of fruitlet growth and our primary focus on fruitlet size estimation, we defined a single "fruitlet" class to capture variability across stages. We created approximately 4,000 annotations using LabelImg software [21] for axis-aligned bounding boxes, and the modified version of YOLO OBB format (https://github.com/heshamaqi/labelImg_OBB/tree/master) for oriented bounding boxes. This dual annotation approach was tested to evaluate potential differences in post-processing. Frame sequences were then divided into a 60-20-20 ratio to enhance the model's ability to generalize across developmental stages while consistently detecting each object.

2.2.1. YOLO setup and performance

Fruitlet detection was carried out using the single-stage YOLO (You Only Look Once) architecture, first introduced by Redmon et al. in 2015 [22]. This algorithm processes the entire image in a single pass, simultaneously handling both bounding box regression and object classification [23]. Over the years, YOLO has undergone continuous refinements, leading to significant improvements in detection capabilities and computational efficiency.

For our analysis, we evaluated several model versions, including YOLOv8, YOLOv10, and the latest YOLOv11, released in September 2024 by Ultralytics [24]. Key training parameters were fixed across experiments: image size (640), batch size (32), and number of workers (8). Detailed test configurations are provided in Table S.1 of the supplementary material.

Among the evaluated models, the pre-trained "large" version YOLOv11l achieved the highest performance in terms of AP and loss, aligning with recent studies that highlight its superiority and reduced parameter requirements [23,25]. Training was conducted on an AWS g5.2xlarge instance equipped with 8 vCPUs, 32.0 GiB of RAM, and an NVIDIA A10G with 24.0 GiB of VRAM.

2.2.2. Depth analysis

To reconstruct the predicted bounding boxes in 3D and estimate object dimensions — width, and height — we applied inverse camera projection, as described by Eq. 1. This process transforms 2D image coordinates back into 3D world coordinates using camera intrinsics, i.e., focal lengths, principal point offsets, distortion coefficients, and video stream dimensions.

After filtering out missing values ($depth = 0$) and background pixels ($depth \geq 0.7$ m), z was obtained by extracting the median depth from the values within each bounding box. This method helped reduce the probability of wrong projections, leading to a more precise mapping from 2D RGB pixels to 3D points in space.

$$\begin{bmatrix} X \\ Y \\ Z \end{bmatrix} = z \cdot \begin{bmatrix} \frac{x - p_x}{f_x} \\ \frac{y - p_y}{f_y} \\ 1 \end{bmatrix} \tag{1}$$

The height and width of each bounding box were then calculated as the distances between the midpoints of its sides. A visualization of this process is shown in Fig. 3B-C. Considering a more elliptical shape, the fruitlet diameter was defined as the smaller of the two dimensions, in accordance with the caliper measurement, which consists in capturing the largest central diameter.

During in-field acquisition, high flower density may result in multiple clusters appearing in a single frame. To isolate the target corymb, the vision system should focus exclusively on the "foreground cluster", corresponding to the framed corymb if the acquisition is correctly performed. Therefore, any additional detected boxes were filtered out using Hierarchical clustering [26]. While this algorithm has been applied to flower segmentation in apple orchards [27], we adapted here to cluster bounding boxes using the Euclidean distance (2-norm) between fruitlets in 3D space.

We selected the complete linkage method, which returns well-defined spherical clusters consistent with corymb shapes. After some tuning of the clustering threshold t, we set 7.5 cm as the maximum distance between opposite fruitlets. As the last step, the bounding boxes of the closest cluster were retrieved based on the centroid distance from the camera, computed as the vector norm of centroid coordinates (x, y, z).

2.3. Frame filtering

2.3.1. Frame selection

Camera-based data collection in complex environments presents several challenges, including human errors and environmental factors, which can disrupt the consistency of object tracking. During rapid surveys with machinery, it's common to accept a certain tolerance for

missed fruits to achieve efficient data acquisition over large areas — especially given the issues of occluded and overlapped targets, as well as the difficulty in distinguishing fruits and background [28,29]. However, in scenarios where high accuracy of each fruit is required and ongoing work operations impact tracking reliability across long frames, counting the number of unique tracklet IDs may lead to a high rate of false positives. Consequently, retrieving individual measurements for each fruitlet becomes difficult and strongly dependent on the accuracy of the survey.

To overcome this problem, we introduced the frame filtering block (Fig. 3D), offering a novel approach to analyzing sequences of video frames when traditional object tracking methods fail to be reliable.

Assuming that each video included frames where all visible fruits were clearly displayed, we implemented a simple yet effective method to identify these most informative frames. The approach analyzed the temporal trend of the detected fruitlet count per video, under the assumption that this number remained consistent across corrected consecutive frames.

Post-processing of the YOLO output started by computing the first-order discrete difference of the number of detected fruitlets N_f across frames, as shown in Eq. 2. This numerical operation, previously used in specific computer vision tasks such as tracking moving animals in dense environments [30], was applied here with a novel adaptation.

$$\Delta N_f(t) = N_f(t+1) - N_f(t) \tag{2}$$

By examining changes in sequential data points, we identified frames where the count remained unchanged and selected those with the highest number of detected fruitlets. The result was a set of n frames f_i — not necessarily consecutive — that display a stable number of detections in the foreground corymb (see Fig. 4). We decided to extract the frames with the maximum number of fruitlets as they are more likely to reflect the true count, though at the risk of a higher false positive rate. A seemingly effective alternative approach would have been to take the most frequent count in the sequence (the mode), which theoretically provides a more robust measure. However, this method proved to be less effective in cluttered videos, where fruitlets are hard to detect.

After extracting the key frames from the video sequence, we iteratively applied the Hungarian algorithm to assign bounding boxes in the current frame to the most likely detections in the next frame, reconstructing the trajectory throughout the video acquisition, similar to other agricultural tracking problems [28,31].

This algorithm constructed a linear assignment model by creating a cost matrix, where each entry represented the absolute difference in diameters between corresponding clusters in the previous and current frames. After identifying the optimal one-to-one assignment, the diameters of the current frame were appended to their respective tracked lists, enabling a dynamic update of fruitlet data over time.

This approach grouped fruitlets with similar sizes, achieving results comparable to those of traditional tracking methods. Ultimately, we computed the median value of each group to determine the individual measures of the target corymb.

2.3.2. Anomaly detection

The frame selection process assumed that each predicted bounding box is a true positive, though this is not always the case. Collecting data from multiple orientations improves accuracy by capturing objects from different perspectives, yet missed detections and false positives may introduce a source of error. False negatives can be mitigated by selecting the most informative frames, while false positives require filtering. However, automatically distinguishing incorrect predictions from actual apple fruitlets remains challenging.

Similar to the hypothesis of count consistency, we can assume that reliable (true) bounding boxes are those that remain “stable” throughout the video. On the other hand, objects detected intermittently or outside the usual framed space may be considered spurious detections, i.e., anomalies.

The dataset obtained through the steps outlined in Section 2.2.2 consisted of a multivariate time series, with the frame number representing the temporal dimension. Table S.2 provides an example of this tabular data, which includes the 3D coordinates of bounding box centers, confidence scores, diameter estimates, fruitlet count (i.e., the number of fruitlets per frame), and frame number. These features were

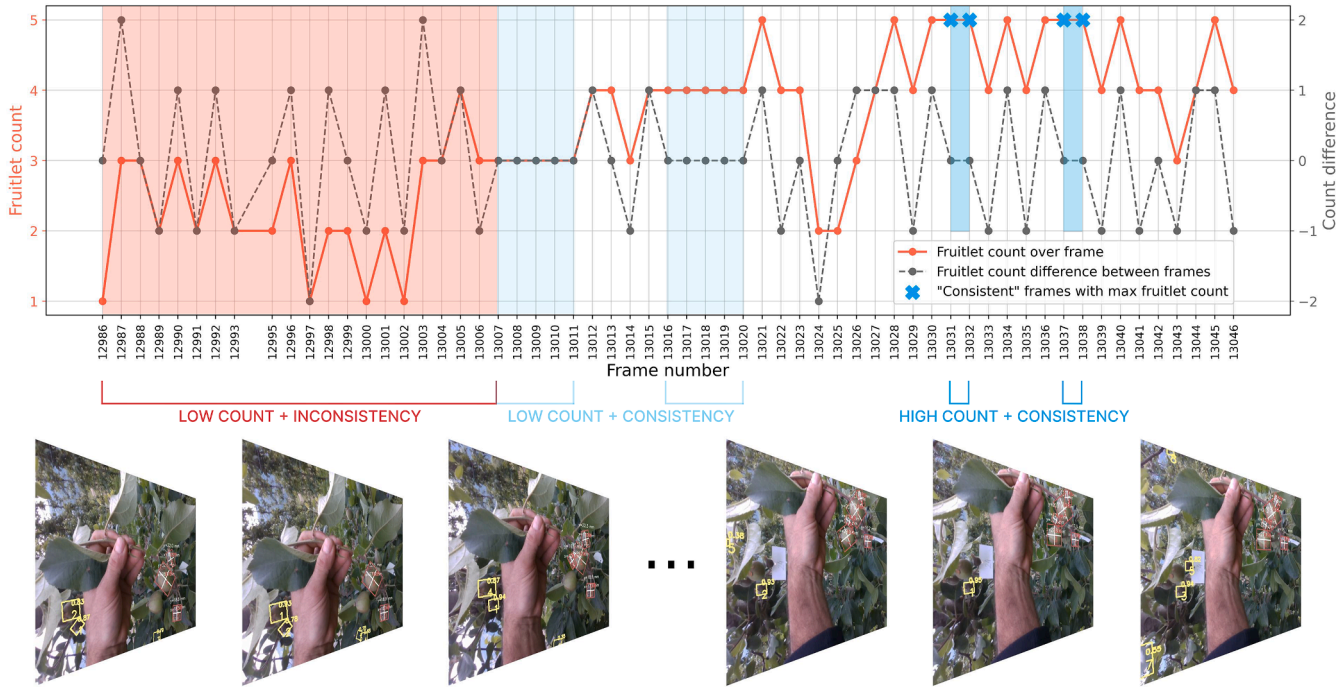

Fig. 4. Identification process of the most informative frames in a video sequence. The upper plot shows the trend of fruitlet count across frames (in red), along with the first-order discrete difference. Colored areas highlight frames where detection is unstable (red), stable but incomplete in detecting all fruitlets (lightblue), and frames with the highest detection count that meet the consistency requirement. These selected frames are used for the subsequent analysis.

used to analyze correlations across frames and identify potential outliers.

Anomaly detection in this kind of complex and high dimensional dataset can be tackled using neural networks, given their ability to learn non-linear relationships between features. Among them, Variational autoencoders (VAEs) are particularly useful for unsupervised anomaly detection when outliers are difficult to define from the existing features, capturing data uncertainty through probabilistic modeling. Typically used for image generation, this type of architecture consists of two main components — an encoder and a decoder — that work together to learn the latent representation of normal data and reconstruct complex data distributions by minimizing the difference between input data and reconstructed output, known as the reconstruction error [32]. The complexity increases further in higher-dimensional spaces, especially if the variables involved are correlated in unusual ways.

Preprocessing included a feature normalization to ensure uniform scaling across all input features. The model architecture is based on a fully connected VAE, where the encoder compresses the normalized data into a lower-dimensional latent space ($Z = 5$), followed by a reparameterization trick and a decoder, which mirrors the structure of the encoder [33]. The model was trained using a combination of binary cross-entropy reconstruction loss and Kullback-Leibler divergence regularization. Spurius bounding boxes were detected using the mean squared error (MSE) between input and reconstructed output, applying a 94th percentile threshold. This conservative approach helped identify more anomalies, though at the risk of excluding correct detections. For further details of the VAE parameters, refer to Table S.3 in the supplementary material.

2.4. Performance assessment

To systematically evaluate the steps of this research, we divided the assessment by considering the three main tasks of our pipeline, as outlined in Section 2.2:

- detection
- size estimation
- counting

While object detection validation relies on standard metrics commonly used in such analysis, for size estimation and counting we employed a combination of statistical and self-defined metrics. This approach was necessary due to the impracticability to compare individual fruitlets within corymbs and our need to provide an objective indication of the system performance for potential users, such as farmers or technicians.

Detection accuracy was evaluated using Average Precision (AP), the standard metric for object detection tasks, which corresponds to the area under the precision-recall curve (see Eq. 3). Precision and recall, which refer respectively to the accuracy of the positive predictions (true positives, TP) and the ability of the model to find all relevant instances (true positive and false negatives, TP+FN), are plotted against each other to highlight the trade-off between them. AP is then computed by integrating these two metrics across different Intersection over Union (IoU) thresholds, which measure how much the predicted box overlaps with the ground truth [34].

Specifically, we evaluated AP@0.5 and AP@[0.5:0.95], with the former calculated at an IoU threshold of 0.5 and the latter averaged over ten IoU thresholds from 0.5 to 0.95 with a step size of 0.05 [35]. These metrics were computed across the seven survey dates to identify the growth stages that were most challenging for the model. This step provided a measure of its generalization capabilities, representing a crucial phase to retrieve reliable and accurate measures.

$$AP = \int_0^1 p(r)dr \tag{3}$$

A second evaluation of YOLO detector — related to both the clustering method and the complexity of data acquisition — was the number of videos where the model failed to detect the correct number of fruitlets in any frame. This index was essential for guiding subsequent validations, particularly for fruitlet counting.

Once the images were processed, the frame selection mechanism analyzed the detection output by clustering nearby fruitlets, removing false positive outliers across frames, selecting the most informative frames, and finally retrieving individual fruitlet measures for each video.

To address sizing accuracy, we measured the difference between the estimated fruitlets and caliper measurements using Root mean squared error (RMSE) and Relative root mean squared error (RRMSE), as defined in Eqs. (4) and (5). Here, y_i represents the ground-truth values, $\hat{y}_i$ the estimated diameter, and N the total number of fruitlets in the corymb. Since individual fruitlets were not labeled, we applied the Hungarian algorithm for each corymb to establish a direct match between the estimated and true values.

RRMSE was specifically computed to account for variations in fruit size by normalizing the RMSE relative to the average observed value $\bar{y}$ within each predefined size range, where M denotes the number of fruitlets in that category.

Ultimately, counting precision was assessed by analyzing the number of unmatched fruitlets, classified as false positives (FP) and false negatives (FN), to compute the mean counting error. This error, calculated as the ratio of unmatched fruitlets to the total videos analyzed, represents the expected counting error per corymb during field acquisition. Additionally, correctly processed videos were identified as those where the estimated fruitlet count exactly matched the ground-truth value of the corymb.

Results are presented both with and without anomaly detection to demonstrate the impact of the Variational Autoencoder on system performance.

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2} \tag{4}$$

$$RRMSE = \sqrt{\frac{\frac{1}{M}\sum_{i=1}^{M}(y_i - \hat{y}_i)^2}{\bar{y}}} \cdot 100 \tag{5}$$

3. Results and discussion

The model achieved an overall accuracy of 0.895 in AP@0.5 on the validation set and 0.894 on the test set, consistent with previous studies. For instance, Wang and He reported an F1 Score of 91.5 [10], while Sapkota et al. reached AP@0.5 and AP@0.75 scores of 0.94 and 0.91, respectively [12]. While data collection was conducted at similar distances to the canopy, our study addressed the additional challenge of obtaining an object detection model that could generalize across different growth stages — marking the first attempt in the apple-growing literature. These detection results provided a solid basis for the following video analysis.

The outcomes on the test set for each survey date are presented in Fig. 5, together with Table S.4 of the supplementary material, which also includes the precision and recall values. As shown, the AP@0.5 generally exceeds 0.9, except for the first survey date, where it falls to 0.87. Similarly, the AP@[0.5:0.95] is generally around 0.8, but decreases to 0.65 during the initial monitoring. This drop can be attributed to the smaller fruit sizes and their flower-like shape, which makes them harder to detect, particularly when partially occluded by leaves. Variability across survey dates was low, indicating a good generalization: precision

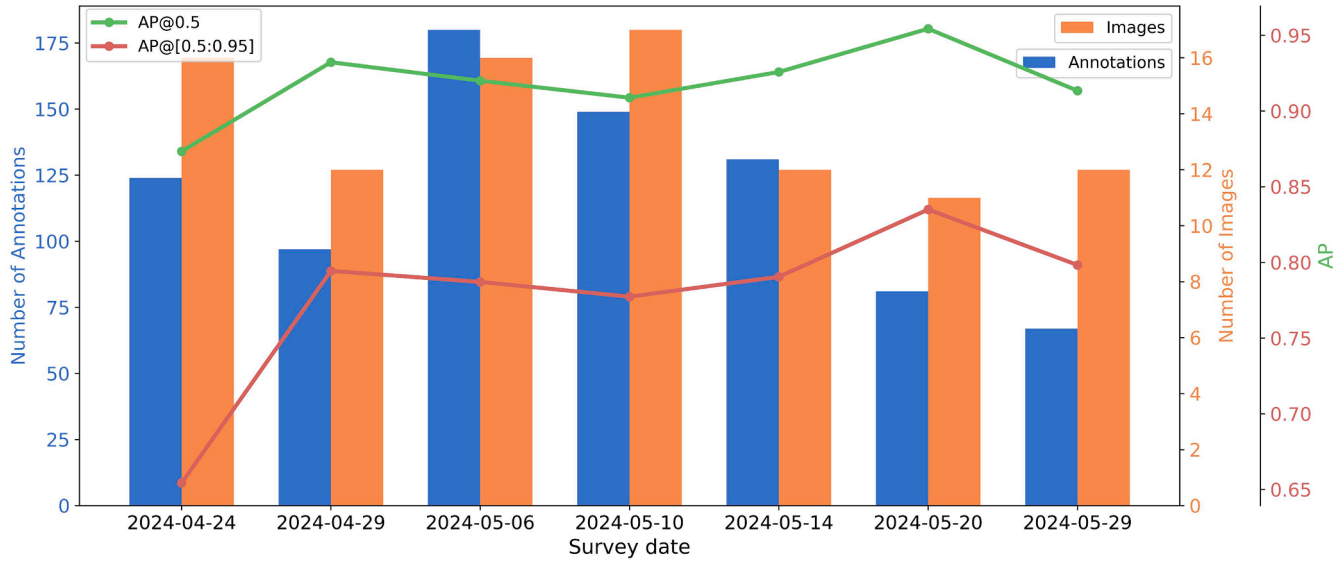

Fig. 5. Overview of the AP performance across survey dates, along with the number of images and annotations of the test dataset.

consistently remained around 0.9, except for the third survey, while Recall exhibited slightly more variation, ranging between 0.8 and 0.9 on average. The higher rate of false negatives, mainly associated with distant objects, was considered negligible, as the system focused on bounding boxes closest to the camera, where predictions were more likely to be correct.

Regarding the frame selection approach, we first identified the number of videos where the model did not manage to detect the correct number of fruitlets in any frame. As discussed in Section 2.4, this quantity is linked to the performance of the object detector, as well as the clustering method and the complexity of data acquisition.

Out of 234 videos, 55 (23,5%) fell into this category. As shown in Table S.5, these errors were mainly concentrated in the earlier survey dates, largely due to the challenge of detecting smaller fruitlets. Nonetheless, some errors also occurred in later surveys, potentially caused by dense occlusions, increased leaf coverage, data acquisition errors, and other inaccuracies in the model training.

Using our methodology with frame filtering and the anomaly detector, the system achieved a complete detection match in 56,4% of the videos (132/234), leaving the remaining 102 with incomplete matches. If we exclude the previously mentioned 55 videos (over 50% of the failures) where the correct number of fruitlets was never visible, the post-processing phase allowed to achieve a complete match in almost 75% of the cases.

The final results included 918 true positives, 136 false negatives and 12 false positives. The mean counting error of the model was approximately 0.63 fruitlets, calculated by dividing the total number of unmatched fruitlets by the 234 videos analyzed. Figure S.1 illustrates the distribution of unmatched detections across different fruitlet diameters. The analysis on the type of buds revealed three key insights:

- The vision system performed less effectively during the early growth stages, when fruitlets were smaller and corymbs contained a higher number of fruitlets.
- Lateral corymbs exhibited the highest rate of incorrect detections, likely due to their limited growth and the distinct size distribution of the fruitlets throughout the phenological stages.
- The model showed strong performance in analyzing fruit clusters originating from apical buds, particularly during the later stages of growth. The difference in unmatched fruitlets between apical and spur corymbs can be attributed to their position within the canopy. The latter tend to be located deeper within the foliage, where dense leaf coverage can obstruct the camera's field of view, making data acquisition more challenging. In contrast, annual corymbs are more exposed, facilitating the acquisition process, and thus the detection. This result is consistent with a recent study [36], which analyzed image acquisition in different apple training systems. Unlike conventional 3D training systems such as Tall Spindle, 2D canopies, like those in Guyot-trained orchards, bear fruits on simpler, shorter structures, reducing occlusions and improving light exposure.

A more detailed analysis of the 136 false negatives revealed the following findings:

- 40% of FN represented the only missing fruitlet in the corymb;
- 22% were part of pairs of fails;
- 4% occurred in corymbs where three fruitlets were missed;
- In only one instance four fruitlets were lost, making it a true outlier.

As far as the evaluation of fruitlet sizing is concerned, Fig. 6A shows the correlation between predicted diameters and the caliper measurements. The three bud types exhibited R^2 values of 0.974, 0.863, and 0.976 respectively, indicating a strong correlation.

This is further supported by the distribution of all RMSEs (see Figure S.2 in the supplementary material), which revealed a mean of 1.05 mm and a mode of 0.72 mm — outperforming a comparable study [12]. The Root mean squared errors were obtained by comparing predicted diameters with observed measurements for each corymb.

Focusing on the predicted - measured plot, the majority of deviations are clearly concentrated in the low size region, which is mostly populated by fruitlets from lateral buds. Given the chosen period for the experimental study, the 5-20 mm range presents more observations, becoming sparser over time due to the phenological fruitlet drop, as previously shown in Fig. 2B. While overestimation errors are common at the beginning of the acquisition period, the system tends to slightly underestimate true dimensions as growth progresses. This behavior is not actually significant, as the plot depicts the absolute size error. To provide a clearer understanding, Fig. 6B shows the weight of these errors across different diameter scales divided by bud types. RRMSE was computed only for diameter ranges that included at least five observations to ensure the reliability of the analysis and minimize the influence of outliers. As a result, no values are reported for apical and spur corymbs in the 0–5 mm and 40–45 mm ranges, nor for lateral bunches above 20 mm. Overall, the error tends to decrease with fruitlet size, reaching a minimum of less than 4% in the 30-35 mm range. Notably, lateral buds show the highest errors — around 35% in the 0–5 mm range

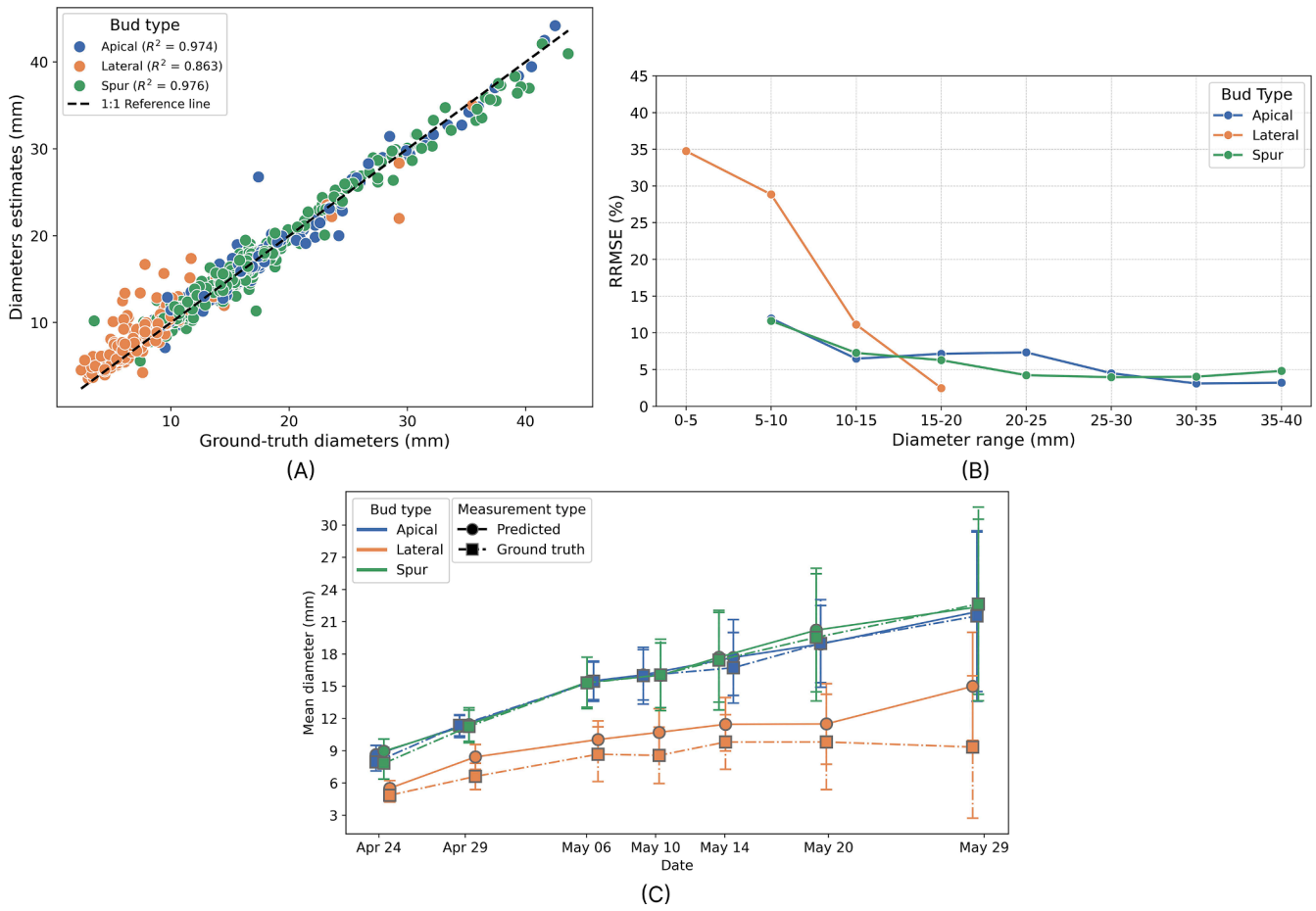

Fig. 6. Results evaluation by bud type. (A) Comparison between ground-truth and predicted diameters. (B) RRMSE across diameter ranges. (C) Comparison between estimated and measured mean diameter trends over time.

and 30% in the 5–10 mm range — corresponding to an average overestimation of approximately 2 mm. Beyond these early stages, the RRMSE stabilizes between 4% and 10%, indicating highly accurate size estimation [37] when comparing ground-truth and predicted pairings.

Finally, Fig. 6C presents a comparison of the growth trends between ground truth and estimated values. Since this analysis focused on the mean diameter of the corymbs, unmatched fruitlets are also included in the computation, leading to higher relative errors. The vertical error bars, which represent the mean standard deviation of the fruitlets in each corymb, give an insight into their size variability during growth.

The increase in mean diameter over time reflects the natural progression of phenological development: for apical and spur buds, this growth pattern is comparable, with the mean fruitlet size increasing from approximately 9 to 22 mm. The model estimates closely align with the ground truth, as indicated by the nearly overlapping trend lines. In contrast, lateral buds exhibit a stronger fruit drop, with a minimal growth of only some millimeters. The vision system struggled to track this behavior accurately, resulting in a consistent slight overestimation — especially evident in the final survey, where the discrepancy became significant. This issue is likely due to the incomplete identification of the corymbs: as far as the early development progressed and the fruitlet drop increased, the model managed to detect only the largest remaining object and failed to account for the smaller and wilting ones. As a result, the mean diameter of the corymb was skewed upward, particularly in the final survey.

An interesting observation was the impact of integrating anomaly detection. Without it, the number of unmatched fruitlets remained almost the same (150 vs. 148), while the distribution between FNs and FPs shifted significantly: undetected fruitlets decreased to 118 while incorrect detections increased to 31. Using the anomaly detector reduced the number of false positives, though at the risk that some true positives could be filtered out, especially if the model was not properly trained. Nevertheless, the utility of this step is clear as omitting it would have resulted in fewer videos with a complete match (125 vs. 132), a higher mean RMSE of 1.15 mm, and a mode of 0.76 mm.

Although a direct comparison is not possible due to differences in deep learning models and the number of labeled fruitlets, we can assert that the use of oriented bounding box annotations led to more accurate diameter size estimates, with a mean RMSE smaller than 0.3 mm. This improvement is linked to the possibility to better fit fruitlet shape, enhancing estimation accuracy.

3.1. Critical assessment

The proposed workflow represents a step toward advanced and objective methods of data collection, with the potential to replace the subjectivity of traditional manual practices, enabling more efficient and scalable orchard monitoring in critical early-season stages. The video acquisition process, designed to be corymb-specific, could be easily integrated into existing field surveys without requiring particular training for agronomists. With further development, the system could be made highly portable using an NVIDIA Jetson paired with a handheld tripod, or alternatively mounted on a robotic arm of unmanned ground vehicles, as tested by Sapkota et al. [12].

Despite the promising findings, the proposed workflow exhibited some limitations along the different blocks. Below, we provide a descriptive summary of these issues along with possible suggestions for improvement, leaving the analytical evaluation for future

experimentations.

- *object detection performance*: while the deep learning model demonstrated good performance on the test set, it did not achieve perfect generalization. Occasional false positives and missed detections were observed, particularly under poor lighting conditions and with small and wilting fruitlets, even across consecutive frames.
- *missing depth data*: blurry frames and partially occluded fruitlets could not always be resolved using the median distance approximation. In such cases, the vision system may fail size estimation despite the correct detection.
- *hierarchical clustering*: the current fruitlet clustering based on bounding box positions may sometimes exclude correct predictions, if considered outside the main cluster. Optimal results were obtained through a recursive trial-and-error process.
- *size estimation approach*: our method relied on bounding box dimensions as a proxy for approximating fruitlet size. While segmentation models and shape fitting techniques [12,38,39] could maybe improve accuracy by considering the fruit's actual shape, their effectiveness strongly depends on actual image data.
- *anomaly detection*: using tabular data from video frames as input to Variational Autoencoder (VAE) is a novel yet effective approach in this domain. However, its unsupervised nature, combined with limited training data, makes it challenging to evaluate its effectiveness and reduce its reliability in filtering false positives detections.
- *hardware settings*: enhancing RGB and depth resolution, along with increasing the frame rate, could likely improve the initial workflow stages but would also impose constraints on the acquisition process, as discussed in Section 2.1.1.

4. Conclusions

In this work, we introduced a novel configuration designed to rapidly and reliably estimate the diameters of apple fruitlets in complex agricultural environments. By leveraging different learning techniques and an RGB-D camera, we captured multiple views of apple corymbs throughout early development stages. These images were processed by a YOLO network, which monitored the growth and changes of fruitlets over time. The model showed consistent detection performance, with AP@0.5 values increasing from 0.87 shortly after fruit set to 0.95 in later stages. The predicted bounding boxes and their estimated diameters were compared across frames to extract the most informative images containing the true number of target fruitlets. Additionally, a VAE was tested to filter out false positive bounding boxes, helping in the post-processing frame analysis.

The system achieved 918 TPs, 136 FNs and 12 FPs. Validation of true positives against traditional caliper measurements revealed high accuracy in size estimation, with a mean RMSE of 1.05 mm. Fruitlet counting resulted in an average error of 0.63 fruitlets per video, highlighting some limitations in tracking and sizing smaller fruitlets, especially in cases affected by leaf occlusion and lighting disturbances. The current results represent a solid base for advancing towards data-driven horticultural practices, suggesting the need for alternative orchard designs, such as the Guyot system, which may facilitate better visibility and access to fruitlets for automated systems. Future work will focus on additional tests to evaluate alternative training systems, expanding the dataset, and optimizing workflow performance.

Funding

This study was carried out within the Interconnected Nord-Est Innovation Ecosystem (iNEST) and received funding from the European Union Next-GenerationEU (PIANO NAZIONALE DI RIPRESA E RESILIENZA (PNRR) – MISSIONE 4 COMPONENTE 2, INVESTIMENTO 1.5 – D.D. 1058 23/06/2022, ECS00000043). This manuscript reflects only the authors' views and opinions, neither the European Union nor the European Commission can be considered responsible for them.

CRediT authorship contribution statement

Giorgio Checola: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Data curation. **Damiano Moser:** Writing – review & editing, Validation, Resources, Conceptualization. **Paolo Sonego:** Writing – review & editing. **Cristian Iob:** Validation, Resources. **Franco Micheli:** Resources, Conceptualization. **Pietro Franceschi:** Writing – review & editing, Supervision.

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.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.atech.2025.100964.

Data availability

The data that support the findings of this study are openly available in Zenodo at https://doi.org/10.5281/zenodo.14844598. The code used to process and analyze the data is available in our GitHub repository at https://github.com/checolag/apple-fruitlet-detection-and-sizing.

References

[1] S.p. Monselise, E.e. Goldschmidt, Alternate bearing in fruit trees. Hortic. Rev, John Wiley & Sons, Ltd, 1982, pp. 128–173, https://doi.org/10.1002/9781118060773.ch5.
[2] A. Botton, G. Eccher, C. Forcato, A. Ferrarini, M. Begheldo, M. Zermiani, S. Moscatello, A. Battistelli, R. Velasco, B. Ruperti, A. Ramina, Signaling pathways mediating the induction of apple fruitlet abscission, Plant Physiol 155 (2011) 185–208, https://doi.org/10.1104/pp.110.165779.
[3] G. Costa, A. Botton, Hierarchy and strategy: how do they affect thinning response? Acta Hortic (2022) 55–59, https://doi.org/10.17660/ActaHortic.2022.1341.8.
[4] D.W. Greene, A.N. Lakso, T.L. Robinson, P. Schwallier, Development of a fruitlet growth model to predict thinner response on apples, HortScience 48 (2013) 584–587, https://doi.org/10.21273/HORTSCI.48.5.584.
[5] J. Jakopic, A. Zupan, K. Eler, V. Schmitzer, F. Stampar, R. Veberic, It's great to be the King: apple fruit development affected by the position in the cluster, Sci. Hortic. 194 (2015) 18–25, https://doi.org/10.1016/j.scienta.2015.08.003.
[6] S. Bargoti, J.P. Underwood, Image segmentation for fruit detection and yield estimation in apple orchards, J. Field Robot. 34 (2017) 1039–1060, https://doi.org/10.1002/rob.21699.
[7] A. Gongal, A. Silwal, S. Amatya, M. Karkee, Q. Zhang, K. Lewis, Apple crop-load estimation with over-the-row machine vision system, Comput. Electron. Agric. 120 (2016) 26–35, https://doi.org/10.1016/j.compag.2015.10.022.
[8] P. Roy, A. Kislay, P.A. Plonski, J. Luby, V. Isler, Vision-based preharvest yield mapping for apple orchards, Comput. Electron. Agric. 164 (2019) 104897, https://doi.org/10.1016/j.compag.2019.104897.
[9] D. Ahmed, R. Sapkota, M. Churuvija, M. Karkee, Machine vision-based crop-load estimation using YOLOv8, arXiv.Org (2023). https://arxiv.org/abs/2304.13282v1 (accessed September 20, 2024).
[10] D. Wang, D. He, Channel pruned YOLO V5s-based deep learning approach for rapid and accurate apple fruitlet detection before fruit thinning, Biosyst. Eng. 210 (2021) 271–281, https://doi.org/10.1016/j.biosystemseng.2021.08.015.
[11] S.R. Khanal, R. Sapkota, D. Ahmed, U. Bhattarai, M. Karkee, Machine Vision system for early-stage apple flowers and flower clusters detection for precision thinning and pollination, IFAC-Pap. 56 (2023) 8914–8919. https://doi.org/10.1016/j.ifacol.2023.10.096.
[12] R. Sapkota, D. Ahmed, M. Churuvija, M. Karkee, Immature green apple detection and sizing in commercial orchards using YOLOv8 and shape fitting techniques, IEEE ACCESS 12 (2024) 43436–43452, https://doi.org/10.1109/ACCESS.2024.3378261.
[13] H. Freeman, G. Kantor, Autonomous apple fruitlet sizing with Next Best View planning, in: 2024 IEEE Int. Conf. Robot. Autom, ICRA, 2024, pp. 15847–15853, https://doi.org/10.1109/ICRA57147.2024.10610226.
[14] G. Eccher, S. Ferrero, F. Populin, L. Colombo, A. Botton, Apple (Malus domestica L. Borkh) as an emerging model for fruit development, Plant Biosyst. - Int. J. Deal.

Asp. Plant Biol. 148 (2014) 157–168, https://doi.org/10.1080/11263504.2013.870254.
[15] A. Kour, R. Bhat, C. Bishnoi, N. Gupta, Flowering, pollination and fruit set. Apples, CRC Press, 2022.
[16] Intel Corporation, Intel® RealSenseTM Product Family D400 Series Datasheet, Intel Corporation, 2023. https://www.intelrealsense.com/wp-content/uploads/2023/03/Intel-RealSense-D400-Series-Datasheet-March-2023.pdf (accessed November 2, 2025).
[17] Intel RealSense, pyrealsense2 – RealSense SDK 2.0 Python compatibility wrapper, (n.d.). https://github.com/IntelRealSense/librealsense/tree/master/wrappers/python.
[18] E.S.L. Gastal, M.M. Oliveira, Domain transform for edge-aware image and video processing. ACM SIGGRAPH 2011 Pap, Association for Computing Machinery, New York, NY, USA, 2011, pp. 1–12, https://doi.org/10.1145/1964921.1964964.
[19] M.-V. Hanke, H. Flachowsky, A. Peil, C. Hättasch, No flower no fruit—Genetic potentials to trigger flowering in fruit trees, Genes Genomes Genomics 1 (2007) 1–20.
[20] R.K. Volz, I.B. Ferguson, E.W. Hewett, D.J. Woolley, Wood age and leaf area influence fruit size and mineral composition of apple fruit, J. Hortic. Sci. 69 (1994) 385–395, https://doi.org/10.1080/14620316.1994.11516468.
[21] Tzutalin, LabelImg. https://github.com/tzutalin/labelImg, 2015.
[22] J. Redmon, S. Divvala, R. Girshick, A. Farhadi, You only look once: unified, real-time object detection, (2016). https://doi.org/10.48550/arXiv.1506.02640.
[23] R. Khanam, M. Hussain, YOLOv11: An overview of the key architectural enhancements, (2024). https://doi.org/10.48550/arXiv.2410.17725.
[24] G. Jocher, J. Qiu, Ultralytics YOLO11. https://github.com/ultralytics/ultralytics, 2024.
[25] A. Sharma, V. Kumar, L. Longchamps, Comparative performance of YOLOv8, YOLOv9, YOLOv10, YOLOv11 and Faster R-CNN models for detection of multiple weed species, Smart Agric. Technol. 9 (2024) 100648, https://doi.org/10.1016/j.atech.2024.100648.
[26] G. James, D. Witten, T. Hastie, R. Tibshirani, others, An introduction to statistical learning, Springer, 2013.
[27] C. Zhang, C. Mouton, J. Valente, L. Kooistra, R. van Ooteghem, D. de Hoog, P. van Dalfsen, P. Frans de Jong, Automatic flower cluster estimation in apple orchards using aerial and ground based point clouds, Biosyst. Eng. 221 (2022) 164–180, https://doi.org/10.1016/j.biosystemseng.2022.05.004.
[28] X. Liu, S.W. Chen, C. Liu, S.S. Shivakumar, J. Das, C.J. Taylor, J. Underwood, V. Kumar, Monocular camera based fruit counting and mapping with semantic data association, IEEE Robot. Autom. Lett. 4 (2019) 2296–2303, https://doi.org/10.1109/LRA.2019.2901987.
[29] Y. Tang, J. Qiu, Y. Zhang, D. Wu, Y. Cao, K. Zhao, L. Zhu, Optimization strategies of fruit detection to overcome the challenge of unstructured background in field orchard environment: a review, Precis. Agric. 24 (2023) 1183–1219, https://doi.org/10.1007/s11119-023-10009-9.
[30] G. Chen, S. Shen, L. Wen, S. Luo, L. Bo, Efficient pig counting in crowds with keypoints tracking and spatial-aware temporal response filtering, in: 2020 IEEE Int. Conf. Robot. Autom, ICRA, 2020, pp. 10052–10058, https://doi.org/10.1109/ICRA40945.2020.9197211.
[31] G.P. Matos, C. Santiago, J.P. Costeira, R.L. Saldanha, E.M. Morgado, Tracking and counting apples in orchards under intermittent occlusions and low frame rates, in: 2024: pp. 5413–5421. https://openaccess.thecvf.com/content/CVPR2024W/Vision4Ag/html/Matos_Tracking_and_Counting_Apples_in_Orchards_Under_Intermittent_Occlusions_and_CVPRW_2024_paper.html (accessed January 17, 2025).
[32] J. An, S. Cho, Variational autoencoder based anomaly detection using reconstruction probability, Spec. Lect. IE 2 (2015) 1–18.
[33] D.P. Kingma, M. Welling, Auto-encoding variational bayes, (2022). https://doi.org/10.48550/arXiv.1312.6114.
[34] M. Everingham, S.M.A. Eslami, L. Van Gool, C.K.I. Williams, J. Winn, A. Zisserman, The Pascal visual object classes Challenge: A retrospective, Int. J. Comput. Vis. 111 (2015) 98–136, https://doi.org/10.1007/s11263-014-0733-5.
[35] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, C. L. Zitnick, Microsoft COCO: Common objects in context, in: D. Fleet, T. Pajdla, B. Schiele, T. Tuytelaars (Eds.), Comput. Vis. – ECCV 2014, Springer International Publishing, Cham, 2014, pp. 740–755, https://doi.org/10.1007/978-3-319-10602-1_48.
[36] G. Bortolotti, K. Bresilla, M. Piani, L.C. Grappadelli, L. Manfrini, 2D tree crops training system improve computer vision application in field: a case study, in: 2021 IEEE Int. Workshop Metrol. Agric. For. MetroAgriFor, 2021, pp. 120–124, https://doi.org/10.1109/MetroAgriFor52389.2021.9628839.
[37] P.D. Jamieson, J.R. Porter, D.R. Wilson, A test of the computer simulation model ARCWHEAT1 on wheat crops grown in New Zealand, Field Crops Res 27 (1991) 337–350, https://doi.org/10.1016/0378-4290(91)90040-3.
[38] G. Bortolotti, M. Piani, M. Gullino, D. Mengoli, C. Franceschini, L.C. Grappadelli, L. Manfrini, A computer vision system for apple fruit sizing by means of low-cost depth camera and neural network application, Precis. Agric. (2024), https://doi.org/10.1007/s11119-024-10139-8.
[39] D. Mengoli, G. Bortolotti, M. Piani, L. Manfrini, On-line real-time fruit size estimation using a depth-camera sensor, in: 2022 IEEE Workshop Metrol. Agric. For. MetroAgriFor, 2022, pp. 86–90, https://doi.org/10.1109/MetroAgriFor55389.2022.9964960.