Looking behind occlusions: a study on amodal segmentation for robust on-tree apple fruit size estimation

Jordi Gené-Mola [a, b,*], Mar Ferrer-Ferrer [a], Eduard Gregorio [a], Pieter M. Blok [b], Jochen Hemming [b], Josep-Ramon Morros [c], Joan R. Rosell-Polo [a], Verónica Vilaplana [c], Javier Ruiz-Hidalgo [c]

[a] Research Group in AgroICT& Precision Agriculture - GRAP, Department of Agricultural and Forest Engineering, Universitat de Lleida (UdL) – Agrotecnio-CERCA Center, Lleida, Catalonia, Spain. (jordi.genemola@udl.cat, marfferrer98@gmail.com, eduard.gregorio@udl.cat, joanramon.rosell@udl.cat)

[b] Wageningen University and Research, 6700 AA Wageningen, the Netherlands (pieter.blok@wur.nl, jochen.hemming@wur.nl)

[c] Department of Signal Theory and Communications, Universitat Politècnica de Catalunya, Barcelona, Catalonia, Spain. (ramon.morros@upc.edu, veronica.vilaplana@upc.edu, j.ruiz@upc.edu)

Abstract

The detection and sizing of fruits with computer vision methods is of interest because it provides relevant information to improve the management of orchard farming. However, the presence of partially occluded fruits limits the performance of existing methods, making reliable fruit sizing a challenging task. While previous fruit segmentation works limit segmentation to the visible region of fruits (known as modal segmentation), in this work we propose an amodal segmentation algorithm to predict the complete shape, which includes its visible and occluded regions. To do so, an end-to-end convolutional neural network (CNN) for simultaneous modal and amodal instance segmentation was implemented. The predicted amodal masks were used to estimate the fruit diameters in pixels. Modal masks were used to identify the visible region and measure the distance between the apples and the camera using the depth image. Finally, the fruit diameters in millimetres (mm) were computed by applying the pinhole camera model. The method was developed with a Fuji apple dataset consisting of 3925 RGB-D images acquired at different growth stages with a total of 15335 annotated apples, and was subsequently tested in a case study to measure the diameter

* *Corresponding author.* E-mail address: jordi.genemola@udl.cat

of Elstar apples at different growth stages. Fruit detection results showed an F1-score of 0.86 and the fruit diameter results reported a mean absolute error (MAE) of 4.5 mm and R^2 = 0.80 irrespective of fruit visibility. Besides the diameter estimation, modal and amodal masks were used to automatically determine the percentage of visibility of measured apples. This feature was used as a confidence value, improving the diameter estimation to MAE = 2.93 mm and R^2 = 0.91 when limiting the size estimation to fruits detected with a visibility higher than 60%. The main advantages of the present methodology are its robustness for measuring partially occluded fruits and the capability to determine the visibility percentage. The main limitation is that depth images were generated by means of photogrammetry methods, which limits the efficiency of data acquisition. To overcome this limitation, future works should consider the use of commercial RGB-D sensors. The code and the dataset used to evaluate the method have been made publicly available at https://github.com/GRAP-UdL-AT/Amodal_Fruit_Sizing.

Keywords: Fruit detection; Fruit measurement; Yield estimation; Fruit visibility; Deep Learning; Precision Agriculture

1. Introduction

In modern fruit production optimization of the use of all inputs is desired. Instead of a treatment or application per field or plot, fruit trees should get precisely the treatment they need. This approach is commonly known as precision fruticulture/horticulture. The goal is to produce more with less, to reduce the inputs such as labour, water, fertilizer and chemicals, and by doing so, reduce potential environmental pollution. The latest sensor developments and data technology allow continuous monitoring of the chosen field and tree parameters with high spatial and temporal resolution. These data provide better information for management decisions.

The research and development of systems for fruit detection has been carried out for a considerable time given its enormous importance for the management of fruit farms (Slaughter and Harrell, 1987). One of the first and main applications is crop yield prediction. An accurate estimate, weeks or months in advance, of the fruit production allows optimal planning of the operations necessary for the management of the crop, as well as its collection, conservation and marketing (Anderson et al., 2019). Another application of fruit detection systems consists in mapping the predicted yield. These maps allow to analyse how the yield is distributed throughout the orchard in order to optimize the management based on the spatial variability

(Longchamps et al., 2022). Automated harvesting is another field of application of fruit detection systems. It is a line of research and development whose beginnings go back a few decades (Moltó et al., 1992). However, the technological advances of the last ten years, together with the greater economic accessibility of components, have contributed to the fact that, at present, it is a very intense research field (Kootstra et al., 2021).

The evolution of fruit detection systems should not only focus on detection but also on estimation of the size (and weight) of the detected fruit (Tijskens et al., 2016; Zhou et al., 2012). This allows estimation of the future yield in mass (kg or tons) and can also provide relevant information such as fruit quality and the optimal time to start harvesting. When measured several times during the growing season, size information can also be used to calculate the fruit growth rate and, together with models and decision support systems, improve orchard management by better managing the irrigation, nutrition and other agricultural tasks such as thinning (Robinson et al., 2008). The automated harvesting of fruit crops is another domain where fruit size estimation is relevant (Gongal et al., 2018). It opens up the possibility of selective harvesting of the fruits according to their size and quality. In addition, harvesting robots need to measure the fruits for more careful and gentler handling (Wang et al., 2017).

Traditionally, fruit sizing on the tree has been carried out manually by means of Vernier callipers or sizing rings. These manual procedures are error-prone, labour-intensive and time-consuming, which in practice limits measurements to a few samples (trees) of the orchard. To overcome these limitations, several optical-based methodologies have been developed for automatic in-field fruit sizing. Thermography and colour cameras have been used to estimate fruit diameters (Stajnko et al., 2004; Wang et al., 2018), but they require the use of calibration targets next to the measured fruits in order to perform image calibration (Lu et al., 2022). To avoid the need for calibration targets, currently there is a trend towards 3D-based sensing methodologies such as light detection and ranging (LiDAR) systems, RGB-D cameras and photogrammetry techniques (Gregorio and Llorens, 2021). LiDAR systems are insensitive to prevailing lighting conditions and advantage can be taken of their radiometric capabilities to estimate fruit size, as demonstrated by Tsoulias et al. (2020). RGB-D cameras are affordable sensors that simultaneously provide colour and depth data, but their performance is affected by high lighting conditions (Gené-Mola et al., 2020). With RGB-D cameras, Wang et al. (2017) were able to estimate the length and width of mangoes with RMSE values of

4.9 and 4.3 mm, respectively. The advent of low-cost photogrammetric software along with advances in computing power are driving the use of structure-from-motion (SfM) and multi-view stereo (MVS) for fruit size estimation (Grilli et al., 2021). In our previous work (Gené-Mola et al., 2021), we proposed a new apple fruit sizing methodology based on SfM and MVS which showed an MAE value of 3.7 mm and a coefficient of determination (R^2) of 0.91. However, it presented high processing times due to the intensive operations on which SfM is based.

One of the major challenges that sensor-based fruit sizing methods need to overcome is the measurement of partially occluded fruits. In addition, fruit occlusions also affect the performance of harvesting robots estimating the 3D grasping point for a robotic gripper. To overcome this challenge, in this work we propose to estimate the shape of partially occluded fruits by means of amodal instance segmentation. Amodal instance segmentation aims to predict the visible and occluded parts of each object of interest in an image (Li and Malik, 2016). It has been mainly applied in the field of mobile robotics and autonomous driving (Qi et al., 2019), while in the agriculture field it has been applied to predict the complete shape of broccoli heads (Blok et al., 2021). To date, the task of fruit instance segmentation has been applied to mask the visible region of fruits (Santos et al., 2020; Wang and He, 2022), which is known as modal instance segmentation. An example of modal segmentation is illustrated in **Fig. 1b** (Modal Mask). Alternatively, with the hypothesis that occluded regions are important to estimate the fruit size, we propose to mask not only the visible but also the occluded regions of the apples, as shown in **Fig. 1c** (Amodal Masks). To this end, we implemented a convolutional neural network (CNN) for simultaneous modal and amodal instance segmentation. The main contributions of this work are: 1) for the first time, simultaneous modal and amodal instance segmentation is applied to fruit images; 2) a robust method for fruit sizing and visibility estimation is proposed, based on the combination of modal and amodal segmentation in RGB-D images; 3) an analysis of the method performance depending on the amount of fruit visibility and the detection confidence is presented; 4) a performance evaluation in different apple varieties (Fuji and Elstar) and growth stages (green and ripe) is provided; and 5) the implemented code and the generated dataset with ground truth annotations are provided.

a) Image b) Modal Masks c) Amodal Masks

Fig. 1. Example of modal and amodal segmentation masks in a Fuji apple image.

After this introduction, the rest of the paper is structured as follows. **Section 2** provides detailed information about the generated dataset, the architecture and training details of the implemented CNN and explains the method to estimate fruit diameter from amodal masks. **Section 3** presents the fruit detection and sizing results, analyses the fruit sizing performance at different levels of fruit visibility, different detection confidences and different growth stages, and evaluates a case study which aims to estimate the mean fruit diameter per tree in a different apple variety and different growth stages. **Section 4** discusses the results and compares them to other methods from the state of the art. Finally, the main conclusions and future works are commented on in **Section 5**.

2. Materials and Methods

2.1. Dataset

Two different datasets were collected and used in this work. The first set was used to train, validate and test the proposed method, while the second was used as a case study carried out in a different orchard and under different conditions to those used to develop the method.

The first set consisted of 3925 annotated RGB-D images acquired in a Fuji apple orchard located in Agramunt (Catalonia, Spain). Data was collected at two different growth stages: in mid-July, corresponding to growth stage BBCH77 (Meier, 2001) when apples were green and about 70% of their final size (**Fig. 2a.left**); and in the first week of October, corresponding to growth stage BBCH85 when apples were at an advanced ripening stage (**Fig. 2a.right**). Raw images were taken with a handheld EOS 60D DSLR camera (Canon Inc. Tokyo, Japan) equipped with a 5184 x 3456 pixels CMOS APS-C sensor. Consecutive photographs were taken with an overlap higher than 75% to facilitate camera alignment and subsequently

estimate the images depth maps (**Fig. 3b**) by applying SfM and MVS using Agisoft Metashape software (Agisoft LLC, St. Petersburg, Russia, v1.6.4.). After generating the depth maps, the RGB-D images used in this work were generated by cropping raw image patches of 1024 x 1024 pixels (**Fig. 3**), obtaining a total of 1560 images from the BBCH77 growth stage and 2365 images from the BBCH85 growth stage.

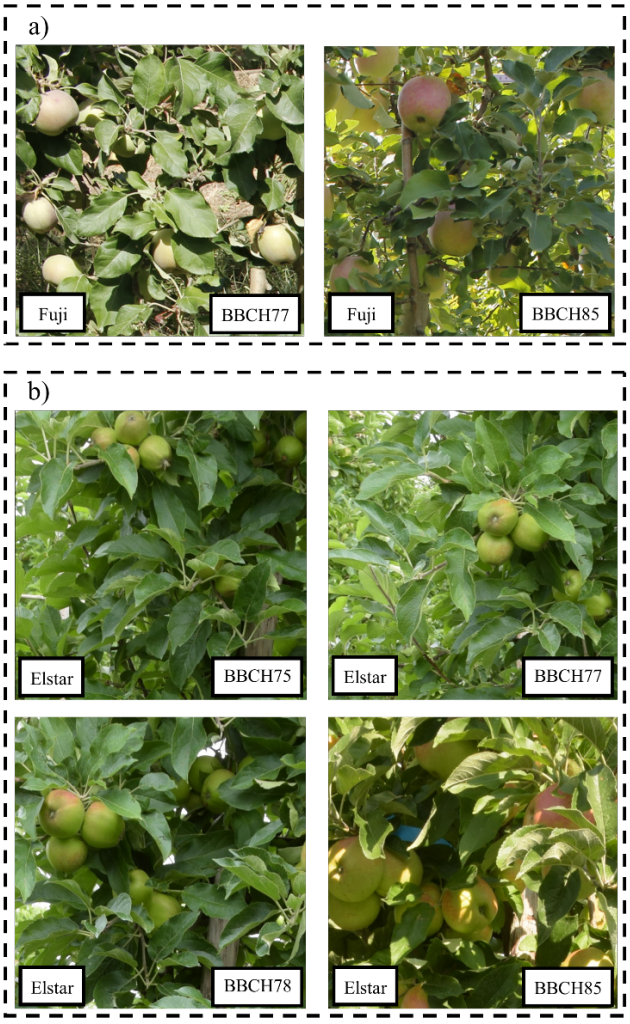
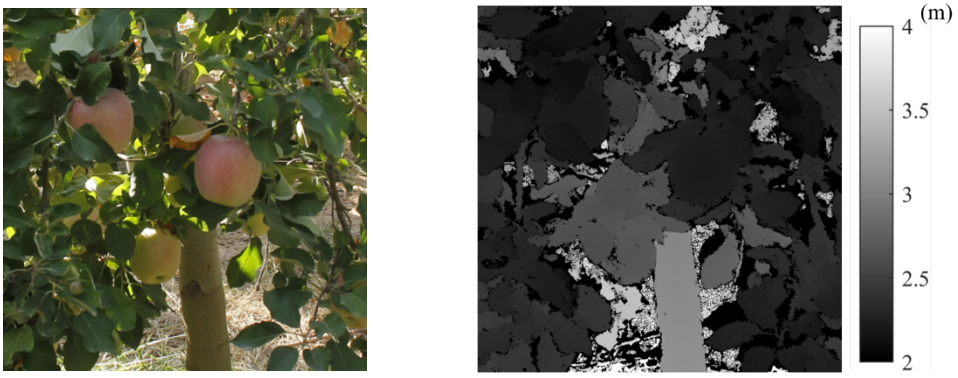


Fig. 2. a) Sample images from the dataset used to train, validate and test the methodology. Data acquired in a Fuji apple orchard at two growth stages: BBCH77 and BBCH85. b) Sample images from the case study carried out in an Elstar apple orchard at four different growth stages: BBCH75, BBCH77, BBCH78 and BBCH85.



(a) (b)

Fig. 3. Example of an RGB-D image from the Fuji dataset. a) Colour image (RGB). b) Depth image (D).

All apple instances visible in the images were annotated with two ground truth masks: modal and amodal. Modal masks include the pixels of the images belonging to the visible, modal, part of each apple (**Fig. 1b**), while the amodal masks refer to the visible and occluded part of each apple (**Fig. 1c**). The modal segmentation ground truth was manually annotated using the VIA annotator software (Dutta and Zisserman, 2019) , while amodal masks were obtained by projecting the apple shape from the 3D space onto the image plane. Since the raw images used in this work had a significant overlap (>75%), the 3D model of the imaged trees was generated using SfM and MVS (same procedure than the explained for generating depth images). The complete 3D model of partially occluded apples was obtained by fitting a sphere of diameter equal to the apple size following the procedure described in Gené-Mola et al. (2021). This complete 3D representation of apples was projected onto the image plane following the pinhole camera model, obtaining the corresponding amodal masks. Finally, the projected masks were manually corrected and refined (if necessary) using the VIA annotator software (Dutta and Zisserman, 2019). Apple ground truth diameter was manually measured in the field using a Vernier calliper. Since apples are not perfectly spherical, the ground truth diameter was considered the major axis. Each apple measure was assigned to an apple identification number (appleID) to have a pairwise correspondence between the measurements and the image annotations. This data annotation resulted in a total of 5837 apple instances annotated in BBCH77 images with a mean apple diameter of 54.0 mm (from 26.9 mm to 71.0 mm) and a total of 9498 apple instances annotated in BBCH85 images with a mean apple diameter of 77.4 mm (from 43.6 mm to 94.8 mm) (**Fig. 4a**). This dataset was randomly split into training (2304 images, of which 1036 from BBCH77

and 1268 from BBCH 85), validation (814 images, of which 275 from BBCH77 and 539 from BBCH 85) and test (807 images, of which 249 from BBCH77 and 558 from BBCH85) sets, obtaining approximately 60%, 20% and 20% of apples instances on each set, respectively. As it can be observed in **Fig. 4**, all dataset splits (training, validation and test) contain data from both maturity stages, from different fruit size and from different fruit visibilities. Since the dataset split was performed randomly, the 60%, 20% and 20% (training, validation and test) proportion was preserved at different apple diameter and visibility intervals (**Fig. 4.b** and **Fig. 4.d**). The dataset has been made publicly available at https://github.com/GRAP-UdL-AT/Amodal_Fruit_Sizing.

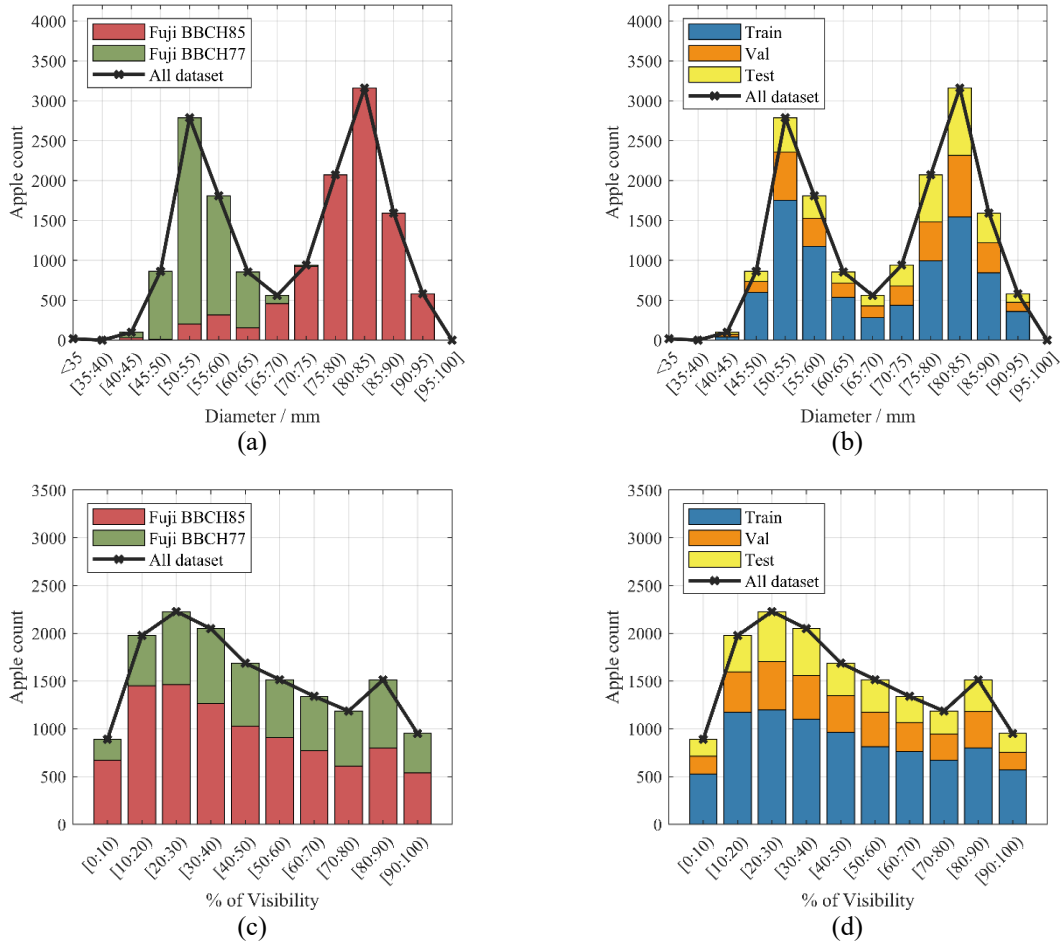


Fig. 4. a) Stacked histogram of apple diameters at different growth stages. b) Stacked histogram of apple diameters at different dataset splits. c) Stacked histogram of apple visibilities at different growth stages. b) Stacked histogram of apple visibilities at different dataset splits. Growth stages: BBCH 85 (red) and BBCH 77 (green). Dataset splits: training (blue), validation (orange) and test (yellow).

The data used for the case study was acquired in an Elstar apple orchard located in Randwijk (the Netherlands). Five different trees were imaged at four different dates (**Table 1**), obtaining data at different growth stages: BBCH75, BBCH77, BBCH78 and BBCH85 (**Fig. 2b**). To have a complete representation of trees, images were acquired from both sides of the row of trees. This allowed to evaluate the performance of the method (presented in **Section 3.3**) depending on the side from which images were acquired. The RGB-D images were obtained following the same procedure as for the previous dataset, but this time a Nikon Z6 camera (24.5 MP) was used. This case study consisted in measuring the mean diameter of the apples grown in each imaged tree. To evaluate the results, the mean diameter ground truth of each tree at each measured date was computed by averaging the manual measurement (with a Vernier calliper) of 15 apple samples randomly selected on each tree.

Table 1. Data used in the case study: number of trees measured and mean diameter measured at different growth stages.

Measurement date	Growth stage	Num. of measured trees	Mean diameter
21/06/2019	BBCH 75	2	40.2 mm
03/07/2019	BBCH 77	5	46.1 mm
16/07/2019	BBCH 78	5	55.3 mm
23/08/2019	BBCH 85	5	72.3 mm

2.2. Deep neural network architecture and training details

The convolutional neural network (CNN) used in this work is the one implemented by Blok et al. (2021). This CNN has the same network architecture as Mask R-CNN (He et al., 2017), except that an additional mask head branch was added to perform the amodal segmentation task (**Fig. 5**).

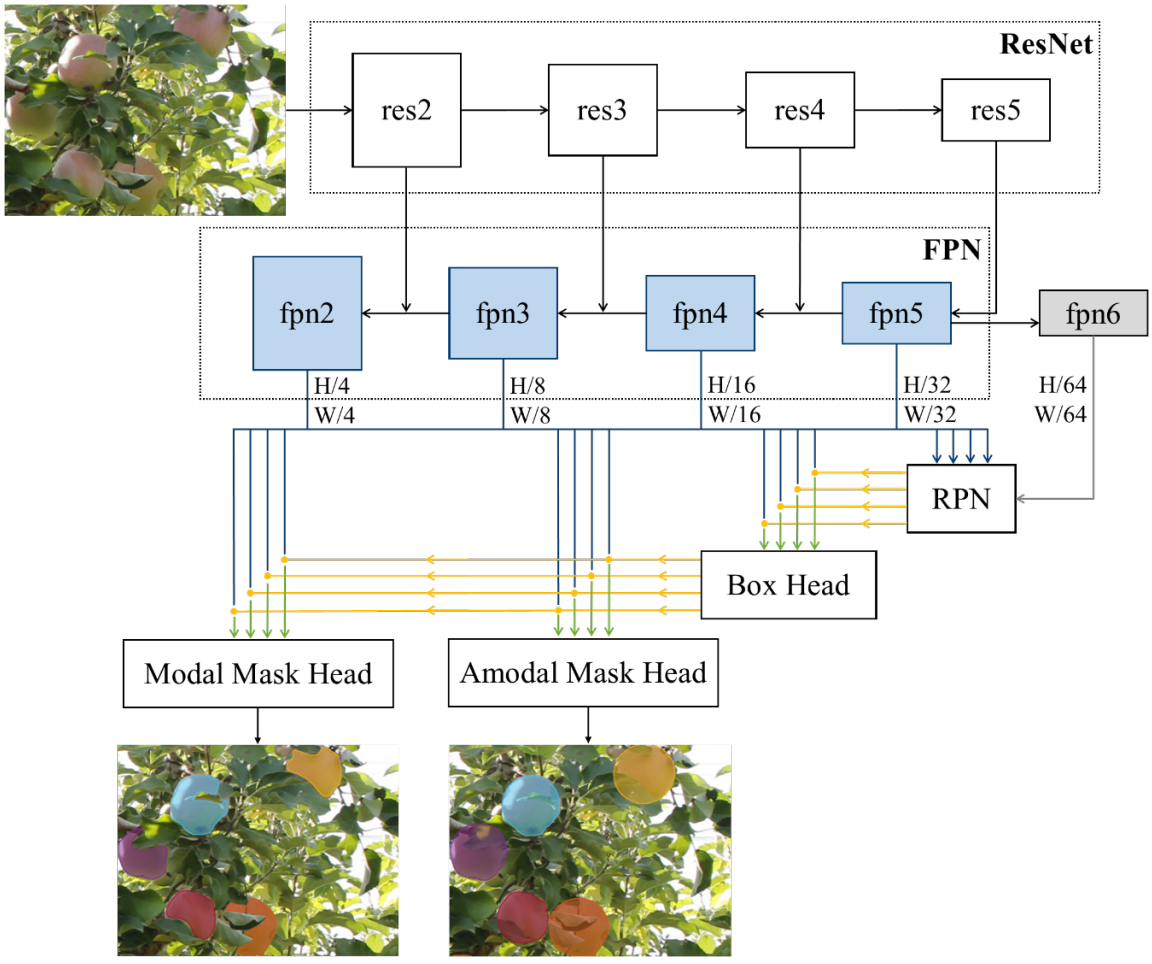


Fig. 5. Architecture of the convolutional neural network used for simultaneous modal and amodal mask segmentation.

The code baseline was the Mask R-CNN architecture from the Detectron2 library (Wu et al., 2019). ResNet-101 (He et al., 2016) was used as a backbone network, followed by the feature pyramid network (FPN) to extract feature maps at different scales. These feature maps were fed into the region proposal network (RPN) to identify promising regions of interest (ROIs) that were likely to contain an apple. The number of region proposals to be produced by the RPN was set to 512, which proved to yield better detection performance than the original number of 256. These ROIs were refined by means of the ROI align layer before sending it to the Box Head, which performed a regression to obtain the final corners of each bounding box. The L1-loss was used to calculate the regression error using the bounding box that encapsulated the amodal mask as detection ground truth. This was done because the amodal region is, by

definition, equal to or larger than the visible region. Thus, the modal and the amodal segmentation were applied inside the same amodal bounding box. The main contribution of the Mask R-CNN architecture of Blok et al. (2021) was the implementation of the two parallel segmentation branches: one for the modal (visible) mask and one for the amodal mask.

To train the network, we applied transfer-learning by initializing the network with the weights of the Mask R-CNN network pre-trained on the Microsoft Common Objects in Context (COCO) dataset (Lin et al., 2014). Then, the CNN was fine-tuned on the apple dataset (**Section 2.1**) using the stochastic gradient descent optimiser with a momentum of 0.9, a weight decay of 0.0001 and a learning rate of 0.02. The image batch size was set to 4, limited by the memory capabilities of the GPU that was used (NVIDIA GeForce GTX 1080 Ti) which has 11 Gb of memory. Data augmentation was applied with random horizontal flip with a probability of 0.5. After the training, the validation loss curve was inspected and the weights-file trained at the 3,000th iteration was selected to prevent overfitting (the network trained at this iteration had the lowest validation loss).

Fruit detection performance was evaluated in terms of precision (P), recall (R), F1-score and average precision (AP). A detection was considered positive if the detection confidence provided by the CNN was higher than the confidence threshold set after analyzing the validation results (**Section 3.1**). Then, a positive detection was classified as true positive (TP) if the intersection over union (IoU) between the ground truth and the detection bounding box was > 0.5. Following the standard definition of these metrics, P was computed as the ratio of TP and the total number of detections and R was computed as the ratio of TP with respect to the total number of ground truth annotated apples. The F1-score was computed as the harmonic mean of P and R, while the AP was computed as the area under the PR curve. The software was written in Python (version 3.0) with Pytorch (version 1.12) and Torchvision (version 0.13) as the deep learning libraries. The code has been made publicly available jointly with the dataset at https://github.com/GRAP-UdL-AT/Amodal_Fruit_Sizing.

2.3. From amodal masks to fruit diameter

The fruit diameter is extracted using both the modal and amodal segmentation of each detected apple. The modal segmentation was used to compute the distance from the camera to the fruit (excluding the occluded pixels, such as leaves, so that these did not interfere with the distance estimation). The distance

from the camera to the fruit in millimetres (mm), z, is computed as the average of the corresponding distances d_i for all pixels N in the modal segmentation (Eq. 1), where d_i is the value of pixel i in the depth image (**Fig. 3b**).

$$z = \frac{1}{N}\sum_{i=1}^{N} d_i \tag{1}$$

The amodal segmentation was used to compute the fruit diameter in pixels. Given the area, A, as the sum of all pixels in the mask that correspond to each apple in the amodal segmentation, the fruit diameter in pixels, d, can be calculated from Eq. 2 as:

$$d = 2\sqrt{\frac{A}{\pi}} \tag{2}$$

The relationship between the fruit diameter in the 2D image (expressed in pixels) and the corresponding diameter in the real 3D world (expressed in mm) is estimated using a pinhole camera model. **Fig. 6** shows a graphical example of the pinhole camera model. Let $X = [x\ y\ z]^T$ be a representation of a 3D point, $p = [u\ v\ w]^T$ the homogeneous coordinates of this point on the 2D image using the pinhole camera, and K the 3x3 camera matrix that represents the intrinsic pinhole camera parameters. The relation between them is expressed as Eq. 3:

$$p = K \cdot X \tag{3}$$

In our case, we consider the world origin to be located at the pinhole with no rotation and, therefore, the camera matrix can be expressed as Eq, 4, where f is the focal length (in pixels) and u_0, v_0 the pixel coordinates of the centre of the image:

$$K = \begin{bmatrix} f & 0 & \mathrm{u}_0 \\ 0 & f & \mathrm{v}_0 \\ 0 & 0 & 1 \end{bmatrix} \tag{4}$$

Knowing the estimated distance from the camera to the apple, z, it is possible to back project any point in the 2D image $[u, v]$ to obtain the 3D coordinates $[x, y, z]$ as Eq.5:

$$x = \frac{u-u_0}{f} z \qquad y = \frac{v-v_0}{f} z \qquad z = z \tag{5}$$

From Eq. 5, the fruit diameter in mm, D, can be computed from the estimation of the diameter of the fruit in pixels in the amodal segmentation:

$$D = \frac{z}{f} d \tag{6}$$

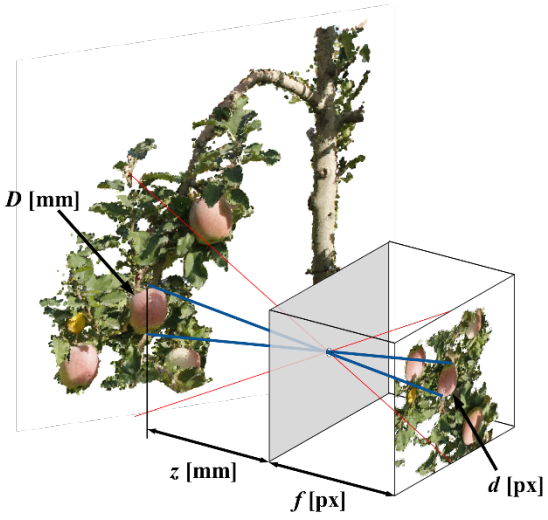


Fig. 6. Representation of image capturing using a pinhole camera. The 3D point cloud of an apple tree branch with four apples is imaged with a pinhole camera illustrated as a box. The parameters used to convert the diameter in pixels (d [px]) into millimetres (D [mm]) are represented: camera focal length (f [px]) and distance from the camera to the measured apple (z [mm]).

The fruit size estimation performance was evaluated in terms of the mean absolute error (MAE), the mean bias error (MBE), the mean absolute percentage error (MAPE), the root mean square error (RMSE) and the coefficient of determination (R^2). The MAE was computed by averaging the absolute differences between ground truth and estimated diameter. The MBE was computed similarly but without the absolute operator. Thus, the MBE sign indicates if the measurement overestimates or underestimates the ground truth. To measure the MAPE, first the percentage error committed in each estimation was calculated and then all percentage errors were averaged. The RMSE was calculated by applying the root square to the average square differences. Due to the square operation, this metric penalizes bigger errors. Finally, the R^2 was measured with the linear regression obtained between the ground truth and the estimated diameters. The predicted modal and amodal masks were not only used to estimate the diameter, but also to estimate fruit visibility (V) by computing the ratio of visible pixels (modal mask area) with respect to the total apple pixels (amodal mask area). This feature is used in **Section 3.2** to evaluate the fruit size estimation at different levels of visibility.

3. Results

3.1. Fruit detection and visibility estimation

Precision, Recall and F1-score at different detection confidence levels in the validation set were used to select the minimum detection confidence in order to maximise the F1-score, which was reported for a confidence of 0.2 with an F1-score of 0.86 (validation set). Based on these validation results, the detection performance in the test dataset was also evaluated setting the minimum confidence value of 0.2 (**Table 2**). Test results were similar to the ones obtained in the validation set (F1-score = 0.86 in both cases). The model performed slightly better for detection of ripe apples (AP_{BBCH85}=0.51; $F1_{BBCH85}$=0.87) than for detection of green apples (AP_{BBCH77}=0.44; $F1_{BBCH77}$=0.84), probably due to the larger colour contrast and the higher portion of BBCH85 apples in the training set.

Table 2. Fruit detection performance in the test dataset in terms of Precision (P), Recall (R), F1-score and average precision (AP) at different maturity stages: BBCH77, BBCH85 and BBCH77+BBCH85. A minimum confidence of 0.2 was set to consider a positive detection.

	BBCH77	BBCH85	BBCH77 + BBCH85
P	0.90	0.94	0.92
R	0.79	0.80	0.80
F1-score	0.84	0.87	0.86
AP	0.44	0.51	0.47

The predicted modal and amodal masks were used not only for fruit sizing, but also to estimate fruit visibility. **Fig. 7** evaluates the linear correlation between the ground truth visibility (computed using ground truth masks) and the predicted one (computed using fruit detections). Results show a high correlation between ground truth and predicted visibility (R^2 = 0.93), showing that the proposed deep learning model can be used as a reliable method to estimate fruit visibility.

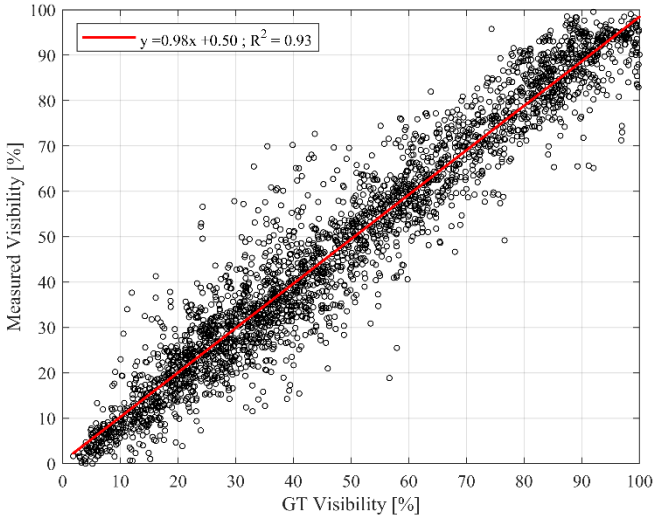


Fig. 7. Linear correlation between ground truth (GT) and estimated visibility.

3.2. Fruit size estimation

The fruit size estimation performance was evaluated with respect to the visibility percentage and the detection confidence score. **Fig. 8** represents the 3D plot of the MAE (Z axis) obtained in the validation set at different combinations of visibility (Y axis) and detection confidence scores (X axis). Results show that both features affected the performance of the measurement, with the percentage of visibility having more influence on the measured errors. When measuring the size of all detections irrespective of the visibility (Visibility > 0%), the detection confidence was useful to identify apples that were likely to be wrongly measured. In consequence, the MAE decreased with confidence, from MAE = 4.5 mm (confidence > 0) to MAE = 3.0 mm (confidence > 0.99). However, the visibility feature showed a higher influence on the MAE, achieving optimal results for visibilities higher than 80%, reporting an MAE of 2.2 mm in the validation set. Diameter errors increased for visibilities close to 100%. We attribute this effect to the low number of samples with visibilities of about 100%, which meant that the presence of outliers in this range of visibilities had a high influence on MAE results.

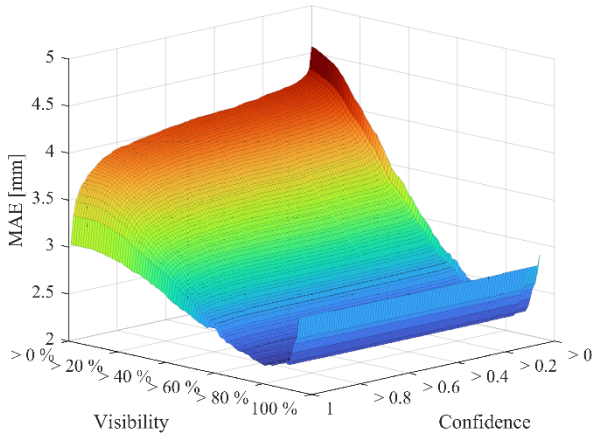


Fig. 8. Diameter estimation mean absolute error (MAE) obtained in the validation set measuring apples detected with a confidence or visibility higher than the specified values in the X and Y axis.

Fig. 9 represents fruit size distributions obtained at different visibility levels using the validation set. In general, even if all the visibility levels are considered (V > 0 %), the predicted distribution for both datasets fits well to the actual ground truth, which shows that the system is already robust with no need for further filtering. Nonetheless, higher visibility ensures optimal results, resulting in almost full overlaps between predicted and ground truth distributions when the minimum visibility threshold is increased. The most accurate overlap between ground truth and the automatically measured fruit size distribution was achieved when measuring fruits with a visibility higher than 60 %. This contrasts with the MAE evaluation, which achieved optimal results for visibilities higher than 80 %. The authors attribute the better distribution for visibilities higher than 60% to the fact that, although having a higher MAE, the number of fruits measured is higher and, in consequence, better fruit distributions are obtained when having more samples measured.

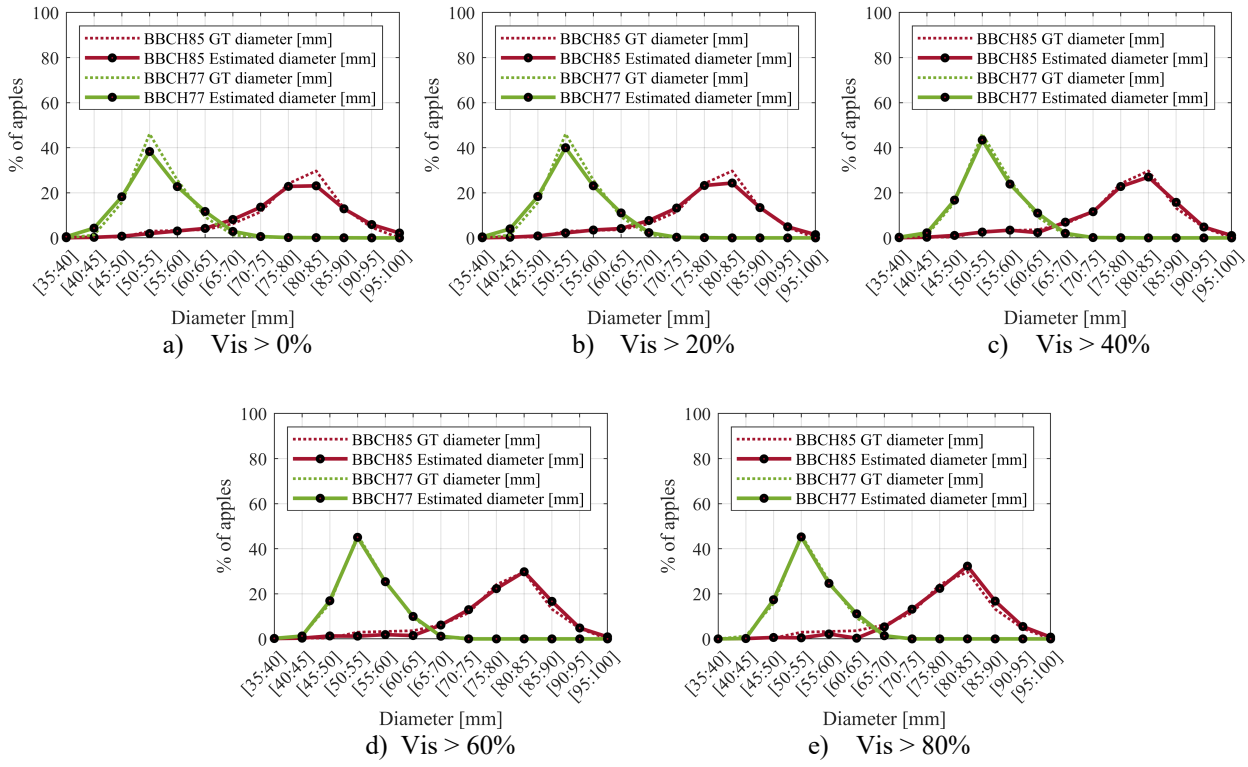


Fig. 9. Comparison between the ground truth diameter distribution (dotted line) and that automatically estimated in the validation dataset (solid line) at different visibility values: 0 % (a), 20 % (b), 40 % (c), 60 % (d) and 80 % (e). Green curves refer to the diameter distributions at BBCH77 growth stage and red curves refer to the diameter distributions at BBCH85 growth stage.

Table 3 presents the test results after automatically measuring the visibility and discarding the measurement of fruits detected with an estimated visibility lower than 60 %, which was the visibility percentage that obtained the best fruit size distributions in the validation set (**Fig. 9**). The proposed method reported an MAE of 2.93 mm. The negative MBE in all columns denotes that the model has a tendency to underestimate the final diameter. However, even the highest error (-0.54 mm) is small enough to be negligible for all practical purposes. The mean error was proportional to the fruit size, obtaining lower errors for smaller apples. In consequence, the percentage error reported similar results (MAPE ~ 4%) at both maturity stages. This linear error propagation is explained by **Eq. 6**, which applies a linear operation to convert the diameter from pixels to mm.

Table 3. Evaluation of the fruit diameter estimation in the test dataset in terms of MAE, MBE, MAPE and RMSE obtained at different growth stages (BBCH77 and BBCH85). Results were obtained by measuring apples automatically detected with a visibility higher than 60 %.

	BBCH77	BBCH85	BBCH77 + BBCH85
MAE (mm)	2.05	3.34	2.93
MBE (mm)	-0.54	-0.02	-0.19
MAPE (%)	3.79	4.27	4.19
RMSE (mm)	2.80	4.59	4.14

Fig. 10 shows that good correlations ($R^2 > 0.8$) between predicted and ground truth diameters were achieved for all levels of visibility, but even higher correlations were reported when the minimum visibility was increased. At very high visibility values the coefficient of determination started to decrease. This effect is a consequence of the small number of samples evaluated at high visibility intervals, because the presence of outlier estimations on which the diameter or visibility was wrongly predicted have more influence to the correlation when evaluating a small number of samples (as it is the case in high visibility intervals).

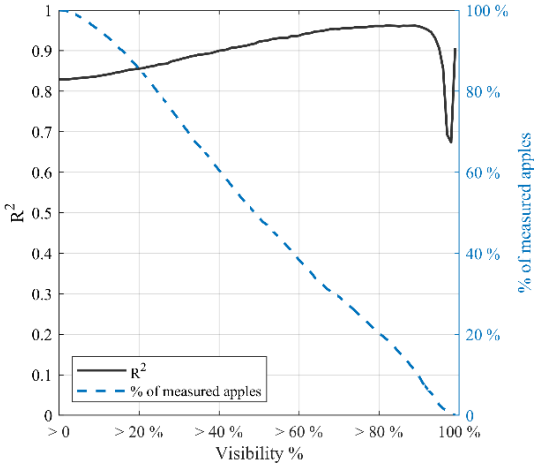
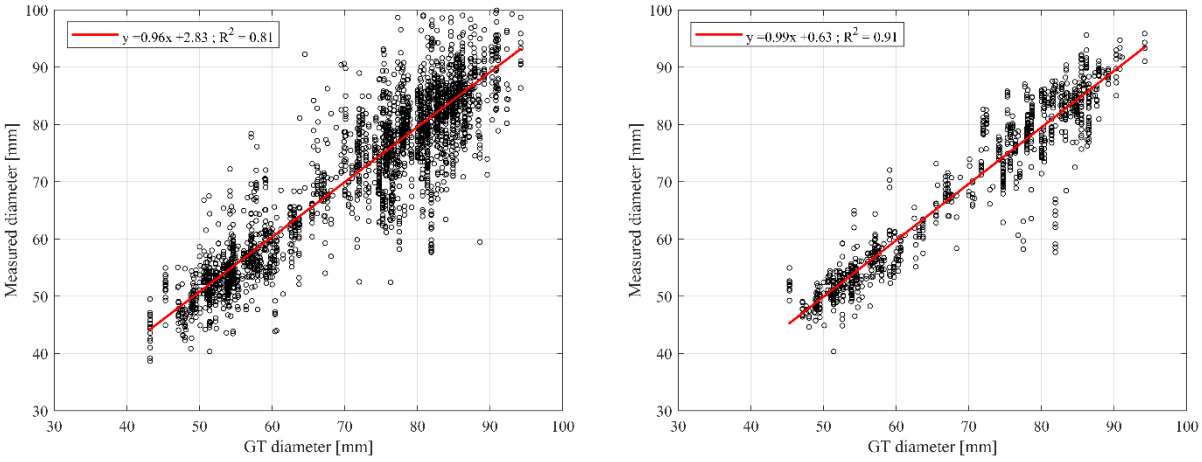


Fig. 10. Evolution of the coefficient of determination (R^2) (black line, left axis) and the percentage of measured apples (blue dashed line, right axis) depending on the visibility threshold used to discard the most occluded apples in the validation set.

Fig. 11 plots the linear correlation between the predicted diameter and the ground truth diameter for all apples detected in the test dataset (**Fig. 11a**) and for apples detected with a visibility higher than 60 % (**Fig. 11b**). The predicted samples fit better to the ground truth when limiting the measurement to apples detected with a visibility higher than 60 % ($R^2 = 0.91$), although acceptable correlation ($R^2 = 0.81$) was also obtained when measuring all detected samples. Close visual inspection shows two main clusters, which correspond

to the different datasets (ripe and not ripe apples).



(a) (b)

Fig. 11. (a) Scatter plot and linear correlation between the ground truth (GT) diameter and the automatically measured diameter of all apples detected in the test set. (b) Scatter plot and linear correlation between the ground truth (GT) diameter and the automatically measured diameter of apples detected in the test set with a visibility higher than 60 %.

In terms of computational speed, the average processing time per image in the test set was 0.830 s/img (frame rate of 1.205 img/s) using a NVIDIA GeForce GTX 1080 Ti GPU. This processing time was distributed as follows: 30.2 % of the time was allocated for fruit detection (0.251 s/img), 32.5 % for fruit sizing (0.270 s/img), and 37.3 % (0.309 s/img) for other processing tasks such as image reading, CPU and GPU data transferring and saving results.

For a qualitative evaluation, **Fig. 12** presents fruit detection and size estimation results in image examples from the test set. Modal masks are solid coloured while amodal masks are coloured with a certain amount of transparency. The fruit visibility was automatically estimated based on modal and amodal detections, and only those fruits presenting a visibility higher than 60 % were measured. Estimated diameters are written in white, while the ground truth is written above in green. The images included in this figure were selected in order to show when the method succeeds and when it fails: the first row (images a and b) contains the two highest scoring images (MAE < 1 mm), the second (images c and d) presents two intermediate scoring images (MAE of 2.2 mm) and the final row (images e and f) the two worst (MAE > 5mm). In all cases almost all the apples were detected, but the highest diameter errors were obtained for

those detections with errors in the modal and amodal mask shapes, or apple clusters that were detected with a unique detection.

Fig. 12. Fruit detection and sizing results. Predicted diameters are provided for detections with an estimated visibility higher than 60 %. Fruit size predictions are written in white and the ground truth in green. First row (a, b) contains examples of the best size estimation result, the second row (c, d) shows two intermediate scoring images and the third row (e, f) shows the two images with the worst estimations.

3.3. Case study: testing the trained model in a different apple variety and for different growth stages

To evaluate the robustness of the method in a different scenario, the model trained with the Fuji dataset was used to monitor the fruit growth of 5 Elstar trees at different growth stages. The apple mean diameter per tree was estimated on different dates using the presented method, and the results were compared to the measures carried out by the farmer.

Table 4 presents the differences between the manual estimation, carried out by the farmer, and our predictive method when measuring the fruits from the east, west and both (east and west) sides of the row of trees. Although the neural network was not trained with the same apple variety as the one used in this case study, the model generalised well and was able to estimate the mean diameter with an MAE of 3.09 mm and 5.23 mm from the west and east row side, respectively, resulting in an average MAE of 4.17mm. This error is 1.24 mm higher than the one obtained in the Fuji dataset (**Section 3.2**). The authors attribute this difference to three main reasons: 1) the network was not trained on images with Elstar apples, and consequently the detection performance is less accurate; 2) the Elstar trees had a denser foliage, which increases the chance of fruit occlusions; 3) part of the error was committed with the manual measurement, due to the small number of samples that were manually measured with respect to the total amount of fruits in the trees. In terms of bias error, the MBE was negative, confirming that the method tends to underestimate the actual fruit diameter. The authors attribute this underestimation of the diameter to the fact that the ground truth was obtained by measuring the major axis, whereas the method estimates the average apple diameter. Finally, in terms of R^2, the results showed a similar performance to those for the Fuji dataset, which shows that, despite a higher error in terms of MAE, there is still a high correlation between the manual measurements and the estimated predictions.

This case study was processed in a computing server equipped with an NVIDIA GeForce GTX 1080 Ti GPU. The total processing time per image was 0.865 s/img (processing frame rate of 1.156 img/s), of which 37.8 % of this time (0.327 s/img) was allocated for fruit detection, modal and amodal segmentation tasks, 26.7 % (0.231 s/img) was required for fruit sizing and occlusion estimation, and the other 35.5 % (0.307 s/img) was required for other image processing tasks such as RGB and depth images reading, data transfer

between CPU and GPU and saving the results. These processing times are similar to those reported in **Section 3.2** for Fuji apple detection and sizing.

Table 4. MAE, MBE, MAPE, RMSE, R^2 obtained in the case study dataset when measuring apple mean diameter per tree using images from the west, east, and west+east sides of the row of trees.

	East	West	East + West
MAE (mm)	5.23	3.09	4.17
MBE (mm)	-4.05	-2.73	-3.50
MAPE (%)	9.55	5.65	7.61
RMSE (mm)	6.27	4.48	5.15
R^2 (mm)	0.86	0.92	0.91

4. Discussion

Regarding the fruit detection task, the results showed a similar performance to other state-of-the-art works based on deep learning methods which reported F1-score results between 0.73 and 0.97 (Koirala et al., 2019). Fruit detection results were robust at different growth stages, although slightly better for ripe apples (F1-score =0.87) than for green apples (F1-score=0.84). The authors attribute the better detection of ripe fruits to the fact that they are bigger and with a different colour to the background (green leaves), which makes ripe fruits easy to differentiate. Tian et al. (2019) also experienced this behaviour, obtaining an F1-score of 0.78 and 0.81 when detecting young and ripe apples, respectively.

The main contribution and novelty of this work is the use of amodal instance segmentation to predict the apple shapes (masks) even in the presence of occlusions. The results show that the CNN was able to robustly estimate the actual shape of fruits that were partially occluded by other elements such as leaves, trunks or branches, or that were placed at the image edges (**Fig. 12**). Amodal masks were used to estimate fruit diameter, giving an MAE error lower than 4.5 mm in the validation set irrespective of fruit visibility and the detection confidence. In addition, the modification of the Mask R-CNN architecture to simultaneously predict modal and amodal masks allowed the method to estimate the percentage of occlusion of each detected apple, showing a coefficient of determination of $R^2 = 0.93$ between the predicted and the ground truth visibility. These results are comparable with those obtained in Gené-Mola et al. (2021)

using shape fitting methods in 3D point clouds. However, the method presented in the present paper has the advantage of being designed for 2D images, which allow higher data acquisition and inference speeds.

In previous research, Wang et al. (2017) demonstrated that fruits from the inner and outer part of the canopy had similar mass. This fact suggests that, to measure the fruit size distribution and the average size it is sufficient to measure a representative sample of fruits, instead of measuring all fruits on the tree. Thus, to select the best candidate detections to be measured two approaches were tested: 1) only measure the fruits that were detected with higher detection confidence; 2) only measure the fruits that were detected with higher visibility. When measuring the fruits detected with higher detection confidence (confidence > 0.99) the MAE error decreased to 3.0 mm. Better results were reported when limiting the measurement to fruits detected with higher visibility (V > 80 %), which gave MAE errors of 2.2 mm in the validation set. While the best MAE results were reported when limiting the measurement to the most visible fruits (V > 80 %), the best fruit distribution estimations were achieved for visibility percentages higher than 60 % (**Fig. 9**). This happened because the small number of fruits with visibilities higher than 80 % was not as representative as the number of fruits visible with a percentage higher than 60%.

Validation results allowed us to identify the optimal parameters to be used. These parameters were applied to evaluate the method using the test dataset, which showed an MAE of 2.93 mm measuring fruit diameters. This is considered an accurate result compared with other state-of-the-art results that measured fruit size in similar conditions and achieved MAE values of between 3.7 mm and 12.4 mm (Gené-Mola et al., 2021; Rakun et al., 2019; Tsoulias et al., 2020). Other works from the literature have also reported similar performances, or even better, but were studied under different conditions. For instance, Wang et al. (2020) reported an MAE error of 0.9 mm, but their study was limited to fully visible fruits that were manually selected. Grilli et al. (2021) obtained an RMSE of < 1 mm and 4 mm when measuring fruits of a synthetic and a laboratory dataset, respectively, but they did not provide results of their fruit sizing method evaluated in commercial orchard conditions.

Finally, the proposed method was tested in a case study to measure the mean apple diameters of different trees at different growth stages. Although this case study was carried out in a different field and with a different apple variety to the one used to train the CNN, the network was able to generalise and successfully detect and measure Elstar apples at different growth stages. An MAE of 4.17 mm and an R^2 of 0.91 were

obtained between the mean diameters measured by the farmer and those automatically measured with our method. The results show better predictions measuring fruits with images acquired from the west side of the row of trees than from the east side. The authors attribute this difference to the fact that fruits from the west side were more visible than fruits grown on the east side. Future works should analyse if this behaviour is general for other orchards or it is a characteristic of the field where the measurements were carried out.

The depth images used in this work were obtained with SfM and MVS techniques, which limits the data acquisition efficiency due to the number of images required to generate high quality depth maps and the computational cost of SfM and MVS algorithms. The acquisition of data with commercial RGB-D sensors will facilitate the data acquisition, but further analysis should be carried out to evaluate the performance of the present method with RGB-D sensors-based data, which provide depth images with less resolution and accuracy than the ones used in the present work.

In terms of inference speed, the CNN had an inference speed of 0.251 s/img (processed with an NVIDIA GeForce GTX 1080 Ti GPU), while the sizing task was performed in 0.270 s/img. These results show that, even with the addition of the amodal head to the Mask R-CNN architecture (**Fig. 5**), the CNN remains computationally efficient, and the method could be deployed in commercial orchards for fruit detection, robotic fruit picking and sizing tasks.

5. Conclusions

This work presented a novel approach to estimate the size of partially occluded apples by combining modal and amodal instance segmentation in RGB-D images. The results show that the method is efficient and robust for fruit detection and size estimation tasks at different growth stages and with different apple varieties (Fuji and Elstar). The fruit detection task performed slightly better for ripe apples than for green ones as ripe apples are larger and have a different colour to the background. In contrast, the fruit sizing task reported lower errors for green apples than for ripe ones as the sizing error is proportional to apple size due to the linear error propagation when converting the estimated diameter from pixels to mm. Besides the fruit detection and sizing tasks, the method was also effective for estimating the percentage of visibility of fruits, presenting a coefficient of determination of $R^2 = 0.93$ between ground truth and estimated visibility. From these results, an analysis of the fruit sizing performance at different visibility intervals showed that the

percentage of visibility is a good confidence parameter of diameter estimation. Thus, the estimated visibility allowed to discard highly occluded apples and estimate the fruit size by only considering those that are likely to be better measured (the most visible). In consequence, the diameter MAE error decreased from 4.5 mm to 2.93 mm when discarding the most occluded apples.

A case study carried out in a different orchard than the one used for training the CNN showed that the model generalizes well with another apple variety and other growth stages, obtaining a coefficient of determination of $R^2 = 0.91$ between the ground truth and the estimated diameter irrespective of the fruit variety and growth stage. In terms of processing time, the fruit detection and sizing task required an average of 0.521 s/image. From this, it is concluded that the method is computationally efficient to be applied in commercial orchards. The main limitation of this study is that it was carried out with depth images acquired with photogrammetry, which limit the data acquisition efficiency. Future works should evaluate the performance of the method with commercial RGB-D sensors, which would facilitate data collection. In addition, future works should consider extending this research to measuring the size of non-spherical fruits such as pears.

Acknowledgements

This work was partly funded by the Departament de Recerca i Universitats de la Generalitat de Catalunya (grant 2021 LLAV 00088), the Spanish Ministry of Science, Innovation and Universities (grants RTI2018-094222-B-I00 [PAgFRUIT project], PID2021-126648OB-I00 [PAgPROTECT project] and PID2020-117142GB-I00 [DeeLight project] by MCIN/AEI/10.13039/501100011033 and by "ERDF, a way of making Europe", by the European Union). The work of Jordi Gené Mola was supported by the Spanish Ministry of Universities through a Margarita Salas postdoctoral grant funded by the European Union - NextGenerationEU. We would also like to thank Nufri (especially Santiago Salamero and Oriol Morreres) for their support during data acquisition, and Pieter van Dalfsen and Dirk de Hoog from Wageningen University & Research for additional data collection used in the case study.

REFERENCES

Anderson, N.T., Underwood, J.P., Rahman, M.M., Robson, A., Walsh, K.B., 2019. Estimation of fruit

load in mango orchards: tree sampling considerations and use of machine vision and satellite imagery. Precis. Agric. 20, 823–839. https://doi.org/10.1007/s11119-018-9614-1

Blok, P.M., van Henten, E.J., van Evert, F.K., Kootstra, G., 2021. Image-based size estimation of broccoli heads under varying degrees of occlusion. Biosyst. Eng. 208, 213–233. https://doi.org/10.1016/J.BIOSYSTEMSENG.2021.06.001

Dutta, A., Zisserman, A., 2019. The VIA Annotation Software for Images, Audio and Video, in: Proceedings of the 27th ACM International Conference on Multimedia. ACM, New York, NY, USA. https://doi.org/10.1145/3343031.3350535

Gené-Mola, J., Llorens, J., Rosell-Polo, J.R., Gregorio, E., Arnó, J., Solanelles, F., Martínez-Casasnovas, J.A., Escolà, A., 2020. Assessing the performance of rgb-d sensors for 3d fruit crop canopy characterization under different operating and lighting conditions. Sensors (Switzerland) 20, 7072. https://doi.org/10.3390/s20247072

Gené-Mola, J., Sanz-Cortiella, R., Rosell-Polo, J.R., Escolà, A., Gregorio, E., 2021. In-field apple size estimation using photogrammetry-derived 3D point clouds: comparison of 4 different methods considering fruit occlusions. Comput. Electron. Agric. 188, 106343. https://doi.org/https://doi.org/10.1016/j.compag.2021.106343

Gongal, A., Karkee, M., Amatya, S., 2018. Apple fruit size estimation using a 3D machine vision system. Inf. Process. Agric. 5, 498–503. https://doi.org/10.1016/j.inpa.2018.06.002

Gregorio, E., Llorens, J., 2021. Sensing Crop Geometry and Structure, in: Kerry, R., Escola, A. (Eds.), Sensing Approaches for Precision Agriculture. Progress in Precision Agricuture. Springer, Cham. https://doi.org/10.1007/978-3-030-78431-7_3

Grilli, E., Battisti, R., Remondino, F., 2021. An advanced photogrammetric solution to measure apples. Remote Sens. 13. https://doi.org/10.3390/rs13193960

He, K., Gkioxari, G., Dollar, P., Girshick, R., 2017. Mask R-CNN. Proc. IEEE Int. Conf. Comput. Vis. 2017-Octob, 2980–2988. https://doi.org/10.1109/ICCV.2017.322

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. 770–778. https://doi.org/10.1109/CVPR.2016.90

Koirala, A., Walsh, K.B., Wang, Z., McCarthy, C., 2019. Deep learning – Method overview and review of use for fruit detection and yield estimation. Comput. Electron. Agric. 162, 219–234. https://doi.org/10.1016/j.compag.2019.04.017

Kootstra, G., Wang, X., Blok, P.M., Hemming, J., van Henten, E., 2021. Selective Harvesting Robotics: Current Research, Trends, and Future Directions. Curr. Robot. Reports 2, 95–104. https://doi.org/10.1007/s43154-020-00034-1

Li, K., Malik, J., 2016. Amodal Instance Segmentation, in: Leibe, B., Matas, J., Sebe, N., Welling, M. (Eds.), ECCV 2016. Lecture Notes in Computer Science, Vol 9906. Springer, Cham. Springer, pp. 677–693. https://doi.org/https://doi.org/10.1007/978-3-319-46475-6_42

Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L., 2014. Microsoft COCO: Common objects in context, in: European Conference on Computer Vision. pp. 740–755. https://doi.org/10.1007/978-3-319-10602-1_48

Longchamps, L., Tisseyre, B., Taylor, J., Sagoo, L., Momin, A., Fountas, S., Manfrini, L., Ampatzidis, Y., Schueller, J.K., Khosla, R., 2022. Yield sensing technologies for perennial and annual horticultural crops: a review. Precis. Agric. 23, 2407–2448. https://doi.org/10.1007/s11119-022-09906-2

Lu, S., Chen, W., Zhang, X., Karkee, M., 2022. Canopy-attention-YOLOv4-based immature/mature apple fruit detection on dense-foliage tree architectures for early crop load estimation. Comput. Electron. Agric. 193, 106696. https://doi.org/10.1016/J.COMPAG.2022.106696

Meier, U., 2001. Growth stages of mono- and dicotyledonous plants, BBCH Monograph. https://doi.org/10.5073/bbch0515

Moltó, E., Plá, F., Juste, F., 1992. Vision systems for the location of citrus fruit in a tree canopy. J. Agric. Eng. Res. 52, 101–110. https://doi.org/10.1016/0021-8634(92)80053-U

Qi, L., Jiang, L., Liu, S., Shen, X., Jia, J., 2019. Amodal instance segmentation with kins dataset. Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. 2019-June, 3009–3018. https://doi.org/10.1109/CVPR.2019.00313

Rakun, J., Stajnko, D., Zazula, D., 2019. Plant size estimation based on the construction of high-density corresponding points using image registration. Comput. Electron. Agric. 157, 288–304.

Robinson, T., Osborne, J., Fargione, M., 2008. Pruning, fertilization, chemical thinning and irrigation affect “Gala” apple fruit size and crop value. Acta Hortic. 772, 135–141. https://doi.org/10.17660/ActaHortic.2008.772.16

Santos, T.T., de Souza, L.L., dos Santos, A.A., Avila, S., 2020. Grape detection, segmentation, and tracking using deep neural networks and three-dimensional association. Comput. Electron. Agric. 170, 105247. https://doi.org/10.1016/j.compag.2020.105247

Slaughter, D.C., Harrell, R.C., 1987. Color vision in robotic fruit harvesting. Trans. ASAE 30 (4), 1144–1148.

Stajnko, D., Lakota, M., Hocevar, M., Hočevar, M., 2004. Estimation of number and diameter of apple fruits in an orchard during the growing season by thermal imaging. Comput. Electron. Agric. 42, 31–42. https://doi.org/10.1016/S0168-1699(03)00086-3

Tian, Y., Yang, G., Wang, Z., Wang, H., Li, E., Liang, Z., 2019. Apple detection during different growth stages in orchards using the improved YOLO-V3 model. Comput. Electron. Agric. 157, 417–426. https://doi.org/10.1016/j.compag.2019.01.012

Tijskens, L.M.M., Unuk, T., Okello, R.C.O., Wubs, A.M., Šuštar, V., Šumak, D., Schouten, R.E., 2016. From fruitlet to harvest: Modelling and predicting size and its distributions for tomato, apple and pepper fruit. Sci. Hortic. (Amsterdam). 204, 54–64. https://doi.org/10.1016/j.scienta.2016.03.036

Tsoulias, N., Paraforos, D.S., Xanthopoulos, G., Zude-Sasse, M., 2020. Apple shape detection based on geometric and radiometric features using a LiDAR laser scanner. Remote Sens. 12, 2481. https://doi.org/10.3390/RS12152481

Wang, D., He, D., 2022. Fusion of Mask RCNN and attention mechanism for instance segmentation of apples under complex background. Comput. Electron. Agric. 196, 106864. https://doi.org/10.1016/j.compag.2022.106864

Wang, D., Li, C., Song, H., Xiong, H., Liu, C., He, D., 2020. Deep learning approach for apple edge detection to remotely monitor apple growth in orchards. IEEE Access 8, 26911–26925. https://doi.org/10.1109/ACCESS.2020.2971524

Wang, Z., Koirala, A., Walsh, K., Anderson, N., Verma, B., 2018. In field fruit sizing using a smart phone application. Sensors (Switzerland) 18. https://doi.org/10.3390/S18103331

Wang, Z., Walsh, K.B., Verma, B.B., 2017. On-tree mango fruit size estimation using RGB-D images. Sensors 17, 2738. https://doi.org/10.3390/s17122738

Wu, Y., Kirillov, A., Massa, F., Lo, W.-Y., Girshick, R., 2019. Detectron2 [WWW Document]. GitHub Repos. URL https://github.com/facebookresearch/detectron2 (accessed 2.10.22).

Zhou, R., Damerow, L., Sun, Y., Blanke, M.M., 2012. Using colour features of cv. “Gala” apple fruits in an orchard in image processing to predict yield. Precis. Agric. 13, 568–580. https://doi.org/10.1007/s11119-012-9269-2