Available online at www.sciencedirect.com

ScienceDirect

journal homepage: www.elsevier.com/locate/issn/15375110

Research Paper

Image-based size estimation of broccoli heads under varying degrees of occlusion

Pieter M. Blok [a,*], Eldert J. van Henten [b], Frits K. van Evert [a], Gert Kootstra [b]

[a] Agrosystems Research, Wageningen University & Research, Wageningen, the Netherlands
[b] Farm Technology Group, Wageningen University & Research, Wageningen, the Netherlands

ARTICLE INFO

Article history:
Received 18 January 2021
Received in revised form
29 April 2021
Accepted 1 June 2021
Published online 18 June 2021

Keywords:
Size estimation
Occlusion
Deep-learning
Agriculture
Data set
Computer vision

The growth and the harvestability of a broccoli crop is monitored by the size of the broccoli head. This size estimation is currently done by humans, and this is inconsistent and expensive. The goal of our work was to develop a software algorithm that can estimate the size of field-grown broccoli heads based on RGB-Depth (RGB-D) images. For the algorithm to be successful, the problem of occlusion must be solved, which is the partial visibility of the broccoli head due to overlapping leaves. This partial visibility causes sizing errors. In this research, we studied the use of deep-learning algorithms to deal with occlusions. We specifically applied the Occlusion Region-based Convolutional Neural Network (ORCNN) that segmented both the visible and the amodal region of the broccoli head (which is the visible and the occluded region combined). We hypothesised that ORCNN, with its amodal segmentation, can improve the size estimation of occluded broccoli heads. The ORCNN sizing method was compared with a Mask R−CNN sizing method that only used the visible broccoli region to estimate the size. The sizing performance of both methods was evaluated on a test set of 487 broccoli images with systematic levels of leaf occlusion. With a mean sizing error of 6.4 mm, ORCNN outperformed Mask R−CNN, which had a mean sizing error of 10.7 mm. Furthermore, ORCNN had a significantly lower absolute sizing error on 161 heavily occluded broccoli heads with an occlusion rate between 50% and 90%. Our software and data set are available on https://git.wur.nl/blok012/sizecnn.

1. Introduction

The in-field estimation of the crop size is an important task in plant phenotyping, growth monitoring and harvesting. Currently, this crop size estimation is mainly done by humans, and this can be inconsistent and expensive. A promising alternative is a sensor system with a software algorithm that can autonomously estimate the size of a crop while it is still on the tree or plant. This system can potentially increase the accuracy and the frequency of the size estimates, while reducing labour costs.

Recent studies focused on the in-field size measurement of apple (Gongal et al., 2018), broccoli (Kusumam et al., 2017), citrus (Lin et al., 2019) and mango (Wang et al., 2017). These studies have two similarities. The first similarity is that all

* Corresponding author.
E-mail addresses: pieter.blok@wur.nl (P.M. Blok), eldert.vanhenten@wur.nl (E.J. van Henten), frits.vanevert@wur.nl (F.K. van Evert), gert.kootstra@wur.nl (G. Kootstra).
https://doi.org/10.1016/j.biosystemseng.2021.06.001

Nomenclature

Abbreviations

3D	three dimensional
AIoU	amodal intersection over union
CNN	convolutional neural network
COCO	common objects in context
Det	detection
Diam	diameter
Diff	difference
DPL	depth-pixel loss rate
Est	estimation
FN	false negative
FP	false positive
GT	ground-truth
IoU	intersection over union
MAD	median absolute error
MAE	mean absolute error
Mask R–CNN	mask region-based convolutional neural network
NMS	non-maximum suppression
OCR	occlusion rate
ORCNN	occlusion region-based convolutional neural network
P	precision
QR	quick response
R	recall
RGB	red, green and blue
RGB-D	red, green, blue and depth
RMSE	root mean squared error
ROI	region of interest
TP	true positive
VIoU	visible intersection over union

Symbols *(units)*

$\cap$	intersection (–)
$\cup$	union (–)
$\widehat{d}$	diameter estimate (mm)
τ_{CNN}	threshold on the confidence level (–)
τ_{IoU}	threshold on the intersection over union (–)
τ_{NMS}	threshold on the non-maximum suppression (–)
a	significance level of the pairwise Wilcoxon test (%)
$\widetilde{\epsilon}$	median diameter error (mm)
$\widetilde{\epsilon}_{mrcnn}$	median diameter error of Mask R-CNN (mm)
$\widetilde{\epsilon}_{orcnn}$	median diameter error of ORCNN (mm)
ϵ	diameter error (mm)
A_d	area of depth pixels of the broccoli head in the registered depth image (pixels)
A_t	area of the visible region of the non-occluded broccoli head (pixels)
A_v	area of the visible region of the occluded broccoli head (pixels)
c	confidence level on the object detection of the convolutional neural network (–)
d	ground-truth diameter (mm)
M_{gt}	area of the ground-truth mask (pixels)
M_p	area of the predicted mask (pixels)
p	p-value of the pairwise Wilcoxon test (–)
r	Pearson's correlation coefficient (–)

researchers used a camera-based system that generated RGB-Depth (RGB-D) images. An RGB-D image consists of a red, green and blue colour image (RGB) and a registered depth image, where each pixel contains the distance measurement between the image plane and the object. In three of the four studies, the colour image was used to detect the crop, while the depth image was used to estimate the crop's size. In Kusumam et al. (2017), the detection and the size estimation were done on a three dimensional (3D) image that was created from the depth image. The second similarity is that all researchers used feature-engineered software algorithms to process the RGB-D images. In Kusumam et al. (2017) and Lin et al. (2019), a machine-learning algorithm was used, but none of the researchers used deep-learning algorithms. Deep-learning algorithms currently provide state-of-the-art performance in the in-field fruit and crop detection (Blok et al., 2021; Ge et al., 2019; Kang & Chen, 2020; Nejati et al., 2020; Yu et al., 2019). In Kamilaris and Prenafeta-Boldú (2018) it was shown that deep-learning algorithms outperformed feature-engineered algorithms in all 22 agricultural case studies. In line with these findings, our research will focus on the crop size estimation with an RGB-D camera and a deep-learning algorithm.

A limitation of the crop-sizing studies of Gongal et al. (2018), Kusumam et al. (2017), Lin et al. (2019) and Wang et al. (2017), was that the algorithms were tested on crops that had no or minimal occlusion, meaning that the results only partially reflected the in-field sizing performance. Usually, an agricultural image scene is dense and cluttered, with many forms of crop occlusion. When there is crop occlusion, a (big) part of the crop is covered by other crop organs, surrounding plants or materials, making it harder for an algorithm to detect and size the crop. This challenge was also acknowledged by Zhang et al. (2020), who did a literature review on deep-learning algorithms that were tested in dense and cluttered agricultural image scenes. In Zhang et al. (2020), it was stated that occlusion is one of the biggest challenges for a deep-learning algorithm when analysing these complex image scenes. The goal of our work was to develop a deep-learning algorithm that can accurately estimate the size of crops even when they are occluded. In our research, we chose broccoli (*Brassica oleracea* var. *italica*) as our model crop, since broccoli heads can be heavily occluded by leaves and weeds.

With 3D software algorithms it is possible to detect and size the occluded broccoli heads. For example, a Frustum Pointnet algorithm (Qi et al., 2018) can be used to detect partially occluded 3D objects. A 3D shape-completion algorithm can be used to estimate the shape of occluded crops, similar to how Ge et al. (2020) estimated the shape of occluded strawberry fruits. However, 3D algorithms can also have limitations, such as a longer analysis time and a less optimised transfer-learning, due to the limited availability of 3D agricultural data sets. Another way to deal with occlusions, is to obtain multi-view images of the same object from multiple

cameras or camera positions. This multi-view imaging has proven its effectiveness in other occluded crop environments, such as sweet pepper (Barth et al., 2016; Lehnert et al., 2019) and cucumber (Boogaard et al., 2020), but a multi-view imaging also has its disadvantages, such as higher hardware costs and a longer image analysis time compared to an analysis on a single image. Therefore, we will investigate deep-learning methods that can estimate the size of occluded broccoli heads from a single RGB-D image of the scene.

Comparable to the algorithms of Gongal et al. (2018), Kusumam et al. (2017), Lin et al. (2019) and Wang et al. (2017), the sizing algorithm is expected to execute two different sub tasks. The first sub task is the image-based detection of the broccoli head. The second sub task is the size estimation of the broccoli head in the registered depth image, using the detection output from the first sub task. The image-based broccoli detection can be accomplished with a special group of deep-learning algorithms: convolutional neural networks (CNN's). Appropriate CNN's for this task are object-detection algorithms and instance-segmentation algorithms. Object-detection algorithms, like Faster R−CNN (Ren et al., 2017) or YOLOv4 (Bochkovskiy et al., 2020), can detect the broccoli head in an RGB image with a rectangular bounding box, similar to how Bender et al. (2020) detected broccoli plants. Other object-detection algorithms were specifically designed to detect circles (YOLO-Tomato (Liu et al., 2020)) or ellipses (BubCNN (Haas et al., 2020)), which might better match the shape of the broccoli head. However, with the bounding box, circle and ellipse detections it is impossible to specify whether the pixels that are inside the detected shape belong to the broccoli head or to objects that occlude the broccoli. Due to this lack of pixel differentiation, an object-detection algorithm requires an additional filtering algorithm to remove the pixels that do not belong to the broccoli head, because these pixels cannot be used for the size estimation in the registered depth image. An alternative approach is to use an instance-segmentation algorithm, like Mask R−CNN (He et al., 2017) or YOLACT++ (Bolya et al., 2020). An instance-segmentation algorithm can segment the broccoli head pixels inside the bounding box. With this additional pixel segmentation, the sizing algorithm will be less dependent on an additional filtering algorithm, making an instance-segmentation algorithm a more appropriate algorithm for an autonomous broccoli sizing system.

When an instance-segmentation algorithm is used to segment an occluded broccoli head, then the pixel segmentation would represent only the visible region of the broccoli head. When this partially completed segmentation is used for the size estimation, there is a chance that the actual size is underestimated. One way to alleviate this problem, is to extend the instance-segmentation algorithm with an additional shape-completion algorithm, like Ge et al. (2020) did, to approximate the bigger shape from the visible region of the broccoli head. An alternative approach is to estimate the bigger shape of the occluded broccoli head with an instance-segmentation algorithm that segments the combined visible and occluded part of the broccoli head. This segmentation is called *amodal segmentation* (Zhu et al., 2017), and this might better reveal the actual shape of the occluded broccoli head. However, with an amodal segmentation some of the segmented pixels would belong to objects that occlude the broccoli head. Obviously, these pixels need to be removed with an additional filtering algorithm to assure an accurate size estimation. In summary, with an instance-segmentation algorithm that either segments the visible broccoli region or the amodal broccoli region, there is a need of an additional shape-completion or filtering algorithm to estimate the size of the broccoli head. The problem is that these algorithms might cause additional sizing errors.

A possible solution is to use an instance-segmentation algorithm that can generate two segmentations: one on the amodal region of the broccoli head and one on the visible region of the broccoli head. The amodal segmentation can be used to estimate the bigger shape of the occluded broccoli head, whereas the visible segmentation can be used to extract the depth values of the broccoli head that are needed to estimate its real-world size. This dual segmentation makes the sizing algorithm less dependent on an additional shape-completion or filtering algorithm, which might improve the size estimate.

Occlusion Region-based Convolutional Neural Network (ORCNN) (Follmann et al., 2018) is an instance-segmentation algorithm that can generate this dual segmentation. ORCNN is an extended Mask R−CNN network with multiple mask head branches, of which one can be trained to segment the visible broccoli pixels and the another one can be trained to segment the amodal broccoli pixels. Because ORCNN generates a pixel segmentation for both the visible and the amodal region, it can be used to predict all kinds of crop shapes.

In this paper, we hypothesised that the size estimation of occluded broccoli heads can be improved when using an algorithm that can segment both the visible and the amodal region of the broccoli head. The objective of our study was to test this hypothesis by comparing the sizing performance of ORCNN with a Mask R−CNN sizing method that was only based on a single segmentation of the visible broccoli pixels. Our research was conducted on a data set of 2560 broccoli images with systematic levels of occlusion. The main contribution of our research is a novel size estimation method that uses a dual image segmentation to better deal with crop occlusions. The secondary contribution is the release of the crop-sizing software and a data set of broccoli images with systematic levels of occlusion.

2. Materials and methods

2.1. *Image data set*

This paragraph highlights how the RGB-D images were acquired in the broccoli fields (section 2.1.1) and how the acquired images were pre-processed and annotated (section 2.1.2). Then, it is explained how the broccoli occlusion rate was calculated in the RGB image (section 2.1.3) and in the registered depth image (section 2.1.4). Finally, it is described how the annotated images were aggregated for CNN training and testing (section 2.1.5).

2.1.1. *Image acquisition*

To the best of our knowledge, there are currently two online-available data sets with RGB-D images of field-grown broccoli (Bender et al., 2019; Kusumam et al., 2016). Unfortunately, the broccoli images of both data sets had no or minor occlusions. Also, Kusumam et al. (2016) did not publish the ground-truth size measurements. Therefore we decided not to use these images, and to acquire two data sets of broccoli images with systematic levels of occlusion. The two data sets were acquired with two different cameras on two broccoli fields that were located in the United States of America (USA) and in the Netherlands. On these fields, two different broccoli cultivars were grown in two different growing seasons. The variations in crop conditions and imaging hardware resulted in a diverse data set for the training and the testing of the algorithms.

The first data set was acquired in 2018 on a broccoli field in Santa Maria (USA). On this field, the broccoli plants of the cultivar Avenger were grown on beds with two crop rows that were 0.31 m apart. The intra-row spacing was 0.20 m. Before the image acquisition, we selected two rows in the broccoli field that were grown on two different beds. In these two rows, we randomly selected 122 occluded broccoli heads with a diameter between 85 and 228 mm (the average diameter was 156 mm). The selected broccoli heads were tagged with a Quick Response (QR) code for visual recognition. Then, the RGB-D images were acquired with a prototype harvesting robot. This robot was equipped with an image acquisition system that acquired top-view RGB-D images of the broccoli crop, see Fig. 1a. The image acquisition system was constructed as an enclosed box for uniform illumination. The acquisition system was equipped with one RGB colour camera (IDS UI-5280FA-C-HQ) with a 8 mm lens (Fujifilm HF8XA-5M), one monochrome stereo-vision camera (IDS Ensenso N35) and 21 light emitting diode (LED) strips (OSRAM VFP2400S-G3-865-03) for artificial illumination, see Fig. 1b. The colour camera was positioned at the centre of the stereo-vision camera, but with a 52 mm vertical offset, Fig. 1c. The distance between the two cameras and the broccoli heads was approximately 0.6 m. At this distance, the camera's field-of-view was 0.62 m (width) by 0.52 m (height). The two cameras were levelled before the image acquisition with a bubble level instrument, to ensure the horizontal and vertical alignment between the cameras and the broccoli heads.

The images of the colour and the stereo-vision camera were simultaneously acquired with a hardware trigger from an electronic encoder wheel that was attached to the front wheel of the robot. This encoder generated a hardware trigger to the cameras for each 0.15 m (+/− 0.01 m error) of relative displacement of the robot. The RGB image (produced by the colour camera) and the depth image (produced by the stereo-vision camera) were registered, creating one RGB-D image of 1280 × 1024 pixels for each hardware trigger. The robot was driven with a constant speed of approximately 0.14 m s^{-1} over the two rows to acquire the images of the 122 selected broccoli heads with its natural occlusion (leaves). In total 947 RGB-D images were captured (four to ten frames per broccoli head). Because the robot moved over the crop, a various range of natural occlusions occurred in the different frames due to changes in camera perspective, see three examples in Fig. 2. After the image acquisition, the leaves that occluded the broccoli heads were removed and the robot was driven again over the same rows to acquire four to ten frames from each broccoli head without any occlusion. Finally, the diameters of the broccoli heads were measured with a circular ruler, Fig. 3b.

The second data set was acquired in 2020 on a broccoli field in Sexbierum (The Netherlands). On this field, the broccoli plants of the cultivar Ironman were grown in single rows that were 0.75 m apart. The intra-row spacing was 0.33 m. The broccoli images were acquired with an Intel Realsense D435 camera that was mounted on a metal frame to acquire top-view RGB-D images of the broccoli crop, see Fig. 3a. The distance between the Realsense camera and the broccoli heads was approximately 0.6 m. At this distance, the camera's field-of-view was 0.79 m (width) by 0.58 m (height). The Realsense camera was levelled with a bubble level instrument before each image acquisition of a broccoli head, to ensure the horizontal and vertical alignment between the camera and the broccoli heads. The RGB-D images were acquired in daylight without artificial illumination. Diffuse light conditions were created with an umbrella. The RGB image and the depth image from the Realsense camera were registered, creating one RGB-D image of 1280 × 720 pixels for each software trigger.

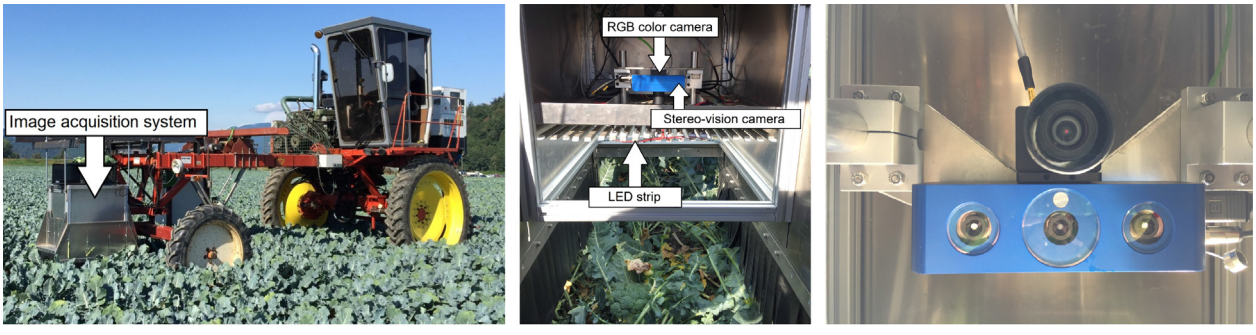


Fig. 1 – (a) Overview of the image acquisition system that was attached to a prototype harvesting robot to acquire top-view RGB-D images of broccoli heads in a field in Santa Maria (USA). (b) The image acquisition system consisted of one RGB colour camera, one monochrome stereo-vision camera and 21 LED strips for artificial illumination. (c) There was a vertical offset of 52 mm between the RGB-camera (upper black camera) and the stereo-vision camera (lower blue camera).

(a) (b) (c)

Fig. 2 – (a) The prototype harvesting robot drove over the broccoli crop to acquire multiple frames from the same broccoli head (which were tagged with a QR code). In this example, the first frame was captured when the broccoli head was in the top of the image. In this frame, the broccoli head was subject to heavy leaf occlusion. (b) Another frame was taken when the broccoli head was in the centre of the image. In this frame, the broccoli head was subject to moderate occlusion. (c) In the last frame, the tagged broccoli head was in the bottom of the image and had a low level of occlusion, because of the changed position of the camera.

(a) (b)

Fig. 3 – (a) The Intel Realsense D435 camera was mounted on a metal frame to acquire top-view RGB-D images of broccoli heads in a field in Sexbierum (The Netherlands). (b) After the image acquisition, the diameter of the broccoli head was measured with a circular ruler.

On the field, 250 occluded broccoli heads were randomly selected from multiple crop rows. The selected broccoli heads had a diameter between 68 and 239 mm (the average diameter was 137 mm). First, one frame of the broccoli head was acquired with its natural occlusion (leaves and weeds), see Fig. 4a. Then, additional frames were acquired of the same broccoli head with different occlusions. The different occlusions were created by cutting a leaf from a neighbouring plant and then placing this leaf over the broccoli head to create a human-made, yet natural-looking occlusion, Fig. 4b. This was repeated for five to ten frames per broccoli head. Afterwards, all leaves were removed from the broccoli plant and a last image frame was acquired without any occlusion, Fig. 4c. In total, 1613 RGB-D images were captured on this broccoli field. After the image acquisition, the diameters of the broccoli heads were measured with the same circular ruler that was used to measure the broccoli heads in the USA, Fig. 3b.

2.1.2. *Image pre-processing and annotation*

All 2560 RGB-D images from the two data sets were re-scaled and zero-padded to a resolution of 1280 × 1280 pixels, see examples in Figures 2 and 4. This zero-padding allowed us to extrapolate the annotations into the black-coloured regions in case the broccoli head was only partially in the field-of-view of the camera, see an example in Fig. 5. The image annotation was done with the LabelMe software (version 4.5.6) (Wada, 2016). First, the image frames with no occlusion were annotated. In these frames, each broccoli head was annotated by two masks: a polygonal mask for the visible broccoli region and a circular mask for the amodal broccoli region. The

Fig. 4 – For each broccoli head in the second data set, five to ten frames were acquired with different occlusions. (a) The first frame was acquired from the broccoli head with its natural occlusion. (b) Then, a randomly clipped leaf from a neighbouring broccoli plant was positioned above the broccoli head to create a human-made, yet natural-looking occlusion. This process was repeated for five to ten different leaf positions to create different occlusions. In this example, the clipped leaf occluded the bottom part of the broccoli head. (c) The last frame was acquired when all occluding leaves had been removed.

Fig. 5 – The image annotation procedure involved the following steps: (a) First, the non-occluded frame was annotated. For each broccoli head, a circular mask was drawn for the amodal region (red circle) and a polygonal mask was drawn for the visible region (blue polygon). The amodal mask of the partially captured broccoli head in the bottom of the image, was drawn into the zero-padded region of the image (black-coloured region). This amodal annotation was done by means of the best guess of the image annotator. (b) Then, all amodal masks were copied to the frames of the same broccoli head with occlusion. The visible masks were independently drawn, because they could not be copied.

circular mask was drawn along the circumference of the broccoli head, Fig. 5a. We chose for a circular mask, because this corresponded to the shape of the circular ruler, which was used to obtain the ground-truth, Fig. 3b. Then, the circular amodal mask was copied to the frames of the same broccoli plant with occlusion, Fig. 5b. This procedure allowed us to precisely annotate the occluded broccoli head with the amodal mask that was drawn in the non-occluded frame. The position of the amodal mask was then checked and corrected by another image annotator when necessary. Finally, the polygonal masks of the visible broccoli regions were annotated for all images. Examples of these visible broccoli annotations are the blue polygons in Fig. 5. The other broccoli heads that were present in the image but not tagged and measured in the field, were also annotated by means of the best guess of the image annotator. Examples of these annotations are visualised in the bottom of the image of Figure 5a and b. These annotations were used to train the CNN's and to calculate the detection metrics. The software procedures of the annotation process can be found on our git repository.

2.1.3. *Calculation of the occlusion rate in the RGB image*

In the RGB image, the pixel area of the visible region of the occluded broccoli head, A_v, was divided with the pixel area of the visible region of the same broccoli head in the non-occluded frame, A_t. This division resulted the occlusion rate (OCR) in the RGB image (Equation (1)). An example of the OCR calculation is the division of the area of the blue polygons in Fig. 5b with the area of the blue polygons in Fig. 5a.

$$OCR = 1 - \frac{A_v}{A_t} \quad (1)$$

The occlusion rates were quantified for the broccoli images that had a natural leaf occlusion and for the broccoli images that had a human-made leaf occlusion. In total, 1197 of the 2560 broccoli images had a natural occlusion. These were the images from the first data set and the frames of the second data set that had a natural occlusion, see an example in Fig. 4a. The broccoli heads were on average 25.9% occluded by leaves and weeds in its natural situation (the standard deviation was 23.6%). The remaining 1363 broccoli images had a human-made leaf occlusion, see two examples in Figure 4b and c. In this human-made situation, the broccoli heads were on average 44.9% occluded with a standard deviation of 30.3%. Figure 6 shows the distribution of the occlusion rate for the two situations.

2.1.4. *Calculation of the pixel loss in the depth image*

Both of the used depth cameras were stereo-vision cameras. These cameras use a left and a right monochrome camera to produce a depth image. Due to the different perspective of the two cameras, some parts of the scene can only be viewed by one camera due to occlusion. For these image parts, the depth cannot be calculated. In our data sets, the depth-pixel loss rate, DPL, was quantified by comparing the area of the broccoli pixels in the RGB image, A_v, with the area of the depth pixels that were present in the same broccoli region in the registered depth image, A_d. The depth-pixel loss rate was calculated with Equation (2). A visual example of the depth-pixel loss rate is the difference between the blue region in Fig. 7a and the green region in Fig. 7b.

$$DPL = 1 - \frac{A_d}{A_v} \quad (2)$$

In all 2560 depth images, the average depth-pixel loss rate was 20.2% (the standard deviation was 19.9%). In the 947 depth images of the Ensenso N35, the average depth-pixel loss rate was 17.5% (the standard deviation was 17.7%). In the 1613 depth images of the Realsense D435, the average depth-pixel loss was 21.7% (the standard deviation was 20.9%). Figure 8 shows the distribution of the depth-pixel loss for the two depth cameras.

2.1.5. *Training, validation, and test set*

The 2560 annotated images from 372 unique broccoli plants were divided into a training set, a validation set and a test set. First, all images of a unique broccoli plant were placed into separate groups. These groups of images were then placed into either the training set, the validation set or the test set, based on a stratified sampling criterion using the measured diameter of the broccoli head. This stratified sampling ensured that all images of the same broccoli plant were placed in either the training set, validation or test set, and that a various range of broccoli diameters would appear in each of the three sets. The 372 unique broccoli plants were split into a training set of 222 plants (60%), a validation set of 75 plants (20%) and a test set of 75 plants (20%). The images that belonged to the unique plants were then put into the three sets, resulting a training set of 1569 images (61.3%), a validation set of 504 images (19.7%) and a test set of 487 images (19.0%).

2.2. *Size estimation of the broccoli heads*

The broccoli size estimation involved two sub tasks: the segmentation of the broccoli head in the RGB image (which will be described in paragraph 2.2.1), and the diameter estimation of the broccoli head in the registered depth image (which will be described in paragraph 2.2.2).

2.2.1. *Broccoli head segmentation with Mask R−CNN and ORCNN*

ORCNN was compared with a conventional Mask R−CNN algorithm to evaluate the effect of the additional amodal segmentation on the sizing performance. In this paragraph, the technical details of the two CNN's are described by means of the network architecture (section 2.2.1.1), the used software

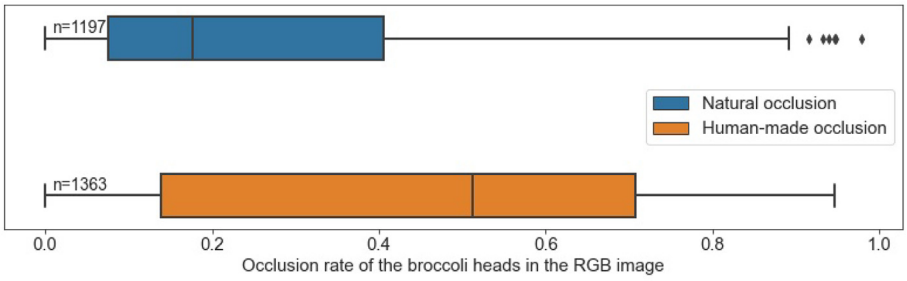


Fig. 6 – A box-and-whisker plot showing the distribution of the broccoli occlusion rates that were calculated in the 2560 RGB images. *n* is the number of RGB images that were used to respectively calculate the natural occlusion rate and the human-made occlusion rate. The line within the box indicates the median of the distribution. 50% of the data is present within the ends of the box, which represent the 25th percentile (first quartile) and the 75th percentile (third quartile). The whiskers indicate the variability outside the first and third quartiles, whereas the dots indicate the outliers.

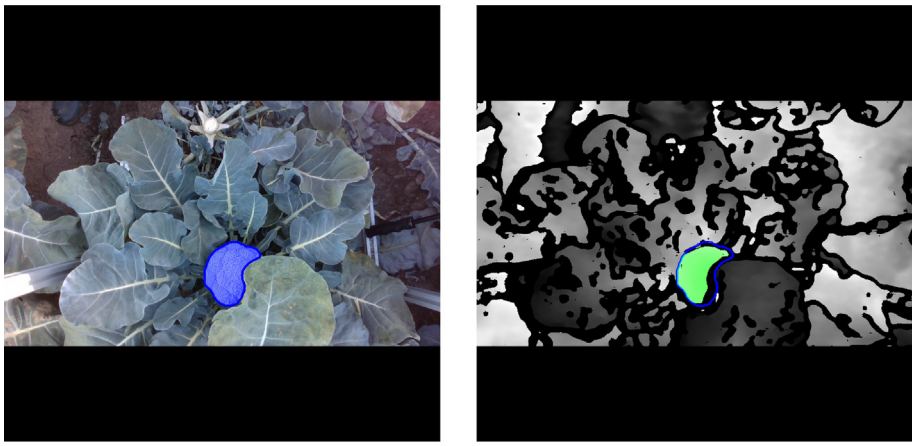


Fig. 7 – (a) The blue polygon visualises the visible mask annotation in the RGB image. (b) The green polygon visualises the pixels with a depth value after copying the visible mask annotation (blue contour) into the registered depth image. The black pixels inside the blue contour are the pixels of the broccoli head that had no depth value. The grey-scale of the depth image was based on the depth values: the pixels with a lighter colour are further from the camera.

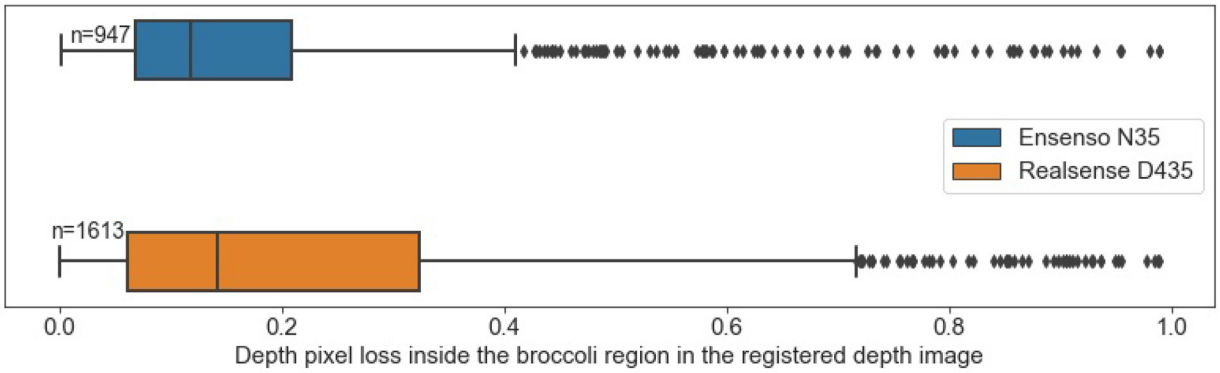


Fig. 8 – The box-and-whisker plots show the distribution of the depth-pixel loss rate for the two depth cameras that were used in our experiments. An explanation of the box-and-whisker plot can be found in Fig. 6. *n* is the number of depth images per camera.

and hardware (section 2.2.1.2), the training procedure (section 2.2.1.3) and the image inference procedure (section 2.2.1.4).

2.2.1.1. *Network architectures*. Mask R–CNN (He et al., 2017) is a neural network that consists of multiple branches. First, there is a backbone, which is a neural network that extracts feature maps at various resolution scales from an image with a feature pyramid network. In our research, the ResNeXt-101 (32 × 8d) (Xie et al., 2017) residual network was used as backbone. After the backbone, there is a region proposal network that proposes regions of interest (ROI) of possible distinct objects from the feature maps. To avoid duplicate ROIs for the same object, non-maximum suppression (NMS) is used that discards the ROIs that overlap with a more confident ROI. Then, the remaining ROIs are realigned with the ROI align layer and transformed into fix-sized feature maps. These feature maps are further processed in two parallel branches in the so-called network head. The first head branch has two fully connected layers, of which one performs object classification and the other one bounding box detection. The second branch, which is the mask head branch, has four 3 × 3 convolutional layers that segment the object pixels inside the bounding box, yielding the mask.

Except for the mask head branch, ORCNN (Follmann et al., 2018) has the same architecture as Mask R–CNN (He et al., 2017). With ORCNN, the object classification and the bounding box detection are trained on the ground-truth class and box of the amodal instance, because this is by definition the largest region. Then, all segmentations are done inside the same amodal bounding box. ORCNN's original network architecture has three mask head branches: one for the visible mask, one for the amodal mask and one for the occlusion mask (which is the difference between the amodal and the visible mask). In our research, the occlusion mask branch was removed, because this mask head was not needed for our sizing application. The visible and the amodal mask head branch that remained, were both based on the mask head branch of Mask R–CNN, indicating that they both had four 3 × 3 convolutional layers, refer to the schematic representation of the network architecture in Fig. 9.

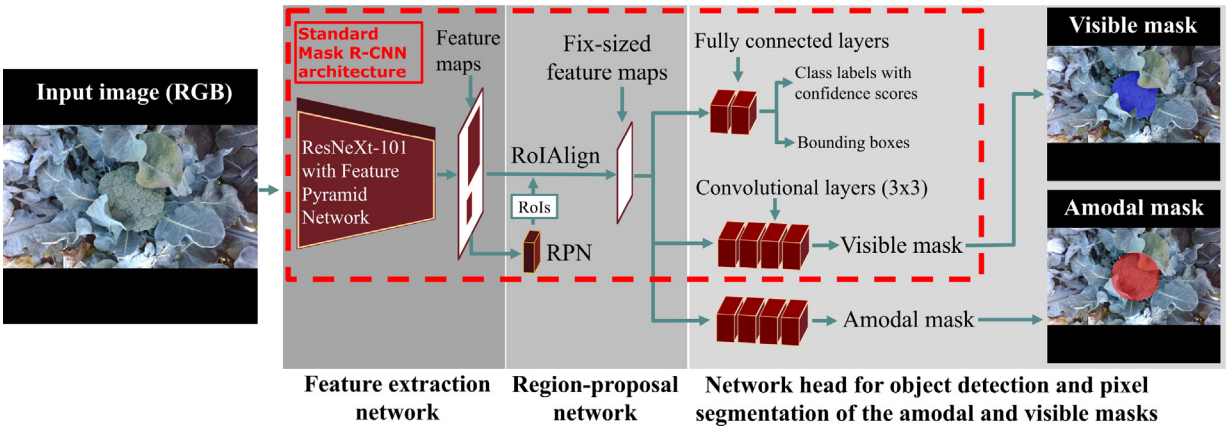


Fig. 9 – Schematic representation of the architecture of the ORCNN network that was used in this research. The part within the dashed red box represents the conventional Mask R–CNN architecture. ORCNN had two mask head branches: one for the visible mask segmentation and one for the amodal mask segmentation. The image was adapted from Follmann et al. (2018) and Shi et al. (2019).

2.2.1.2. *Software and hardware*. The software code of the online ORCNN repository of Lam (2020) was used. This code was based on the Mask R–CNN code of Detectron2 (Wu et al., 2019). From the code, we removed all code references of the occlusion mask. The ORCNN network that remained only outputted the visible and the amodal mask. The code of this network can be found at our git repository.

The Mask R–CNN code was implemented from the ORCNN code. First, the ORCNN code was duplicated. In this duplicated code, we disabled all software references of the amodal mask head branch. The software that remained had only one mask head branch for the visible mask segmentation, and the rest of the code was exactly the same as the ORCNN code. This allowed a fair comparison between the conventional Mask R–CNN and ORCNN.

Both networks were installed on a computer with an Intel Core i7 8700K processor (32 GB DDR4 RAM). The computer was equipped with two graphical processing units (GPU) (one NVIDIA GeForce GTX 1080 Ti and one NVIDIA GeForce GTX 1070 Ti) to accelerate the CNN training and testing. The operating system of the computer was Ubuntu Linux (version 18.04). CUDA (version 10.1) was used as the computational back-end. Both codes were deployed in Python (version 3.8) with Pytorch (version 1.4) and Torchvision (version 0.5) as the deep-learning libraries.

2.2.1.3. *Training procedure*. Transfer-learning was used to initialise the weights of both networks with the weights of Mask R–CNN that was trained on the Microsoft Common Objects in Context (COCO) data set (Lin et al., 2014). Then, the CNN's were fine-tuned on our own training data. The training procedures of Mask R–CNN and ORCNN were exactly the same. Both networks were trained with the stochastic gradient descent optimiser with a momentum of 0.9 and a weight decay of $1.0 \cdot 10^{-4}$. The image batch size was two. The training procedures used the same data augmentations: a random horizontal flip of the image (with a probability of 0.5) and an image resizing along the shortest edge of the image (while maintaining the aspect ratio of the image). Both augmentations were the default data augmentations of the Mask R–CNN code of Detectron2 (Wu et al., 2019).

Both networks were trained for 15000 iterations. The first 1000 iterations served as warm-up, where a lower learning rate of $4.0 \cdot 10^{-5}$ was used that slowly build up to the initial learning rate of $2.0 \cdot 10^{-2}$. This learning rate build-up was applied to stabilise the learning process in the initial phase of the training. Between the 1000th and the 7000th iteration, the initial learning rate of $2.0 \cdot 10^{-2}$ was used. Then, a 0.1 step-based learning rate decay became effective, causing the learning rate to be $2.0 \cdot 10^{-3}$ between the 7000th and the 11000^{th} iteration. The decay was again applied at the 11000^{th} iteration, causing the learning rate to be $2.0 \cdot 10^{-4}$ between the 11000^{th} and the last iteration.

At every 20th iteration, the training and the validation loss were calculated (the loss summarises the classification, localisation and segmentation error). The training loss was calculated on the 1569 training images and the validation loss was calculated on the 504 validation images. These validation images were not used to train the neural network weights, but to inspect whether the network was overfitting. Network overfitting occurs when the network weights are too specifically optimised on the training images, making it harder to generalise on the validation images, leading to an increase in the validation loss. After the training, the network weights with the lowest validation loss were selected.

2.1.1.4. *Image inference procedure*. The selected network weights were used to either segment the visible mask (when using Mask R–CNN) or to segment the visible and the amodal masks (when using ORCNN). The mask segmentation with Mask R–CNN and ORCNN was done with a fixed threshold on the confidence level ($\tau_{CNN} = 0.5$) and a fixed threshold on the non-maximum suppression (NMS) ($\tau_{NMS} = 0.01$). With this NMS threshold, all instances were removed that overlapped with a more confident instance, resulting just one instance segmentation per broccoli head. This approach was considered valid since the broccoli heads grew solitary and did not overlap each other.

2.2.2. Diameter estimation

The second sub task in the broccoli size estimation, was the calculation of the real-world diameter in the registered depth image, using the segmentation output of Mask R–CNN or ORCNN. In this paragraph, the software method that was used for this calculation is described. The diameter estimation method consisted of three algorithms: a circle fit (section 2.2.2.1), a histogram filtering (section 2.2.2.2), and a pixel-to-millimetre conversion (section 2.2.2.3).

2.2.2.1. Circle fit on the segmented mask. The first algorithm involved a circle fit procedure to estimate the diameter of the broccoli head in pixels. With Mask R–CNN, the circle fit procedure was applied on the visible mask, Fig. 10. With ORCNN, the circle fit procedure was applied on the amodal mask, Fig. 11.

The circle fit procedure involved several sub methods. First, the pixel contour of the mask was extracted. From that contour, the convex hull shape was obtained. The convex hull excluded the concave points of the original mask contour, which were considered irrelevant for the circle fit. From the contour points of the convex hull, a circle was fitted with the least squares method of Kanatani and Rangarajan (2011) using the software of Klear (2019). From the fitted circle, the centre point coordinates and the radius were extracted.

One thing we noticed when testing the circle fit algorithm on our training and validation images, was that the least squares method sometimes discarded the contour points of broccoli florets that grew outside the contour of the broccoli head. This discarding of contour points resulted in an underestimation of the broccoli diameter. To prevent this, an additional software method was implemented, which selected the biggest radius from either the least squares circle fit or the minimum enclosing circle fit. This minimum enclosing circle was drawn on all contour points of the mask using the OpenCV software library (version 4.2). With this enclosing circle fit procedure, the extended broccoli florets were included in the circle. After the selection of the biggest radius between the least squares circle and the minimum enclosing circle, the pixel diameter was calculated by multiplying the radius by 2.

2.2.2.2. Histogram filtering in the depth image. Before the pixel diameter could be converted to a real-world diameter, a histogram filtering was applied on the depth image. This filtering algorithm removed any mis-segmented pixel from the visible mask that did not belong to the broccoli head. A wrong segmentation, even if it represented only a few pixels, could potentially cause an offset of centimetres or even decimetres when these pixels are transferred to the registered depth image (because a mis-segmented pixel can belong to a leaf that can be decimetres higher than the broccoli head). A depth offset can cause an inaccurate pixel-to-millimetre conversion and an inaccurate diameter estimation.

First, the pixel segmentation of the visible broccoli region, Fig. 12a, was masked onto the depth image, resulting a depth mask as visualised in Fig. 12b. All pixels with a depth value inside the depth mask were put into a histogram with ten bins, see Fig. 12c. The histogram bin with the highest number of depth pixels was selected, assuming that this would represent the majority of the depth pixels of the broccoli head. From the selected bin, the lowest and the highest depth value were extracted. These depth values were averaged, resulting the depth value of the centre of the bin. From this depth value, 5 cm was subtracted to obtain the lowest depth value of the broccoli head. Then, 10 cm was added to this lowest depth value to obtain the highest depth value of the broccoli head. This addition was based on the assumption that the maximum depth range of a broccoli head could not be more than 10 cm. The other depth pixels that fell outside the selected depth range were removed, refer to the red-coloured bin in Fig. 12c. The depth pixels that belonged to that bin(s) were considered as depth outliers.

Because the broccoli depth range could vary between a small-sized and a big-sized broccoli head, an additional filtering method was implemented. This filtering method started with the creation of another histogram from the selected depth pixels of the broccoli head, see Fig. 12d. The histogram was normalised and also consisted of ten bins. From the ten bins, only the bins with a value of 0.04 or higher were selected. The other bins were removed, since less than

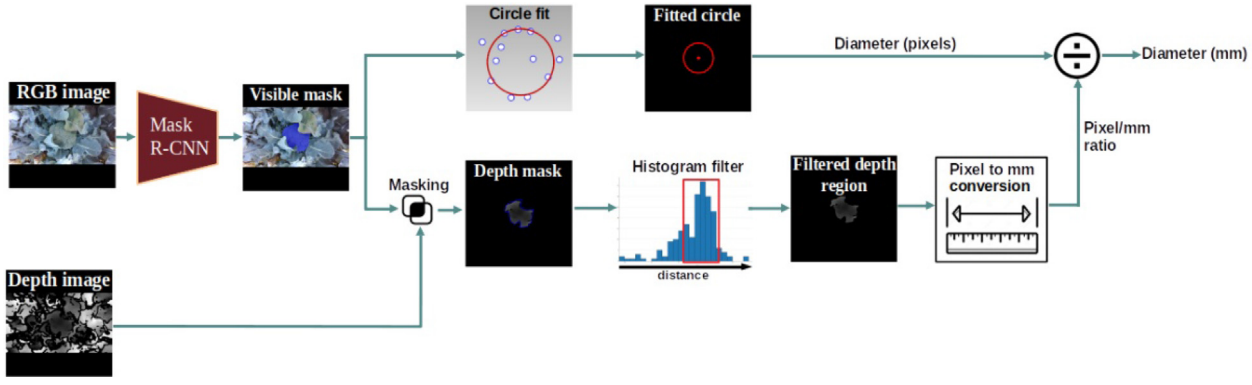


Fig. 10 – Schematic representation of the diameter estimation method using the segmentation output of Mask R–CNN.

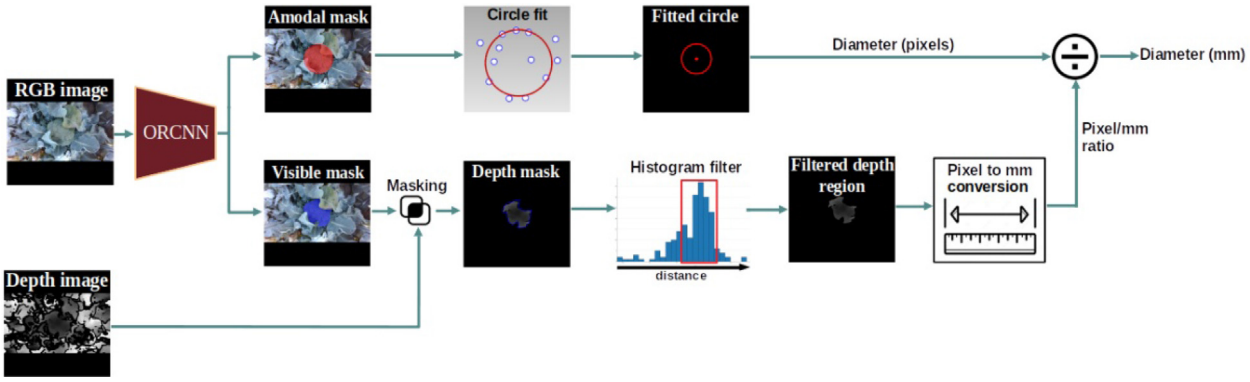


Fig. 11 – Schematic representation of the diameter estimation method using the segmentation output of ORCNN.

4% of the broccoli depth pixels were within that bin. These depth pixels were considered as depth outliers, refer to the red-coloured bins in Fig. 12d. From the selected bins (the blue-coloured bins in Fig. 12d), the overall lowest depth value was selected, which was used to calculate the pixel-to-millimetre conversion.

2.2.2.3. *Pixel-to-millimetre conversion*. With the third algorithm, the real-world diameter of the broccoli head was estimated in millimetres (mm). This estimation was done with a pixel-to-millimetre conversion factor. The factor was calculated from the lowest depth values of the broccoli head, after the histogram filtering. These lowest depth values were expected to represent the depth of the contour of the broccoli head, where also the ground-truth measurement was done, refer to Fig. 3b. Examples of these lowest depth values are the grey-coloured circles in the point-cloud of Fig. 13. All pixels inside the depth mask that had the same depth value as the lowest depth value were selected. From this sub-selection of depth pixels, two pixels were randomly selected and the Euclidean distance between them was calculated (in pixels).

With the use of the camera-intrinsics of the stereo-vision cameras, the 3D real-world coordinates of the two selected pixels were calculated (in x, y, z coordinates). Because the camera was horizontally levelled and the pixels were sampled at the same depth, these pixels approximated the same horizontal depth plane of the broccoli head. The Euclidean distance between the two selected pixels was calculated in millimetres, see Fig. 13. The pixel-to-millimetre conversion factor was calculated by dividing the earlier obtained pixel distance with the real-world millimetre distance. Finally, the diameter of the broccoli head was estimated in millimetres by dividing the pixel diameter of the fitted circle with the pixel-to-millimetre conversion factor, Figs. 10 and 11.

2.3. *Evaluation*

The performance of Mask R–CNN and ORCNN was evaluated with three metrics. The first metric was the detection performance, which specified the ability of each CNN to detect the broccoli heads that were present in the RGB image (section 2.3.1). The second metric was the segmentation performance, which specified the ability of each CNN to segment the pixels of the broccoli head (section 2.3.2). The third metric was the sizing performance, which specified the ability of each sizing method to estimate the real-world diameter of the broccoli head (section 2.3.3).

2.3.1. *Detection performance*

The detection performance was evaluated on the 487 RGB images of the test set. In total, 637 broccoli heads were annotated in these images, which were the 487 broccoli heads that were tagged and measured in the field, and 150 broccoli heads that were annotated in the image but not measured in the field. These 150 broccoli heads were only partially captured in the image, due to the camera's field-of-view, see an example of such a broccoli in the bottom of Figure 5a and b. These 150 broccoli heads were not part of the size experiment, because they were only partially captured in the image. The detection results were calculated for the total number of broccoli heads (637) and for the two subsets (that respectively consisted of 487 and 150 broccoli heads).

A threshold (τ_{CNN}) of 0.5 on the CNN's confidence level (c) was used to determine whether there was a detection ($c \geq \tau_{CNN}$) or not ($c < \tau_{CNN}$). A threshold (τ_{IoU}) of 0.5 on the Intersection over Union (IoU) was used to determine whether the visible mask segmentation was a broccoli head (IoU $\geq \tau_{IoU}$) or background (IoU $< \tau_{IoU}$). The IoU is a measure for the pixel overlap between the ground-truth mask, M_{gt}, and the predicted mask, M_p (Equation (3)), and varies between zero (no pixel overlap) and one (complete pixel overlap). The IoU was calculated on the visible mask, because this was the only common output between Mask R–CNN and ORCNN.

$$\text{IoU} = \frac{|M_{gt} \cap M_p|}{|M_{gt} \cup M_p|} \tag{3}$$

where $|\cdot|$ gives the total number of broccoli pixels.

With the thresholds on the confidence level and the IoU, the number of true positives ($c \geq \tau_{CNN}$ and IoU $\geq \tau_{IoU}$), false positives ($c \geq \tau_{CNN}$ and IoU $< \tau_{IoU}$) and false negatives ($c < \tau_{CNN}$) were determined. A true positive was a broccoli head that was segmented as a broccoli head, a false positive was background that was segmented as a broccoli head, and a false negative was a broccoli head that was not segmented. With the total number of true positives (TP), false positives (FP), and false negatives (FN), the precision P (Equation (4)) and the recall R (Equation (5)) were calculated for both Mask R–CNN and ORCNN. The precision indicated the percentage of correct

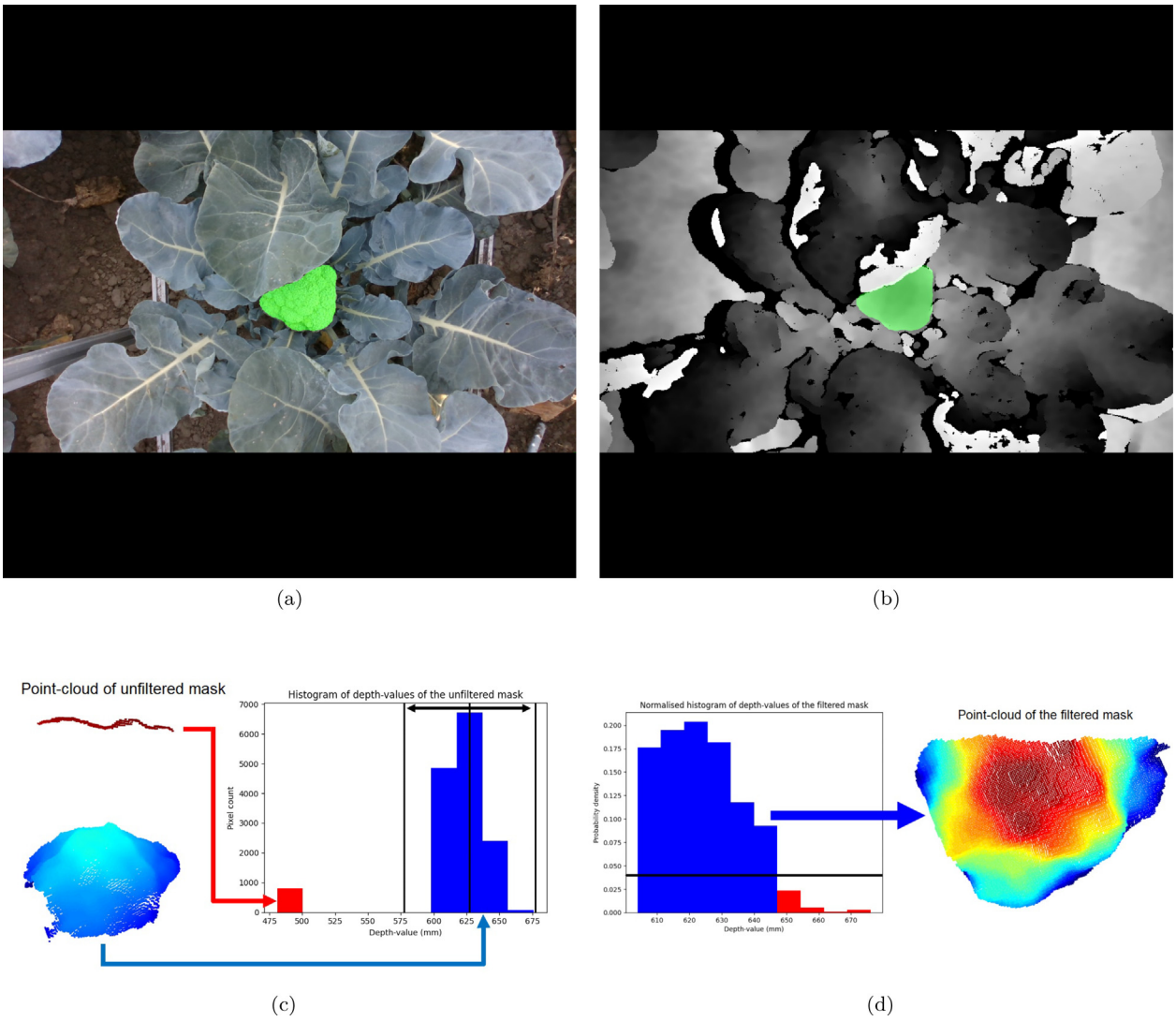


Fig. 12 – The histogram filtering procedure explained: (a) The segmentation of the visible mask (green pixels) in the RGB image. (b) A depth mask was created, by masking the visible mask onto the registered depth image. (c) The depth values of the pixels inside the depth mask were put into a histogram with 10 bins. The bin with the highest number of depth pixels was selected from the histogram. Then, the depth value of the centre of the selected bin was determined (see middle black vertical line in the histogram). From this depth value, 5 cm was subtracted and 5 cm was added to obtain the depth range of the broccoli head. The bins within this depth range (visualised by the left and right black vertical line in the histogram) were selected. The other bin, visualised in red, was removed because it fell outside the depth range. In this example, the removed bin represented the depth values of a higher positioned leaf (refer to the red coloured part of the unfiltered point cloud). (d) The depth values of the selected bins were put into another normalised histogram with 10 bins. The bins with a value of 0.04 (horizontal black line) or higher were selected (these bins are coloured blue). The other bins were removed (red-coloured bins). The depth pixels that remained after the filtering were presumed to belong to the broccoli head (the rainbow-coloured point-cloud is a representation of the broccoli head after the histogram filtering).

detections and the recall indicated the percentage of broccoli heads that were successfully detected by the CNN's.

$$P = \frac{TP}{TP + FP} \tag{4}$$

$$R = \frac{TP}{TP + FN} \tag{5}$$

2.3.2. *Segmentation performance*

The segmentation performance was evaluated on the 487 broccoli heads that were tagged and measured in the field, because only these heads had an accurate ground-truth annotation of the amodal region of the broccoli head.

The 487 broccoli heads were assigned into ten groups based on their occlusion rate, in the range of 0–1 with steps of 0.1. The segmentation performance of Mask R–CNN and ORCNN

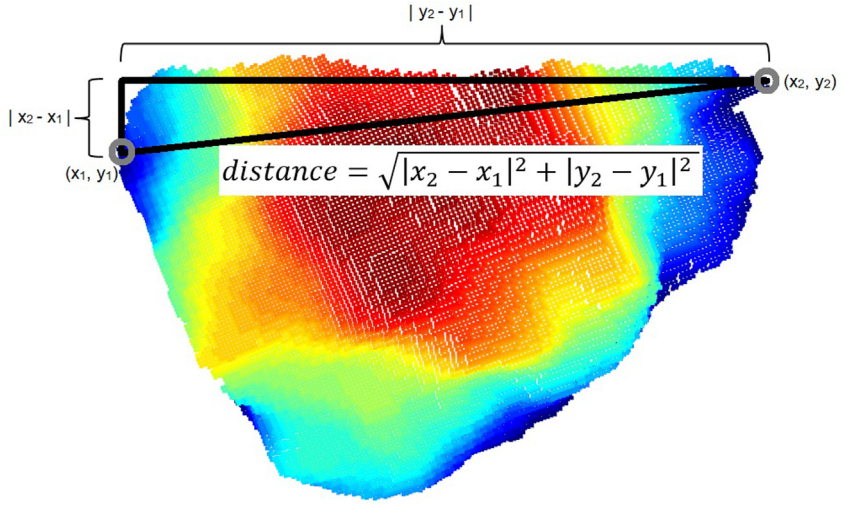


Fig. 13 – The rainbow-coloured point-cloud is a representation of the broccoli head after the histogram filtering. The colour-scale of the point-cloud is based on the depth values: the red and orange-coloured points are closer to the camera compared to the cyan and blue-coloured points. On this point-cloud, the pixel-to-millimetre conversion factor was calculated from two of the lowest depth values (visualised by the two grey-coloured circles). These two depth values were sampled at the same depth (same z coordinate) and were expected to represent the depth of the contour of the broccoli head, where the ground-truth measurement was done. The pixel-to-millimetre conversion factor was obtained from the division of the Euclidean pixel distance and the Euclidean millimetre distance between these two points.

was calculated for the broccoli heads that belonged to an occlusion group (and this was repeated for each occlusion group). The segmentation performance was also calculated for all broccoli heads, irrespective of their occlusion rate.

The segmentation performance was evaluated with the IoU (Equation (3)), which was calculated for both the visible and the amodal mask. Because Mask R–CNN did not output an amodal mask, its amodal IoU was calculated between the circle that was fitted on the visible mask and the amodal annotation that was used to train ORCNN. To allow a fair comparison, the amodal IoU of ORCNN was also calculated between the circle that was fitted on the amodal mask and the amodal annotation.

A pairwise Wilcoxon test (Wilcoxon, 1945) with a significance level of 5% ($\alpha = 0.05$) was employed for the visible and the amodal IoU values to test whether there were statistical differences between the Mask R–CNN segmentation and the ORCNN segmentation. We used the Wilcoxon test, because it can deal with non-normally distributed data, like the IoU.

2.3.3. *Size estimation performance*

The size estimation performance was also expressed for the ten occlusion rate groups (in the range of 0–1 with steps of 0.1). The performance metric was the diameter error (ϵ), which was the difference between the diameter estimate of the CNN sizing methods ($\widehat{d}$) and the diameter measurement that was done in the field, (d) (Equation (6)).

$$\epsilon = \widehat{d} - d \tag{6}$$

The diameter error was evaluated by means of the median error ($\tilde{\epsilon}$), the median absolute error (MAD, Equation (7)), the mean absolute error (MAE, Equation (8)) and the root mean squared error (RMSE, Equation (9)).

$$\text{MAD} = \text{median}(|\epsilon_i - \tilde{\epsilon}|) \tag{7}$$

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|\epsilon| \tag{8}$$

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\epsilon^2} \tag{9}$$

A pairwise Wilcoxon test with a significance level of 5% ($\alpha = 0.05$) was applied on the absolute diameter errors, $|\epsilon|$, to test whether there were statistical differences between the Mask R–CNN sizing method and the ORCNN sizing method. Finally, the Pearson correlation coefficient, r, was calculated to investigate the relation between the amodal segmentation performance and the absolute diameter error.

3. Results

3.1. *Detection results*

The detection results of Mask R–CNN and ORCNN are summarised in Table 1 and Table 2. The detection results are summarised for the two subsets of broccoli heads and for the total number of broccoli heads in the test images. The first subset represented the 487 broccoli heads that were measured in the field and that were used for the size experiment. Both Mask R–CNN and ORCNN detected all broccoli heads without false positive detections, indicating that both CNN's reached a

Table 1 – Mask R–CNN detection results on the test images. The abbreviations indicate the number of ground-truth annotations (GT), the number of detections by the CNN (Det), the number of true positives (TP), the number of false positives (FP), the number of false negatives (FN), the precision (P) and the recall (R).

Subset	GT	Det	TP	FP	FN	P	R
1. Broccoli heads used in the size experiment	487	487	487	0	0	100.0%	100.0%
2. Broccoli heads not used in the size experiment	150	143	133	10	17	93.0%	88.7%
All broccoli heads in the test images	637	630	620	10	17	98.4%	97.3%

Table 2 – ORCNN detection results on the test images. The meaning of the abbreviations can be found in the caption of Table 1.

Subset	GT	Det	TP	FP	FN	P	R
1. Broccoli heads used in the size experiment	487	487	487	0	0	100.0%	100.0%
2. Broccoli heads not used in the size experiment	150	151	136	15	14	90.1%	90.7%
All broccoli heads in the test images	637	638	623	15	14	97.6%	97.8%

recall and a precision of 100.0% on these broccoli heads, see Tables 1 and 2.

The second subset contained the 150 broccoli heads that were only partially captured in the image, due to the camera's field-of-view. Of these 150 partially-captured broccoli heads (which were not used in the size experiment), 133 heads were detected by Mask R–CNN, Table 1 and 136 heads were detected by ORCNN, Table 2. Mask R–CNN had 10 false positive detections and ORCNN had 15 false positive detections. Mask R–CNN had a precision of 93.0% and a recall of 88.7% on these 150 broccoli heads. ORCNN had a precision of 90.1% and a recall of 90.7%.

The final calculation was done on the total number of broccoli heads in the test images. Mask R–CNN detected 620 of the 637 broccoli heads, Table 1, and ORCNN detected 623 of the 637 broccoli heads, Table 2. Mask R–CNN had in total 10 false positive detections and ORCNN had 15 false positive detections. With Mask R–CNN, the precision was 98.4% and the recall was 97.3%, Table 1. ORCNN had a precision of 97.6% and a recall of 97.8%, Table 2.

3.2. *Segmentation results*

The broccoli segmentation performance was calculated on the 487 broccoli heads that were measured in the field and used in the size experiment. Table 3 summarises the Intersection over Union (IoU) values for the visible mask segmentations of Mask R–CNN and ORCNN for the ten occlusion rate groups. Table 3 also summarises the statistical results of the pairwise Wilcoxon test. For eight occlusion rates, ORCNN had a significantly lower IoU on the visible region of the broccoli head compared to Mask R–CNN. However, the effect size was small,

Table 3 – Statistics on the Intersection over Union (IoU) values for the visible mask segmentations of Mask R–CNN and ORCNN. The statistics are expressed for the ten occlusion rate groups. The last row summarises the IoU statistics for all broccoli heads that were used in the size experiment.

Occlusion rate (n = number of broccoli heads)	Mean IoU on the visible mask (standard deviation)		*p*-value Wilcoxon test
	Mask R–CNN	ORCNN	
0.0–0.1 (n = 147)	0.97 (0.01)	0.96 (0.01)	0.00 (****)
0.1–0.2 (n = 60)	0.96 (0.01)	0.95 (0.01)	0.00 (****)
0.2–0.3 (n = 33)	0.95 (0.01)	0.95 (0.01)	0.09 (ns)
0.3–0.4 (n = 35)	0.94 (0.02)	0.93 (0.02)	0.00 (****)
0.4–0.5 (n = 47)	0.93 (0.03)	0.93 (0.03)	0.00 (****)
0.5–0.6 (n = 35)	0.92 (0.03)	0.91 (0.04)	0.00 (****)
0.6–0.7 (n = 65)	0.90 (0.05)	0.88 (0.05)	0.00 (****)
0.7–0.8 (n = 41)	0.87 (0.06)	0.85 (0.07)	0.00 (****)
0.8–0.9 (n = 20)	0.83 (0.08)	0.79 (0.09)	0.00 (***)
0.9–1.0 (n = 4)	0.84 (0.10)	0.75 (0.11)	–
All (n = 487)	0.93 (0.05)	0.92 (0.06)	0.00 (****)

- (too few samples), ns = not significant ($p > 0.05$), * ($0.01 < p \leq 0.05$); **($0.001 < p \leq 0.01$); ***($0.0001 < p \leq 0.001$); ****($p \leq 0.0001$).

Table 4 – Statistics on the Intersection over Union (IoU) values for the amodal mask segmentations of Mask R–CNN and ORCNN, expressed for the ten occlusion rate groups. The last row summarises the IoU statistics for all broccoli heads that were used in the size experiment.

Occlusion rate (n = number of broccoli heads)	Mean IoU on the amodal mask (standard deviation)		*p*-value Wilcoxon test
	Mask R–CNN	ORCNN	
0.0–0.1 (n = 147)	0.95 (0.03)	0.95 (0.03)	0.00 (**)
0.1–0.2 (n = 60)	0.93 (0.04)	0.94 (0.04)	0.00 (**)
0.2–0.3 (n = 33)	0.89 (0.06)	0.90 (0.07)	0.17 (ns)
0.3–0.4 (n = 35)	0.86 (0.09)	0.92 (0.04)	0.00 (**)
0.4–0.5 (n = 47)	0.82 (0.11)	0.89 (0.06)	0.00 (****)
0.5–0.6 (n = 35)	0.80 (0.13)	0.87 (0.09)	0.00 (**)
0.6–0.7 (n = 65)	0.75 (0.18)	0.86 (0.11)	0.00 (****)
0.7–0.8 (n = 41)	0.63 (0.22)	0.82 (0.12)	0.00 (****)
0.8–0.9 (n = 20)	0.47 (0.22)	0.80 (0.11)	0.00 (****)
0.9–1.0 (n = 4)	0.28 (0.25)	0.72 (0.14)	–
All (n = 487)	0.83 (0.18)	0.90 (0.09)	0.00 (****)

- (too few samples), ns = not significant ($p > 0.05$), * ($0.01 < p \leq 0.05$); **($0.001 < p \leq 0.01$); ***($0.0001 < p \leq 0.001$); ****($p \leq 0.0001$).

with IoU differences between 0.01 and 0.09. For the four most heavily occluded broccoli heads, which had an occlusion rate between 90% and 100%, the Wilcoxon test could not be applied, because there were too few test samples.

Table 4 summarises the IoU values for the amodal mask segmentations of Mask R–CNN and ORCNN for the ten occlusion rate groups. For eight occlusion rates, ORCNN had a significantly higher IoU on the amodal region of the broccoli head compared to Mask R–CNN. For the broccoli heads with

an occlusion rate between 0% and 60%, the effect size was small, with IoU differences of maximally 0.07. For the broccoli heads with an occlusion rate between 60% and 90%, the IoU differences were between 0.11 and 0.33. Again, we could not apply the Wilcoxon test for the broccoli heads with an occlusion rate between 90% and 100%, because there were only four samples.

3.3. *Size estimation results*

In Fig. 14, the diameter errors of the Mask R–CNN sizing method and the ORCNN sizing method are plotted in histograms. With Mask R–CNN, 323 of the 487 diameter estimates were underestimated (66.3%) and 164 of the 487 estimates were overestimated (33.7%). The median diameter error of Mask R–CNN, $\tilde{\epsilon}_{\mathrm{mrcnn}}$, was −2.4 mm, indicating that the majority of the estimates were underestimated. With ORCNN, the diameter estimates were more balanced, as there were 209 underestimations (42.9%) and 278 overestimations (57.1%). The median diameter error of ORCNN, $\tilde{\epsilon}_{\mathrm{orcnn}}$, was 1.1 mm, indicating that the majority of the estimates were overestimated and that the median error was smaller than the one from Mask R–CNN.

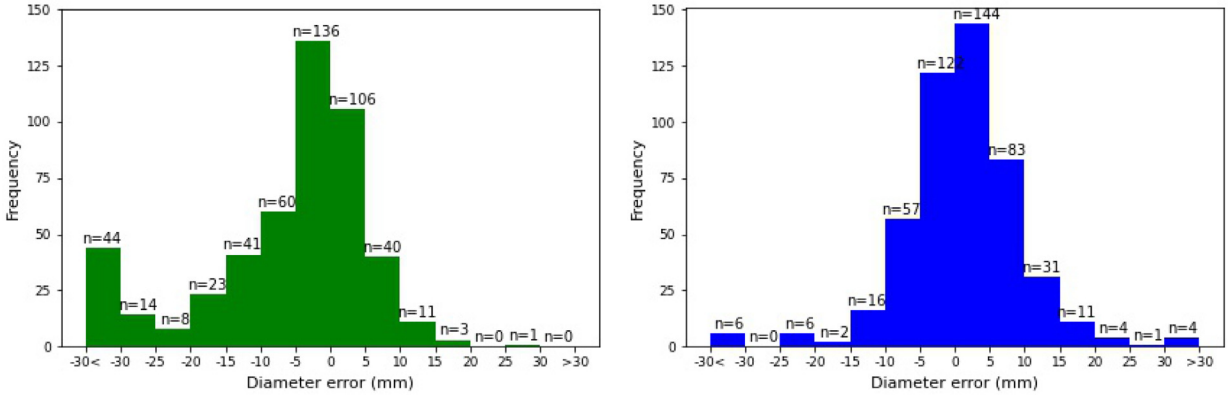
Figure 15 shows the cumulative percentages of the absolute diameter errors. Three error margins (10 mm, 20 mm and 30 mm) are marked by the dashed vertical lines. With the Mask R–CNN sizing method, 342 of the 487 diameter estimates (70.2%) were within 10 mm from the ground-truth diameter. With the ORCNN sizing method, 406 of the 487 diameter estimates (83.4%) were within 10 mm from the ground-truth diameter. The number of diameter estimates that were within 20 mm and within 30 mm from the ground-truth, were respectively 420 (86.2%) and 443 (91.0%) with Mask R–CNN, and 466 (95.7%) and 477 (97.9%) with ORCNN. With Mask R–CNN, 44 of the 487 diameter estimates (9.0%) deviated more than 30 mm from the ground-truth. With ORCNN, this number was 10 (2.1%).

Table 5 summarises the median absolute diameter error (MAD) and the root mean square diameter error (RMSE) for the two sizing methods and the ten occlusion rate groups. The Mask R–CNN sizing method had a MAD of 7.9 mm and a RMSE of 18.7 mm on all broccoli heads. With the ORCNN sizing method, the MAD was 6.7 mm and the RMSE was 9.7 mm.

Table 6 summarises the mean absolute diameter error (MAE) and the statistics of the pairwise Wilcoxon test. The Mask R–CNN sizing method had a MAE of 10.7 mm on all broccoli heads. With the ORCNN sizing method, the MAE was 6.4 mm. The Wilcoxon test revealed that the ORCNN sizing method had a significantly lower absolute diameter error than Mask R–CNN on 161 broccoli heads with an occlusion rate between 50% and 90%, Table 6. For these occlusion rates, the diameter error differences were between 3.5 and 28.4 mm. For the four most heavily occluded broccoli heads, which had an occlusion rate between 90% and 100%, the Wilcoxon test could not be applied, because there were too few test samples.

In Fig. 16, the relation between the amodal segmentation performance and the absolute diameter error is plotted. The Pearson's correlation coefficient was −0.86, indicating that there was a negative correlation between the amodal segmentation performance and the absolute diameter error (the lower the amodal IoU the higher the absolute diameter error, and vice versa). The effect of the amodal segmentation on the sizing performance is also visualised in Figs. 17 and 18. On these two heavily occluded broccoli heads, the ORCNN sizing method had a higher amodal IoU and a lower absolute diameter error compared to the Mask R–CNN sizing method.

With the Mask R–CNN sizing method, the biggest diameter error was −123.1 mm, Fig. 19. The biggest diameter error of the ORCNN sizing method was −74.2 mm, Fig. 20, and this error was found on the same image frame as Fig. 19. The main cause of both errors was the underestimation of the amodal region of the broccoli head. Additionally, there was an inaccurate pixel-to-millimetre conversion, because there were no valid depth pixels of the broccoli head. All depth pixels in Figures 19b and 20b belonged to an occluded leaf, causing that the pixel-to-millimetre conversion factor was calculated on the higher-positioned leaf instead of the broccoli head. With the Mask R–CNN sizing method, the inaccurate size conversion contributed 11.3% to the total diameter error. With the ORCNN sizing method, the contribution was higher: 44.1%. An analysis on the five biggest diameter errors of ORCNN revealed that two more errors were primarily caused by such an



(a) (b)

Fig. 14 – (a) Histogram of the diameter error of the Mask R–CNN sizing method on the 487 broccoli heads. (b) Histogram of the diameter error of the ORCNN sizing method. *n* is the number of diameter estimates.

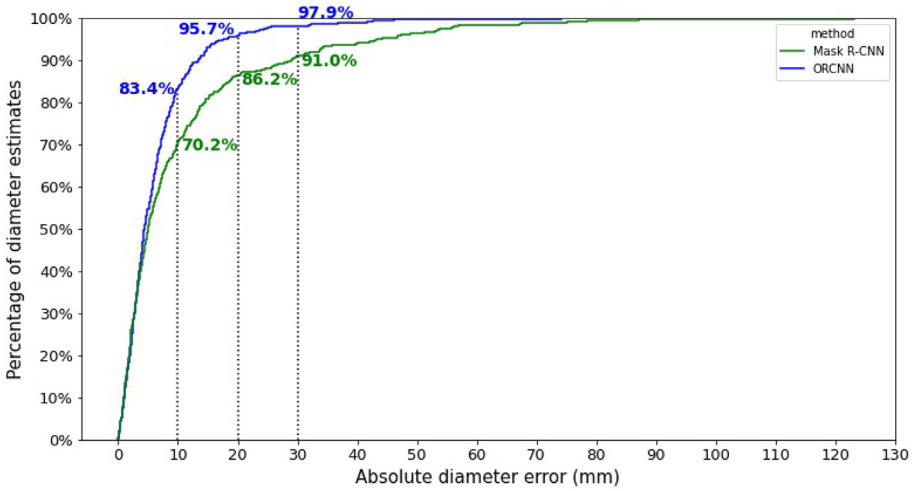


Fig. 15 – Cumulative percentages of the absolute diameter errors of the two sizing methods. The three dashed vertical lines indicate the error margins of 10 mm, 20 mm and 30 mm. The coloured numbers summarise the cumulative percentages of the absolute diameter estimates that were within 10 mm, 20 mm and 30 mm from the ground-truth diameter.

Table 5 – The median absolute diameter error and the root mean square diameter error (RMSE) for the two sizing methods. Both errors were calculated for the ten occlusion rate groups and for the total number of broccoli heads (last row).

Occlusion rate (n = number of broccoli heads)	Median absolute diameter error (mm)		RMSE of the diameter (mm)	
	Mask R–CNN	ORCNN	Mask R–CNN	ORCNN
0.0–0.1 (n = 147)	4.3	4.9	4.7	4.9
0.1–0.2 (n = 60)	3.6	3.6	4.1	4.5
0.2–0.3 (n = 33)	6.2	5.6	6.8	6.7
0.3–0.4 (n = 35)	8.9	8	8.4	7.5
0.4–0.5 (n = 47)	7.4	5.6	11.2	7.9
0.5–0.6 (n = 35)	9.6	7.3	12.8	8.9
0.6–0.7 (n = 65)	9.6	7.6	21.3	11.1
0.7–0.8 (n = 41)	23.6	14.1	31.5	15.9
0.8–0.9 (n = 20)	34.2	13.4	49.1	19.9
0.9–1.0 (n = 4)	41.5	6.3	88.5	38.5
All (n = 487)	7.9	6.7	18.7	9.7

Table 6 – Statistics on the mean absolute diameter error for the two sizing methods, expressed for the ten occlusion rate groups and the total number of broccoli heads (last row).

Occlusion rate (n = number of broccoli heads)	Mean absolute diameter error (mm) (standard deviation)		p-value Wilcoxon test
	Mask R–CNN	ORCNN	
0.0–0.1 (n = 147)	3.6 (3.1)	3.9 (2.9)	0.11 (ns)
0.1–0.2 (n = 60)	3.2 (2.5)	3.8 (2.4)	0.05 (ns)
0.2–0.3 (n = 33)	5.4 (4.1)	5.4 (4.0)	0.64 (ns)
0.3–0.4 (n = 35)	7.0 (4.8)	6.1 (4.5)	0.39 (ns)
0.4–0.5 (n = 47)	8.8 (7.0)	6.3 (4.8)	0.06 (ns)
0.5–0.6 (n = 35)	10.1 (7.9)	6.6 (6.0)	0.03 (*)
0.6–0.7 (n = 65)	16.5 (13.5)	7.8 (7.8)	0.00 (****)
0.7–0.8 (n = 41)	25.4 (18.6)	12.2 (10.2)	0.00 (***)
0.8–0.9 (n = 20)	42.9 (24.0)	14.5 (13.6)	0.00 (***)
0.9–1.0 (n = 4)	77.3 (43.2)	27.0 (27.5)	–
All (n = 487)	10.7 (15.3)	6.4 (7.3)	0.00 (****)

- (too few samples), ns = not significant ($p > 0.05$), * ($0.01 < p \leq 0.05$); **($0.001 < p \leq 0.01$); ***($0.0001 < p \leq 0.001$); ****($p \leq 0.0001$).

inaccurate pixel-to-millimetre conversion. When analysing these two errors of respectively −46.2 mm and −41.6 mm, the inaccurate pixel-to-millimetre conversion contributed 82.6% and 88.7% to the total diameter error.

4. Discussion

Our images were acquired with two different cameras on two fields where different broccoli cultivars were grown in different growing seasons. Although there was variation in our data sets, it is not guaranteed that our algorithms will generalise sufficiently on broccoli images from other fields with different cultivars. It is also acknowledged that the image variation was not as extensive as for example the research of Blok et al. (2021). In the research of Blok et al. (2021), the generalisation performance of Mask R–CNN was evaluated on 600 broccoli images that originated from three cultivars, five growing seasons and 11 broccoli fields that were located in three different countries. Despite the lack of such a comprehensive evaluation, it is expected that our algorithms can be efficiently retrained on new data sets to achieve image generalisation. This expectation is based on the research of Blok et al. (2021), in which image generalisation was reached on a new broccoli cultivar by adding 30 images of that cultivar into the training set. To further enhance the retraining process on other data sets, we have made our software, data set and trained algorithms available.

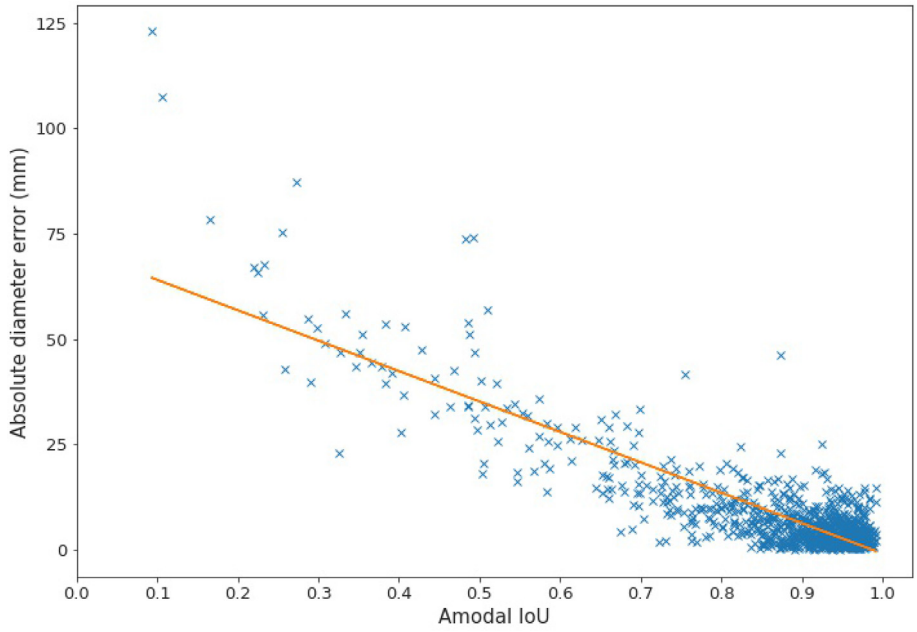


Fig. 16 – A scatter plot showing the relation between the amodal IoU and the absolute diameter error for all 974 estimates of the two CNN sizing methods. The orange line visualises the regression line between the two outputs. The Pearson's correlation coefficient (*r*) was −0.86, indicating that there was a negative correlation between the amodal IoU and the absolute diameter error.

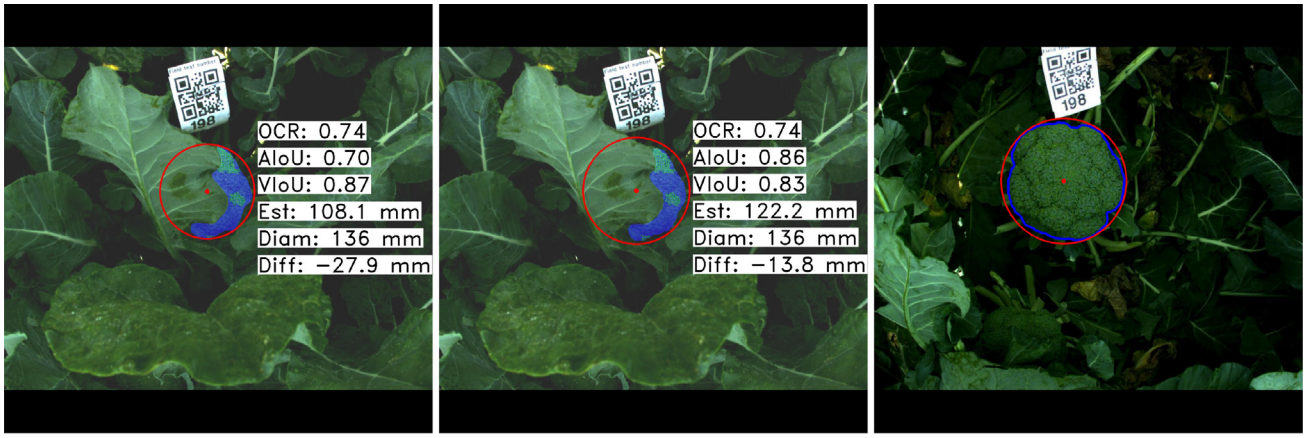


Fig. 17 – (a) The diameter estimation (Est) of the Mask R–CNN sizing method was 108.1 mm on a broccoli head with a measured diameter (Diam) of 136 mm. The diameter error (Diff) was −27.9 mm. The amodal IoU (AIoU) between the circle fit and the amodal annotation was 0.70. The visible IoU (VIoU) was 0.87. The broccoli head had an occlusion rate (OCR) of 74%. (b) On the same image, ORCNN had an amodal IoU of 0.86 and a diameter error of −13.8 mm. In image a & b, the blue pixels visualise the visible mask segmentation. The green pixels are the pixels that had a depth value and that remained after the histogram filtering. The red circle visualises the circle that was fitted on either the visible or the amodal mask. (c) The same broccoli head after removal of all leaves (no occlusion). The amodal mask annotation is visualised by the red circle and the visible mask annotation is visualised by the blue polygon.

In our data sets, 1363 of the 2560 broccoli images (53.2%) had a human-made leaf occlusion. These leaf occlusions were created by the same person, indicating that they may have been subject to some degree of subjectivity. Nevertheless, these human-made occlusions provided natural-looking examples of different levels of leaf occlusion, allowing the systematic evaluation of the algorithms on different occlusion rates. An additional advantage of the leaf modification was that it allowed an accurate annotation of the amodal region of the occluded broccoli head. A similar annotation method can also be used on other occluded crops, which could solve the problem of incorrect annotation of occluded image scenes (a problem that was identified in the research of Zhang et al. (2020)).

Fig. 18 – (a) The diameter estimation (Est) of the Mask R–CNN sizing method was 60.1 mm on a broccoli head with a measured diameter (Diam) of 109 mm. The diameter error (Diff) was −48.9 mm. The amodal IoU (AIoU) between the circle fit and the amodal annotation was 0.31. The visible IoU (VIoU) was 0.84. The broccoli head had an occlusion rate (OCR) of 89%. (b) On the same image, ORCNN had an amodal IoU of 0.85 and a diameter error of −6.6 mm. (c) The same broccoli head after removal of all leaves (no occlusion). The amodal mask annotation is visualised by the red circle and the visible mask annotation is visualised by the blue polygon.

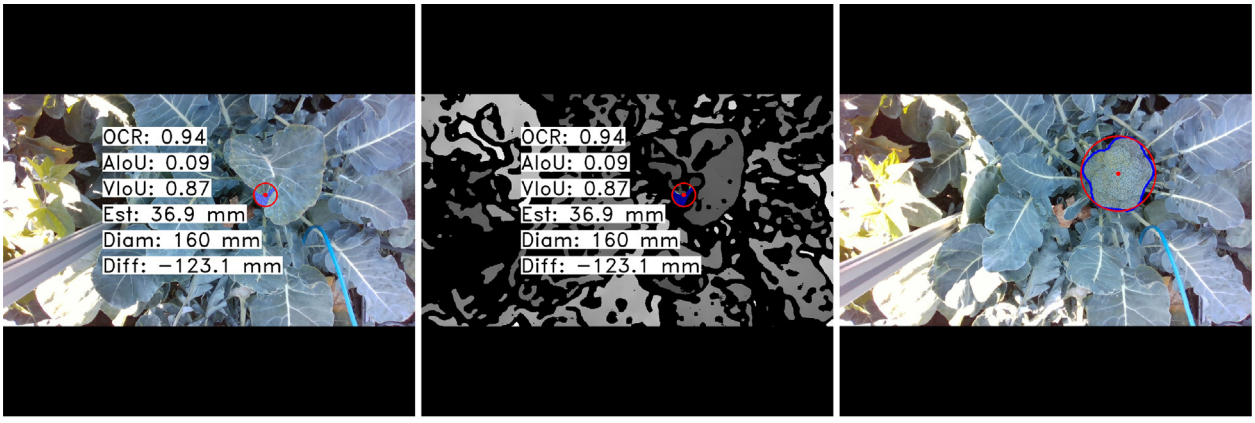


Fig. 19 – (a) The biggest diameter error (Diff) of the Mask R–CNN sizing method was −123.1 mm. The main cause of this error was the underestimation of the amodal region (the amodal IoU was 0.09) of the heavily occluded broccoli head (the occlusion rate (OCR) was 94%). (b) In the registered depth image, all depth pixels belonged to a higher-positioned leaf. This ultimately caused an inaccurate pixel-to-millimetre conversion, which contributed for 11.3% to the total diameter error. (c) The RGB image of the same broccoli head after removal of all leaves (no occlusion). The amodal mask annotation is visualised by the red circle and the visible mask annotation is visualised by the blue polygon.

On both of our data sets, Mask R–CNN and ORCNN detected all broccoli heads that were tagged and measured in the field. For the other broccoli heads that were annotated (but not measured in the field) the detection results were at most 11.3% lower. This is acceptable, as most of these broccoli heads were only partially captured in the field-of-view of the camera. These broccoli heads are likely to be detected in a subsequent frame when the image acquisition device moves further. When comparing our detection results to other broccoli-detection studies (Blok et al., 2021; Kusumam et al., 2017), it can be concluded that both of our CNN's reached a state-of-the-art detection performance, especially since the majority of our broccoli heads were (heavily) occluded.

ORCNN had a significantly lower segmentation performance on the visible region of the broccoli head for eight occlusion rates, although the absolute IoU differences were small. The lower IoU might have been caused by the expansion of the loss function of ORCNN with the additional loss component for the amodal mask. The addition of this extra loss component may have resulted in a reduced minimisation of the other loss components, such as the visual mask. Still, the less optimised visible mask of ORCNN did not seem to negatively affect the overall sizing performance.

ORCNN had a significantly higher segmentation performance on the amodal region of the broccoli head for eight occlusion rates. Especially for the broccoli heads with an

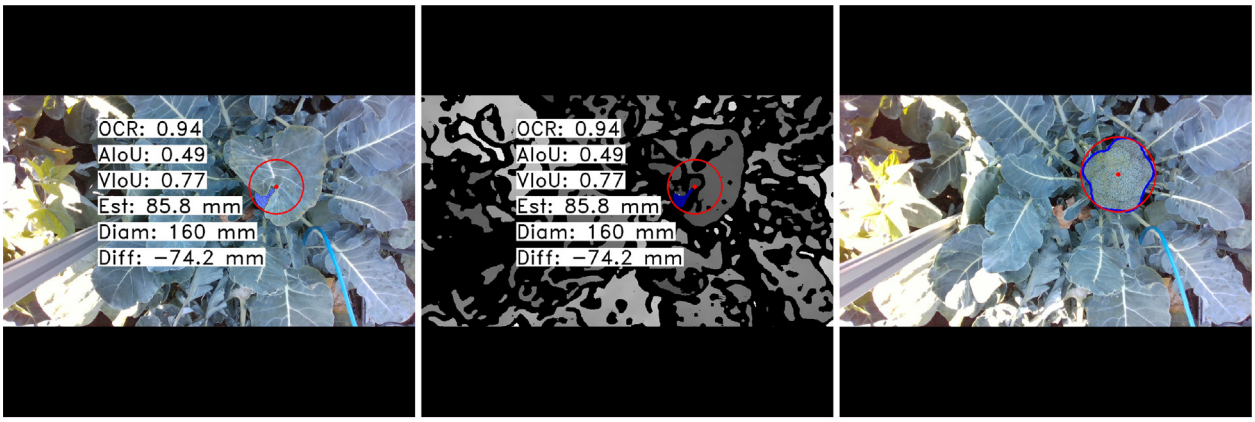


Fig. 20 – (a) The biggest diameter error (Diff) of the ORCNN sizing method was −74.2 mm. This error was found on the same image as Fig. 19. The main cause of this error was the underestimation of the amodal region (the amodal IoU was 0.49) of the heavily occluded broccoli head (the occlusion rate (OCR) was 94%). (b) In the registered depth image, all depth pixels belonged to a higher-positioned leaf. This ultimately caused an inaccurate pixel-to-millimetre conversion, which contributed for 44.1% to the total diameter error. (c) The RGB image of the same broccoli head after removal of all leaves (no occlusion). The amodal mask annotation is visualised by the red circle and the visible mask annotation is visualised by the blue polygon.

occlusion rate higher than 60%, there were large differences between the amodal IoU of ORCNN and Mask R–CNN. For similar amodal predictions with a circle, it could have been sufficient to alter the Mask R–CNN network so that it could detect a circle instead of a bounding box (and then do the pixel segmentation inside the estimated circle). Yet, the flexibility of ORCNN to predict all kinds of crop shapes makes it a more versatile algorithm for use on a variety of crops.

ORCNN significantly improved the diameter estimate of 161 broccoli heads with an occlusion rate between 50% and 90%. Therefore, the ORCNN sizing method should be preferred over the Mask R–CNN sizing method, especially when the size estimation has to be done in broccoli fields where there is more vegetative growth or where the broccoli plants are more densely planted (causing more occlusion).

With the ORCNN sizing method, there was an increase of 13.2% on the number of broccoli heads that were estimated within 10 mm from the ground-truth diameter. Additionally, there was an increase of 9.5% of estimates that were within 20 mm from the ground-truth. The more accurate size estimate of ORCNN could potentially result in less food waste and a higher financial return when the algorithm is used for robotic harvesting. In future research, we want to evaluate the performance of a broccoli harvesting robot that is equipped with the ORCNN sizing method.

The ORCNN sizing method had a median absolute diameter error of 6.7 mm. This error was respectively 2.7 mm and 6.1 mm lower than the median absolute diameter error of the convex hull and the bounding box estimator of Kusumam et al. (2017), who also investigated the in-field size estimation of broccoli heads. A difference is that both estimators of Kusumam et al. (2017) only used the visible part of the 3D point cloud, which ultimately resulted in an underestimated diameter estimate. In our research, ORCNN did not underestimate the broccoli size, as we found a positive median diameter error of 1.1 mm on all broccoli heads. In a similar crop-sizing study by Lin et al. (2019), the diameters of 80 citrus fruits were estimated with the same 3D sizing method as Kusumam et al. (2017) (unfortunately Lin et al. (2019) did not specify whether they used the convex hull or the bounding box estimator). On the relatively smaller citrus fruits, the median diameter error and the median absolute diameter error were respectively −1.0 mm and 4.0 mm, indicating that the estimates of Lin et al. (2019) were slightly better than our ORCNN estimates. An important difference is that Lin et al. (2019) did their evaluation on a lower number of citrus fruits with minimal occlusions, whereas our results were obtained on a test set with heavier occlusions and more broccoli heads.

Despite the promising results of the ORCNN sizing method, there is still room for improvement. An analysis on the five biggest diameter errors of ORCNN revealed that three of the errors were primarily caused by an inaccurate pixel-to-millimetre conversion. This problem was caused by the lack of valid depth pixels of the broccoli head, which in turn was caused by the loss of depth pixels due to the leaf occlusion and the stereo-vision principle of the RGB-D cameras. In all 2560 depth images, on average one fifth of the broccoli depth pixels were lost compared to the broccoli region in the registered RGB image. A way to alleviate this depth pixel loss is to use a camera with a different depth perceiving technique, such as Laser Imaging Detection And Ranging (LIDAR) or Time-of-Flight (ToF). While the depth pixel loss of our stereo-vision cameras sometimes negatively influenced the sizing performance, the main cause of the diameter errors remained the inaccurate estimation of the amodal region of

the broccoli head (this observation was also supported by the negative correlation between the amodal segmentation performance and the absolute diameter error).

5. Conclusions

With ORCNN, the segmentation of the amodal region of the broccoli head significantly improved. ORCNN provided a better estimate of the shape of an occluded broccoli head compared to Mask R–CNN, which estimated the amodal region with a circle fit on the visible broccoli region. With the significantly better amodal segmentation, the ORCNN sizing method achieved a 4.3 mm lower mean absolute diameter error on 487 broccoli heads. Furthermore, with ORCNN there was a 13.2% increase on the number of broccoli heads that were estimated within 10 mm from the ground-truth diameter. The ORCNN sizing method had also a significantly lower absolute diameter error on 161 broccoli heads with an occlusion rate between 50% and 90%. We conclude that ORCNN improved the size estimation of the heavily occluded broccoli heads in our data sets. We encourage other researchers to use our software and data set to further develop methodologies that can deal with crop occlusion.

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.

Acknowledgements

We would like to thank Tony Wisdom and Wiebe Goodijk for providing the broccoli fields on which we acquired the RGB-D images. We thank Toon Tielen, Thierry Stokkermans and Hyejeong Kim for their help with the image acquisition. Thanks to Felipe Schadeck Fiorentin and Kevin Yao for the image annotation.

REFERENCES

Barth, R., Hemming, J., & van Henten, E. J. (2016). Design of an eye-in-hand sensing and servo control framework for harvesting robotics in dense vegetation. *Biosystems Engineering, 146*, 71–84. https://doi.org/10.1016/j.biosystemseng.2015.12.001. Special Issue: Advances in Robotic Agriculture for Crops.

Bender, A., Whelan, B., & Sukkarieh, S. (2019). *Ladybird cobbitty 2017 brassica dataset*. https://doi.org/10.25910/5c941d0c8bccb

Bender, A., Whelan, B., & Sukkarieh, S. (2020). A high-resolution, multimodal data set for agricultural robotics: A ladybird's-eye view of brassica. *Journal of Field Robotics, 37*(1), 73–96. https://doi.org/10.1002/rob.21877

Blok, P. M., van Evert, F. K., Tielen, A. P. M., van Henten, E. J., & Kootstra, G. (2021). The effect of data augmentation and network simplification on the image-based detection of broccoli heads with mask r-cnn. *Journal of Field Robotics, 38*(1), 85–104. https://doi.org/10.1002/rob.21975

Bochkovskiy, A., Wang, C.-Y., & Liao, H.-Y. M. (2020). *Yolov4: Optimal speed and accuracy of object detection*. https://arxiv.org/abs/2004.10934.

Bolya, D., Zhou, C., Xiao, F., & Lee, Y. J. (2020). *Yolact++: Better real-time instance segmentation. IEEE transactions on pattern analysis and machine intelligence*. https://doi.org/10.1109/tpami.2020.3014297. page 1–1.

Boogaard, F. P., Rongen, K. S., & Kootstra, G. W. (2020). Robust node detection and tracking in fruit-vegetable crops using deep learning and multi-view imaging. *Biosystems Engineering, 192*, 117–132. https://doi.org/10.1016/j.biosystemseng.2020.01.023

Follmann, P., König, R., Härtinger, P., & Klostermann, M. (2018). *Learning to see the invisible: End-to-end trainable amodal instance segmentation*. https://arxiv.org/abs/1804.08864.

Ge, Y., Xiong, Y., & From, P. J. (2019). Instance segmentation and localization of strawberries in farm conditions for automatic fruit harvesting. *IFAC-PapersOnLine, 52*(30), 294–299. https://doi.org/10.1016/j.ifacol.2019.12.537, 6th IFAC Conference on Sensing, Control and Automation Technologies for Agriculture AGRICONTROL 2019.

Ge, Y., Xiong, Y., & From, P. J. (2020). Symmetry-based 3d shape completion for fruit localisation for harvesting robots. *Biosystems Engineering, 197*, 188–202.

Gongal, A., Karkee, M., & Amatya, S. (2018). Apple fruit size estimation using a 3d machine vision system. *Information Processing in Agriculture, 5*(4), 498–503. https://doi.org/10.1016/j.inpa.2018.06.002

Haas, T., Schubert, C., Eickhoff, M., & Pfeifer, H. (2020). Bubcnn: Bubble detection using faster rcnn and shape regression network. *Chemical Engineering Science, 216*, 115467. https://doi.org/10.1016/j.ces.2019.115467

He, K., Gkioxari, G., Dollár, P., & Girshick, R. (2017). Mask r-cnn. In *2017 IEEE international conference on computer vision (ICCV)* (pp. 2980–2988). https://doi.org/10.1109/ICCV.2017.322

Kamilaris, A., & Prenafeta-Boldú, F. X. (2018). Deep learning in agriculture: A survey. *Computers and Electronics in Agriculture, 147*, 70–90. https://doi.org/10.1016/j.compag.2018.02.016

Kanatani, K., & Rangarajan, P. (2011). Hyper least squares fitting of circles and ellipses. *Computational Statistics & Data Analysis, 55*(6), 2197–2208. https://doi.org/10.1016/j.csda.2010.12.012

Kang, H., & Chen, C. (2020). Fruit detection, segmentation and 3d visualisation of environments in apple orchards. *Computers and Electronics in Agriculture, 171*, 105302. https://doi.org/10.1016/j.compag.2020.105302

Klear, M. (2019). *Simple circle fitting library for Python*. https://github.com/AlliedToasters/circle-fit.

Kusumam, K., Krajník, T., Pearson, S., Duckett, T., & Cielniak, G. (2016). *3d datasets of broccoli in the field*. https://lcas.lincoln.ac.uk/nextcloud/shared/agritech-datasets/broccoli/broccoli_datasets.html.

Kusumam, K., Krajník, T., Pearson, S., Duckett, T., & Cielniak, G. (2017). 3d-vision based detection, localization, and sizing of broccoli heads in the field. *Journal of Field Robotics, 34*(8), 1505–1518. https://doi.org/10.1002/rob.21726

Lam, W. (2020). *ORCNN in Detectron2*. https://github.com/waiyulam/ORCNN.

Lehnert, C., Tsai, D., Eriksson, A., & McCool, C. (2019). 3d move to see: Multi-perspective visual servoing towards the next best view within unstructured and occluded environments. In *2019 IEEE/RSJ international conference on intelligent robots and systems (IROS)* (pp. 3890–3897). https://doi.org/10.1109/IROS40897.2019.8967918

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). Springer.

Lin, G., Tang, Y., Zou, X., Li, J., & Xiong, J. (2019). In-field citrus detection and localisation based on rgb-d image analysis. *Biosystems Engineering, 186*, 34–44. https://doi.org/10.1016/j.biosystemseng.2019.06.019

Liu, G., Nouaze, J. C., Mbouembe, P. L. T., & Kim, J. H. (2020). *Yolo-tomato: A robust algorithm for tomato detection based on yolov3* (Vol. 20). Basel, Switzerland): Sensors. https://doi.org/10.3390/s20072145

Nejati, M., Penhall, N., Williams, H., Bell, J., Lim, J., Ahn, H. S., & MacDonald, B. (2020). *Kiwifruit detection in challenging conditions*. https://arxiv.org/abs/2006.11729.

Qi, C. R., Liu, W., Wu, C., Su, H., & Guibas, L. J. (2018). *Frustum pointnets for 3d object detection from rgb-d data*. https://arxiv.org/abs/1711.08488.

Ren, S., He, K., Girshick, R., & Sun, J. (2017). Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 39*(6), 1137–1149. https://doi.org/10.1109/TPAMI.2016.2577031

Shi, W., van de Zedde, R., Jiang, H., & Kootstra, G. (2019). Plant-part segmentation using deep learning and multi-view vision. *Biosystems Engineering, 187*, 81–95. https://doi.org/10.1016/j.biosystemseng.2019.08.014

Wada, K. (2016). *labelme: Image polygonal annotation with Python*. https://github.com/wkentaro/labelme.

Wang, Z., Walsh, K. B., & Verma, B. (2017). On-tree mango fruit size estimation using rgb-d images. *Sensors, 17*(12), 2738. https://doi.org/10.3390/s17122738

Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometric Bulletin, 1*(6), 80–83. https://doi.org/10.2307/3001968

Wu, Y., Kirillov, A., Massa, F., Lo, W.-Y., & Girshick, R. (2019). *Detectron2*. https://github.com/facebookresearch/detectron2.

Xie, S., Girshick, R., Dollár, P., Tu, Z., & He, K. (2017). Aggregated residual transformations for deep neural networks. In 2017 *IEEE conference on computer vision and pattern recognition (CVPR)* (pp. 5987–5995). https://doi.org/10.1109/CVPR.2017.634

Yu, Y., Zhang, K., Yang, L., & Zhang, D. (2019). Fruit detection for strawberry harvesting robot in non-structural environment based on mask-rcnn. *Computers and Electronics in Agriculture, 163*, 104846. https://doi.org/10.1016/j.compag.2019.06.001

Zhang, Q., Liu, Y., Gong, C., Chen, Y., & Yu, H. (2020). Applications of deep learning for dense scenes analysis in agriculture: A review. *Sensors, 20*(5), 1520. https://doi.org/10.3390/s20051520

Zhu, Y., Tian, Y., Metaxas, D., & Dollár, P. (2017). Semantic amodal segmentation. In 2017 *IEEE conference on computer vision and pattern recognition (CVPR)* (pp. 3001–3009). https://doi.org/10.1109/CVPR.2017.320