The effect of data augmentation and network simplification on the image-based detection of broccoli heads with Mask R-CNN

Journal of Field Robotics

Blok, Pieter M.; Evert, Frits K.; Tielen, Antonius P.M.; Henten, Eldert J.; Kootstra, Gert

https://doi.org/10.1002/rob.21975

Received: 20 November 2019 | Revised: 27 April 2020 | Accepted: 14 June 2020

DOI: 10.1002/rob.21975

REGULAR ARTICLE

WILEY

The effect of data augmentation and network simplification on the image-based detection of broccoli heads with Mask R-CNN

Pieter M. Blok[1,2] | Frits K. van Evert[1] | Antonius P. M. Tielen[3] | Eldert J. van Henten[2] | Gert Kootstra[2]

[1]Agrosystems Research, Wageningen University & Research, Wageningen, The Netherlands

[2]Farm Technology Group, Wageningen University & Research, Wageningen, The Netherlands

[3]Greenhouse Horticulture, Wageningen University & Research, Wageningen, The Netherlands

Correspondence

Pieter M. Blok, Agrosystems Research, Wageningen University & Research, Droevendaalsesteeg 1, 6708 PB Wageningen, The Netherlands.
Email: pieter.blok@wur.nl

Funding information

Tony Wisdom (Skagit Valley Farm)

Abstract

In current practice, broccoli heads are selectively harvested by hand. The goal of our work is to develop a robot that can selectively harvest broccoli heads, thereby reducing labor costs. An essential element of such a robot is an image-processing algorithm that can detect broccoli heads. In this study, we developed a deep learning algorithm for this purpose, using the Mask Region-based Convolutional Neural Network. To be applied on a robot, the algorithm must detect broccoli heads from any cultivar, meaning that it can generalize on the broccoli images. We hypothesized that our algorithm can be generalized through network simplification and data augmentation. We found that network simplification decreased the generalization performance, whereas data augmentation increased the generalization performance. In data augmentation, the geometric transformations (rotation, cropping, and scaling) led to a better image generalization than the photometric transformations (light, color, and texture). Furthermore, the algorithm was generalized on a broccoli cultivar when 5% of the training images were images of that cultivar. Our algorithm detected 229 of the 232 harvestable broccoli heads from three cultivars. We also tested our algorithm on an online broccoli data set, which our algorithm was not previously trained on. On this data set, our algorithm detected 175 of the 176 harvestable broccoli heads, proving that the algorithm was successfully generalized. Finally, we performed a cost-benefit analysis for a robot equipped with our algorithm. We concluded that the robot was more profitable than the human harvest and that our algorithm provided a sufficient basis for robot commercialization.

KEYWORDS

agriculture, computer vision, learning, perception, sensors

1 | INTRODUCTION

In agriculture, numerous tasks depend on human labor. This labor is getting more expensive and more scarce, which causes problems for tasks that are done by hand, such as the selective harvest of crops. Selective hand-harvest involves the visual assessment of the crop, followed by the harvest of only those specimens that have reached the desired size, quality, or maturity. A crop that is selectively harvested by hand, is broccoli (*Brassica oleracea* var. *italica*). In the Netherlands, broccoli is usually hand-harvested three times in one growing season (Kwin, 2018). Cost studies show that the hand-harvest of broccoli can take up to 107 man-hours per hectare and

23% of the total production costs (Kwin, 2018). Motivated by the scarcity and the costs of human labor, broccoli growers search for alternative ways of selective harvesting. A promising alternative is an agricultural robot that can selectively harvest broccoli. A critical factor that hampers the development of a broccoli harvesting robot, is the lack of an automatic detection system that can replace human visual perception.

Several studies on the automatic detection of broccoli can be found in literature. Ramirez (2006) was the first who detected broccoli heads, using Red–Green–Blue (RGB) color images and texture-based analysis. Unfortunately, the data set of Ramirez (2006) was limited to 13 RGB images, which is too small to draw a conclusion on the applicability of the algorithm in the open field conditions. Blok, Barth, and van den Berg (2016) used a Laws' texture filter on RGB images to detect broccoli heads from two different cultivars. They included an additional color analysis for the maturity evaluation. Despite a promising precision of 99.5%, the researchers observed a recall of 91.2%, which corresponded to 20 false-negatives on 228 broccoli heads. The false-negatives were caused by the fixed thresholds on the texture and the color features that could not generalize sufficiently on broccoli heads whose texture or color differed from the chosen thresholds. Generalization is a common challenge in image analysis, and includes the ability of an algorithm to perform on new images (Goodfellow, Bengio, & Courville, 2016).

Machine learning can provide better image generalization than threshold-based algorithms (Kamilaris & Prenafeta-Boldú, 2018). Kusumam, Krajník, Pearson, Duckett, and Cielniak (2017) detected broccoli heads in RGB-Depth (RGB-D) images with three-dimensional (3D) vision using a viewpoint feature histogram (VFH), a support vector machine (SVM) classifier and a temporal filter. The average precision (AP) was 95.2% on 600 images of the broccoli cultivar Ironman and the AP was 84.5% on 1169 images of the broccoli cultivar Titanium, indicating that their algorithm did not generalize sufficiently on images of different broccoli cultivars. A limitation of Kusumam et al. (2017) is that their machine-learning algorithm was based on a predefined set of image features whose generalization capability was found to be limited on images of different broccoli cultivars.

Image generalization can be further improved with deep learning. A deep learning network that is commonly used for image analysis, is a convolutional neural network (CNN). CNNs internally optimize the feature extraction during training (LeCun, Bengio, & Hinton, 2015). Kamilaris and Prenafeta-Boldú (2018) showed that CNNs outperformed predefined feature-engineered machine learning in all 22 agricultural case studies. Bender, Whelan, and Sukkarieh (2019) researched broccoli and cauliflower detection with Faster Region-based CNN (Faster R-CNN; Ren, He, Girshick, & Sun, 2017) and reported a promising 95% mean average precision (mAP). Unfortunately, this study focused on individual plant detection and did not investigate the broccoli head detection, which is essential for the selective harvest. Jiang, Shuang, Li, Paterson, and Robertson (2018) showed that the detection performance on cabbage and cauliflower was almost doubled when using a Mask Region-based CNN (Mask R-CNN) instead of a threshold-based algorithm. Mask R-CNN (He, Gkioxari, Dollár, & Girshick, 2017) is an upgrade of Faster R-CNN and performs instance segmentation (a combination of object detection and pixel segmentation). Mask R-CNN allows the instance-aware segmentation of distinct objects even if they are overlapping or occluded by other objects (Romera-Paredes & Torr, 2016). Instance-aware segmentation is a desirable feature for the precise size measurement of broccoli, because broccoli heads can be partially occluded by leaves. Therefore, we focused our research on Mask R-CNN.

To generalize Mask R-CNN on the broccoli images, the network must not overfit during training. Network overfitting occurs when an overly complex model is fitted on the training data set and the model fails to generalize on new data (Rosebrock, 2018). Network overfitting can be resolved with regularization. Regularization involves any modification to a learning algorithm that reduces the generalization error, possibly at the expense of increased training error (Goodfellow et al., 2016). There are two types of regularization: explicit and implicit. Explicit regularization involves alterations to the network architecture that constrain the capacity of the neural network. Common explicit regularization methods are drop-out (random disconnection of neurons), weight decay (penalizing large weights), and network simplification (removal of network layers). Implicit regularization is applied during the training process without constraining the capacity of the neural network. Two examples of implicit regularization are early stopping and data augmentation. Early stopping is the termination of the training process whenever the generalization error increases (the generalization error is the difference between the training and the validation error). Data augmentation involves a wide range of image synthesis techniques that generate new training samples from the original ones by applying image transformations. With data augmentation, the network is trained on constantly changing versions of the input images, allowing the network to learn more robust features.

Most regularization research solely focused on data augmentation (Perez & Wang, 2017; Shijie, Ping, Peiyi, & Siping, 2017; Zhu, Aoun, Krijn, Vanschoren, & Campus, 2018). Hernández-García and König (2018) studied the combined effect of drop-out, weight decay, and data augmentation and found that data augmentation led to the highest increase in accuracy. However, their research investigated All-CNN (Springenberg, Dosovitskiy, Brox, & Riedmiller, 2014) and wide residual network (WRN; Zagoruyko & Komodakis, 2016) that have architectures that are less complex than Mask R-CNN. The higher complexity of the Mask R-CNN network might imply the need for other regularization strategies. In our research, we studied the effects of network simplification and data augmentation on the image generalization of Mask R-CNN.

We hypothesized that through network simplification and data augmentation, Mask R-CNN can be generalized on images of multiple broccoli cultivars. The first objective of our study was to test this hypothesis using images of three broccoli cultivars taken with a prototype broccoli harvesting robot. The primary contribution of our research is a quantitative analysis of the effect of network simplification and data augmentation on the image generalization of Mask R-CNN.

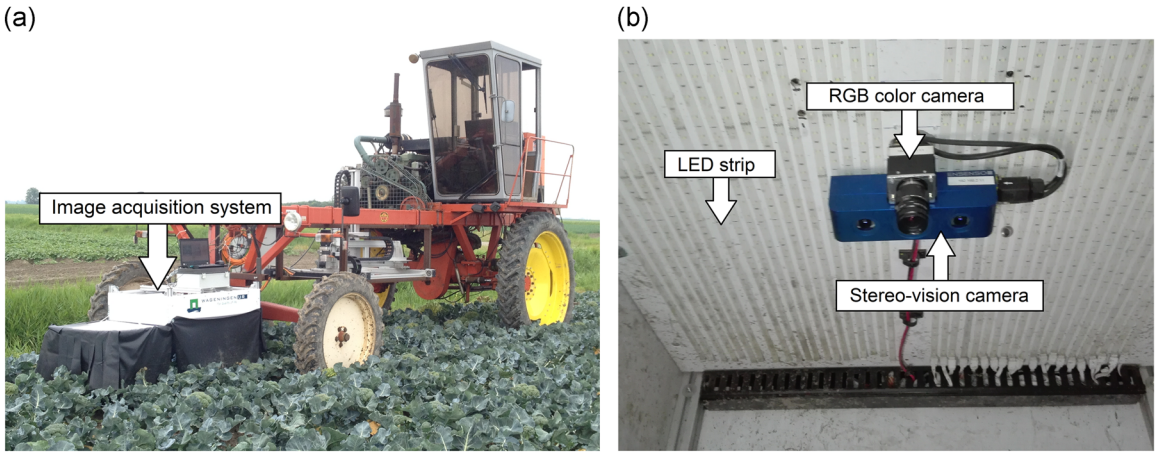

FIGURE 1 (a) Overview of the image acquisition system that was attached to the prototype robot to acquire broccoli images in the Netherlands. (b) The Dutch image acquisition system consisted of one RGB color camera, one stereo-vision camera, and 40 LED strips for artificial illumination. LED, light-emitting diode; RGB, red–green–blue [Color figure can be viewed at wileyonlinelibrary.com]

Eventually, a robot leads to new benefits and new costs. The benefits derive primarily from the savings on labor costs. The csts derive primarily from the robot investment. If the benefits are higher than the costs, then there is a basis for robot commercialization. The second objective of our study was to perform a cost-benefit analysis for a selective broccoli harvesting robot equipped with our Mask R-CNN algorithm. The secondary contribution of our research is a cost-benefit analysis for a selective broccoli harvesting robot that has to work in the field.

2 | MATERIALS AND METHODS

2.1 | Image data set

This section highlights the image acquisition systems that were used (Section 2.1.1), the broccoli images that were acquired in the field (Section 2.1.2), how these images were annotated (Section 2.1.3), the feature variability between images of different broccoli cultivars (Section 2.1.4) and how the annotated images were aggregated for Mask R-CNN training and testing (Section 2.1.5).

2.1.1 | Image acquisition systems

We used a prototype robot that consisted of an image acquisition system that acquired top view images of one row of the broccoli crop. Two different image acquisition systems were used, because the robot was first tested in the Netherlands (Figure 1a) and then in the United States (Figure 2a). Although both systems were constructed as an enclosed box for uniform illumination, they had a different RGB color camera and light-emitting diode (LED) illumination (Table 1). In both systems, the white balance of the color

FIGURE 2 (a) Overview of the image acquisition system that was attached to the prototype robot to acquire broccoli images in the United States. (b) The U.S. image acquisition system consisted of one RGB color camera, one stereo-vision camera, and 21 LED strips for artificial illumination. LED, light-emitting diode. RGB, red–green–blue [Color figure can be viewed at wileyonlinelibrary.com]

TABLE 1 Overview of the cameras and LEDs that were used to acquire images in the Netherlands and the United States

	The Netherlands	The United States
Camera specifications		
Camera	AVT Prosilica GC2450	IDS UI-5280FA-C-HQ
Image resolution (pixels)	2448 × 2050	2456 × 2054
Lens	Kowa LM12JCM	Fujifilm HF8XA-5M
Focal length (mm)	12	8
Field of view (m) at 0.5 m	0.53 × 0.44	0.62 × 0.52
Image scene overlap (%)	66	71
LED specifications		
LED	Paulmann 70209 YourLED	OSRAM VFP2400S-G3-865-03
Number of LED strips	40	21
Color temperature (K)	6000	6500
Luminous flux (lm)	13,500	144,900

Abbreviation: LED, light-emitting diode.

cameras was set and fixed with a color calibration plate (X-Rite ColorChecker Classic). A stereo-vision camera (IDS Ensenso N35) was installed to acquire depth images to estimate the size of the broccoli heads (Figures 1b and 2b). In both systems, the color and the stereo-vision camera were hardware triggered by 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.

2.1.2 | Broccoli images

With the Dutch image acquisition system (Figure 1a,b), nearly 14,000 broccoli images of the Ironman cultivar and 13,000 broccoli images of the Steel cultivar were captured on nine different broccoli fields in the province of Friesland in four consecutive years (2014–2017). On all fields, the broccoli plants were grown in single rows that were 0.75 m apart. The intra-row spacing was 0.33 m. With the American image acquisition system (Figure 2a,b), 14,000 broccoli images of the Emerald-Crown cultivar were acquired in 2018 on one broccoli field in Washington State. On this field, the broccoli plants were grown in single rows that were 0.61 m apart. The intra-row spacing was 0.23 m.

2.1.3 | Ground-truth annotation

From the image data set, 3000 images with broccoli heads were randomly selected, 1000 images for each cultivar. Two instructed research assistants annotated all 4706 broccoli heads that were found in the 3000 images. All broccoli heads were annotated, regardless of their size, to make the annotations suitable for other computer vision tasks. Overripe broccoli heads, which were visually distinguishable by yellow-colored flower buds, were not annotated because these broccoli heads do not have to be harvested. The annotation process involved three steps (Figure 3). First, a bounding box was drawn for each broccoli head using labelImg (T. Lin, 2019). Second, the bounding box was zoomed to full screen to allow the precise pixel annotation of the contour of the broccoli head. Third, the pixel annotations were placed back into the original bigger image to generate the binary masks for each image. When finished, another research assistant independently validated and if necessary corrected the bounding boxes and the masks.

2.1.4 | Image feature similarity

To calculate the image feature similarity, we first resized the 3000 images with zero-padding to a resolution of 1024 × 1024 pixels. This resolution equals to the processing resolution of Mask R-CNN (Abdulla, 2017). Then, we quantified the image feature similarity between the broccoli heads using three color and four texture features. The color features were hue, saturation, and lightness (HSL), and the texture features were energy, correlation, homogeneity, and contrast that were all extracted from the Grey-Level Co-occurrence Matrix (GLCM; Haralick, Shanmugam, & Dinstein, 1973). The three-color features and four texture features corresponded to the features that were used in related research (Blok et al., 2016; Ramirez, 2006). We calculated the average feature value for each annotated broccoli head. A normalized histogram with 101 bins was generated for every feature. As such, seven feature histograms were generated per broccoli cultivar. Then, we quantified the level of similarity between the feature histograms of the three cultivars using the χ^2 distance (Equation 1). A low χ^2 distance indicates a high feature-similarity between the broccoli heads of two cultivars, and a high chi-squared distance indicates a low feature-similarity.

$$\chi^2 = \frac{1}{2}\sum_{i=1}^{n}\frac{(p(i)-q(i))^2}{p(i)+q(i)}, \quad (1)$$

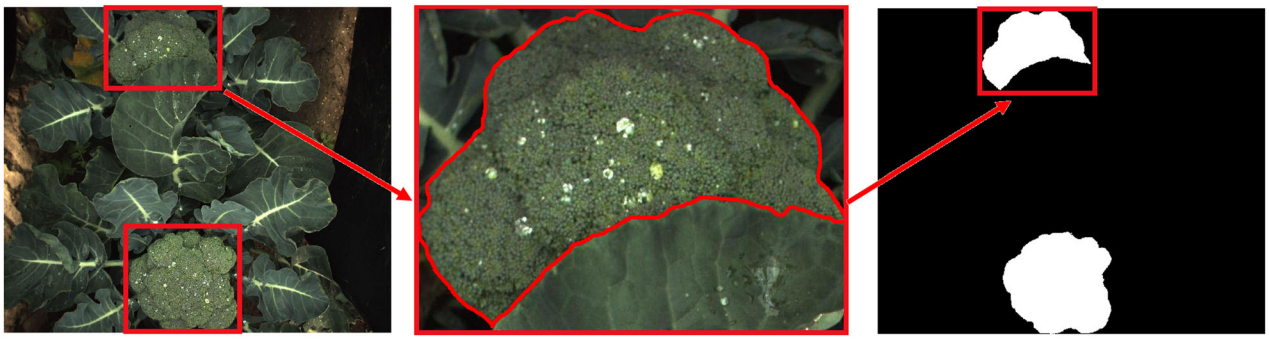

FIGURE 3 Ground-truth pixel annotation of the RGB broccoli images. First, a bounding box (red rectangle) was drawn to encapsulate the region of each broccoli head. Then, the bounding box was zoomed to full screen to allow precise contour delineation (red line) of the broccoli head. After contour-closing, the inner area was automatically filled with pixels and placed back into the original image to generate the binary masks. The white pixels depict the broccoli head and the black pixels the background. RGB, red–green–blue [Color figure can be viewed at wileyonlinelibrary.com]

where p is the feature histogram of one cultivar and q the feature histogram of another cultivar. n is the number of histogram bins.

Table 2 shows that the broccoli heads of the Ironman cultivar and the Steel cultivar were similar in texture, but different in color (hue; see Figure 4 for some examples). The broccoli heads of Ironman and Emerald-Crown were similar in color (hue), but different in texture. Broccoli heads of Steel and Emerald-Crown were the least similar and differed in both texture and color (hue).

2.1.5 | Train, validation, and test sets

After the image feature inspection, we randomly divided each cultivar subset of 1000 images into one training set of 600 images, one validation set of 100 images and three test sets of 100 images each. The training set was used to adjust the network weights of Mask R-CNN during training. The validation set was used during training to control the network's learning rate to minimize the chance of overfitting (see Section 2.2.2). The three test sets were completely independent of the training process and were used in four different experiments to evaluate the performance of the network. We used three different test sets, because the outcome of an experiment influenced the choice of the algorithm in the next experiment (see Section 2.5).

Besides the three test sets, we extracted an independent test set of 300 broccoli images from the online data set of Bender et al. (2019; https://doi.org/10.25910/5c941d0c8bccb). The images of Bender et al. (2019) were acquired on a weekly basis on one broccoli field in Australia (New South Wales), where the broccoli plants were grown in single lines (see Figure 5a,b for two examples). Bender et al. (2019) did not report the cultivar of the broccoli crop, so we did not train our algorithm on these images, but only used them to test the generalization of our trained algorithm. All 399 broccoli heads in the 300 images were annotated, regardless of their size, using the procedure of Section 2.1.3.

2.2 | Mask R-CNN

This section highlights the network architecture of Mask R-CNN (Section 2.2.1), the Mask R-CNN software that was used (Section 2.2.2), and the Mask R-CNN training methodology (Section 2.2.3).

2.2.1 | Network architecture

Mask R-CNN (He et al., 2017) is a network that consists of multiple branches (Figure 6). 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 (FPN). Usually, the backbone is a variant of the Resnet residual network (He, Zhang, Ren, & Sun, 2016). After the backbone, there is a region proposal network

TABLE 2 The χ^2 distance of the three color features and the four texture features between the broccoli heads of two cultivars

	χ^2 distance of the color features			χ^2 distance of the GLCM-texture features			
Cultivar comparison	Hue	Saturation	Lightness	Energy	Correlation	Homogeneity	Contrast
Ironman–Steel	0.91	0.16	0.50	**0.01**	**0.04**	**0.06**	**0.03**
Ironman–Emerald Crown	**0.09**	0.68	0.84	0.45	0.48	0.61	0.59
Steel–Emerald Crown	0.92	0.74	0.91	0.47	0.39	0.60	0.60

Note: The bold values represent χ^2 distances lower than 0.1, which correspond to a high feature similarity. GLCM, Grey-Level Co-occurrence Matrix.

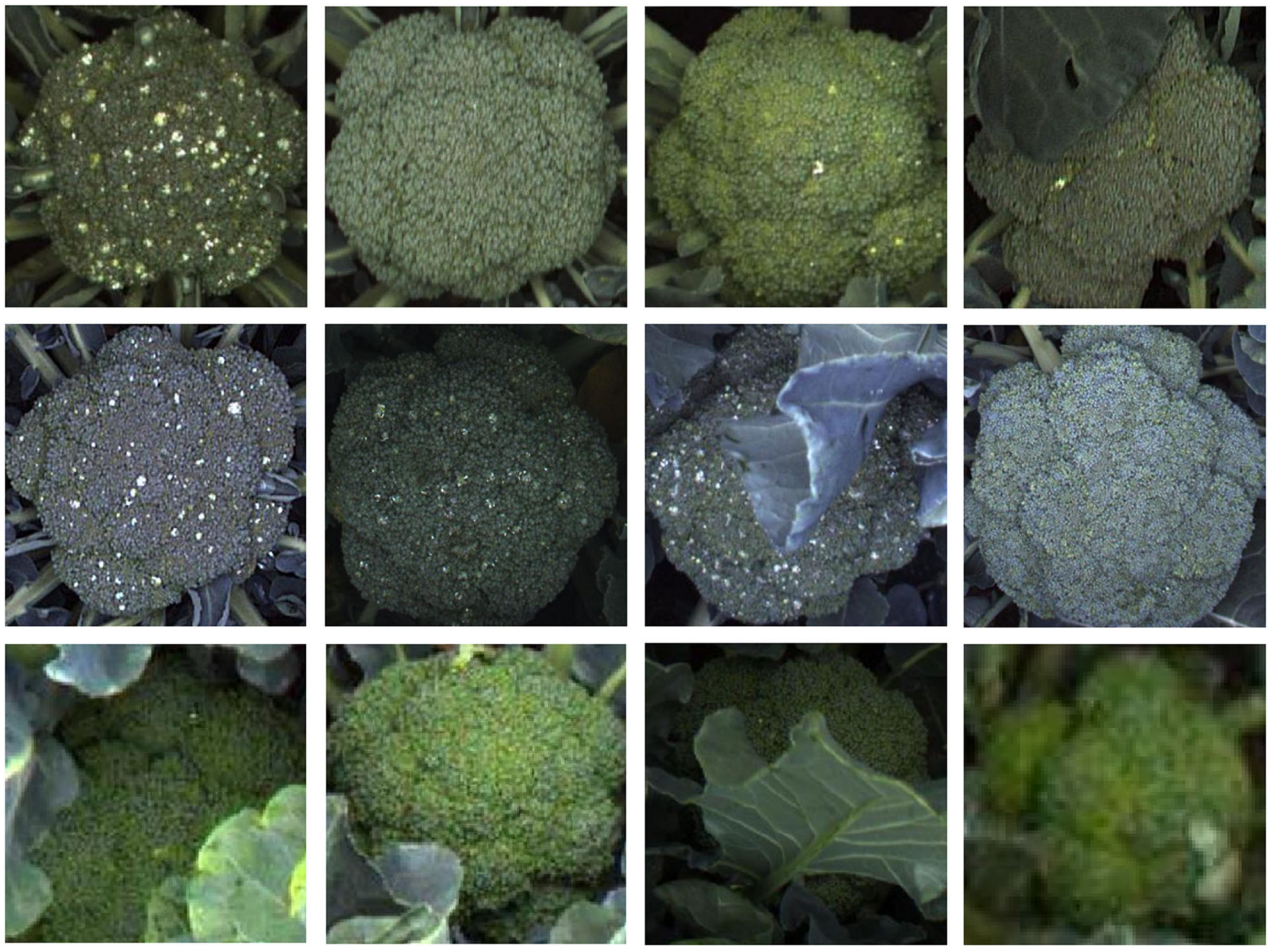

FIGURE 4 Mosaic of four random broccoli heads for each cultivar (top row: Ironman, middle row: Steel, bottom row: Emerald-Crown). The broccoli head images were cropped from a bigger image (see an example in Figure 3). The broccoli heads had different sizes and some were occluded by leaves. The pixel quality of some of the small-sized broccoli heads was less than the pixel quality of the harvestable broccoli heads because the small-sized heads were deeper into the crop and more remote from the camera (an example is a right image on the bottom row) [Color figure can be viewed at wileyonlinelibrary.com]

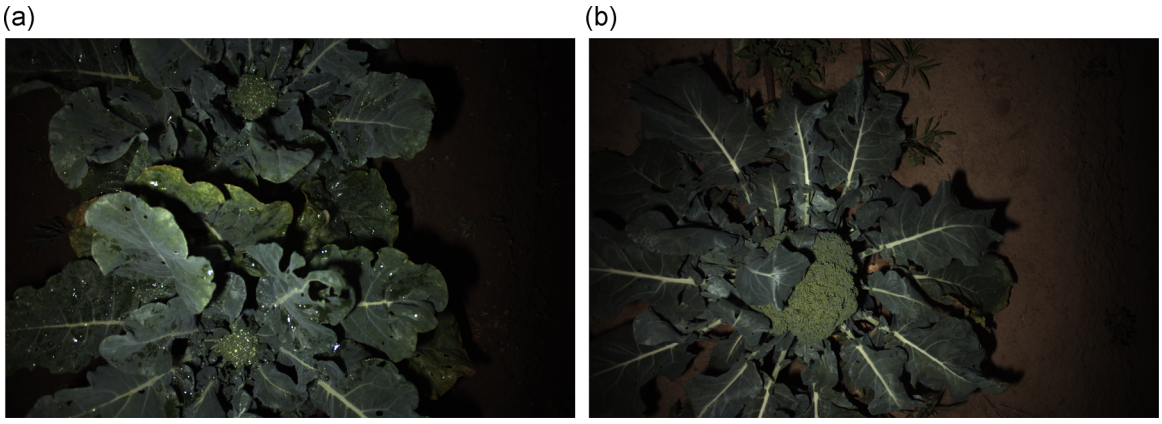

FIGURE 5 (a) An image from a wet broccoli crop taken from the data set of Bender et al. (2019). (b) An image from a dry broccoli crop taken from the data set of Bender et al. (2019) [Color figure can be viewed at wileyonlinelibrary.com]

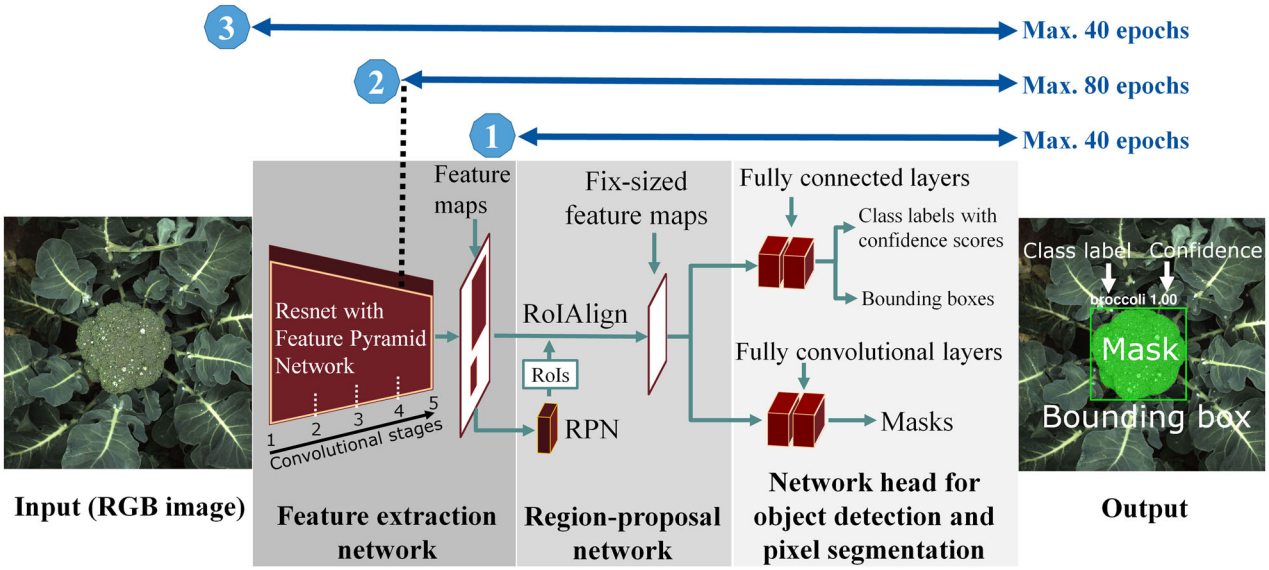

FIGURE 6 Schematic representation of the neural network architecture of Mask R-CNN (image adapted from Shi, van de Zedde, Jiang, & Kootstra, 2019). The numbers 1–3 indicate the training stages that were used to train the network (see Section 2.2.3). Mask R-CNN, Mask Region-based convolutional neural network [Color figure can be viewed at wileyonlinelibrary.com]

(RPN) that proposes regions of interest (ROI) of distinct objects from the feature maps. To avoid duplicate ROIs for the same object, non-maximum suppression (NMS) is used to discard the ROIs that overlap with a more confident ROI. The ROIs that remain after the NMS are realigned with the ROI Align layer. Then, the ROIs are transformed into fix-sized feature maps, which are further processed in two parallel branches in the so-called network head. The first branch has two fully connected (FC) layers, of which one performs object classification and the other bounding box refinement by regression. The second branch has two fully convolutional layers that segment the object pixels inside the bounding box, yielding the mask (Figure 6).

2.2.2 | Software

For our research, we used the code of the online Mask R-CNN repository of Matterport (version 2.1; Abdulla, 2017). The Mask R-CNN code was installed on a computer with an Intel Core i9-7940X processor (64GB DDR4 RAM) and two 12GB graphical processing units (GPU; NVIDIA GeForce RTX 2080 Ti). The operating system of the computer was Ubuntu Linux (version 16.04). Tensorflow (GPU version 1.7.0), CUDA (version 9.0), and CUDNN (version 7.0.5) were used as computational backend. The Mask R-CNN code was deployed in Python (version 3.6) using the Keras library (version 2.1.6) for the deep learning. We added two additional Keras functions to the code to minimize the network overfitting. Network overfitting occurs when the network weights are too specifically parametrized on the images of the train set, making it harder to generalize on the images of the independent validation set, leading to an increase in the validation loss (the loss summarizes the classification, localization and segmentation error). The two Keras functions automatically detected this increase in validation loss and then tried to resolve the overfitting. The first function ("ReduceLROnPlateau"; Keras, 2019) automatically lowered the learning rate by a factor of two whenever the validation loss increased during five consecutive epochs (the learning rate was not allowed to become smaller than 10^{-4}). The second function ("EarlyStopping"; Keras, 2019) was used to automatically stop the training process when the validation loss increased during ten consecutive epochs. An epoch is one complete network training pass through the entire training data set. Finally, the code was altered so that Mask R-CNN performed a binary classification on our required classes (broccoli head and background).

2.2.3 | Network training

Mask R-CNN was trained with the stochastic gradient descent using an image batch size of one. We also used transfer-learning to initialize the network weights of Mask R-CNN with the weights of another Mask R-CNN that was trained on a different data set. This transfer-learning allowed us to use the learned feature maps from the other Mask R-CNN algorithm so that our Mask R-CNN could be effectively trained on our broccoli data set. The transfer-learning was done with a Mask R-CNN that was trained on the Microsoft Common Objects in Context (COCO) data set (T.-Y. Lin et al., 2014) that also contained a broccoli class. However, we could not use this broccoli class directly, because the COCO images contained broccoli heads in dishes instead of broccoli heads in the field.

We trained Mask R-CNN in three stages (Figure 6), similar to Abdulla (2017). In the first training stage, only the upper layers of

Mask R-CNN (RPN and network head) were trained for a maximum of 40 epochs (depending on the "EarlyStopping" function). In the second training stage, the upper layers (RPN and network head) were trained together with the upper half of the Resnet backbone, which included the fourth and fifth convolutional Resnet stage (Figure 6). The second stage was trained for a maximum of 80 epochs (depending on the "EarlyStopping" function). In the third training stage, the complete Mask R-CNN network was trained for a maximum of 40 epochs (depending on the "EarlyStopping" function). With this three-staged training methodology, we gradually optimized the feature layers of the COCO transfer-learned Mask R-CNN to our own data set. In each training stage, the initial learning rate was 10^{-3}. This initial learning rate could be automatically lowered until 10^{-4} depending on the "ReduceLROnPlateau" function. The weight decay (L2-regularization) was set to 10^{-4}.

2.3 | Network simplification and data augmentation

We focused our research on network simplification (Section 2.3.1) and data augmentation (Section 2.3.2).

2.3.1 | Network simplification

We investigated network simplification by simplifying a deep residual backbone with 101 hidden layers (Resnet101) to a shallower version of the same residual backbone with 50 hidden layers (Resnet50). For both residual backbones, the transfer learning was employed with a Resnet50 or a Resnet101 that was trained on the Microsoft COCO data set.

2.3.2 | Data augmentation

We investigated three different types of data augmentation and compared it to no data augmentation at all. The first data augmentation consisted of three geometric transformations: image rotation, image cropping/partitioning and image scaling (Figure 7). These geometric transformations, hereinafter referred as "G," were reported as the most common data augmentation for deep learning in agriculture (Kamilaris & Prenafeta-Boldú, 2018). The second data augmentation consisted of four photometric transformations: light transformations, color transformations, texture enhancement, and texture blur (Figure 7). These transformations, hereinafter referred as "P," tried to resolve the dissimilarity of the color and the texture features between the three broccoli cultivars (Table 2). The third data augmentation combined the three geometric transformations and the four photometric transformations and will be referred as "GP."

With data augmentation, each training image was transformed with a randomly chosen transformation from the augmentation set of G, P, or GP, using the Imgaug software library (version 0.2.8; Jung, 2019). The Imgaug operators were parameterized so that the transformed images represented visually realistic images (Figure 7). For each type of data augmentation, 600 transformed images were randomly created during each training's epoch and these images were used for training (the original images were not trained). In case of no data augmentation, then the 600 original images were used during each training's epoch.

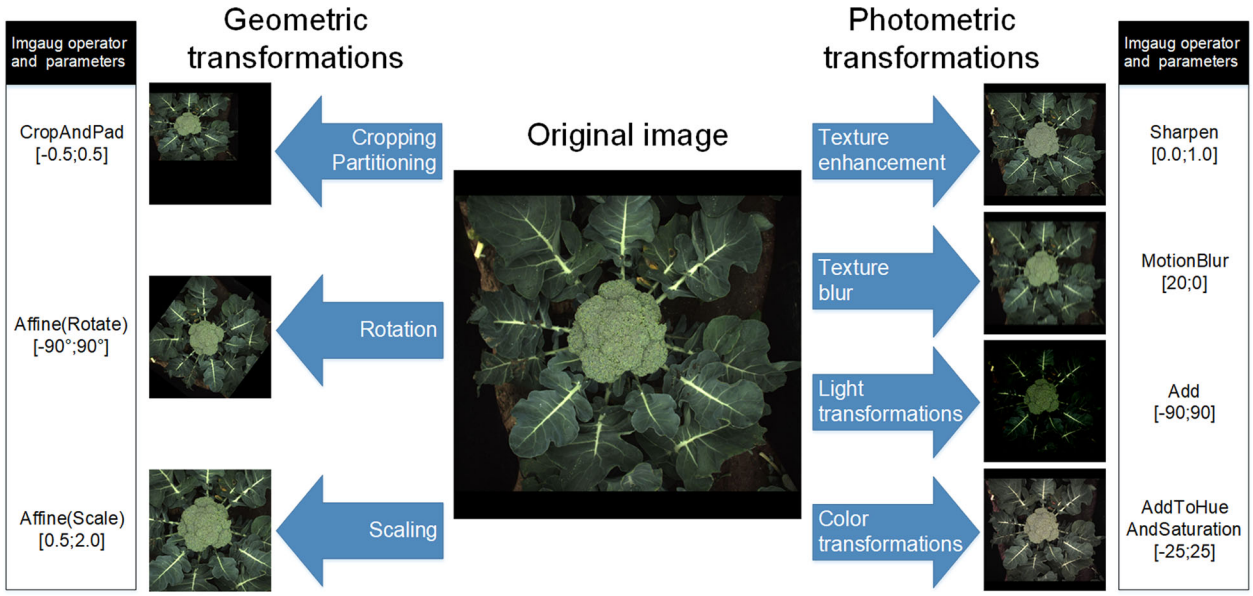

FIGURE 7 Examples of the three geometric (left) and the four photometric (right) image transformations and how these transformations were parameterized with the Imgaug operators in the data augmentation of Mask R-CNN. Mask R-CNN, Mask Region-based convolutional neural network [Color figure can be viewed at wileyonlinelibrary.com]

2.4 | Experimental setup

We have set up four experiments. In the first three experiments, we studied the effects of network simplification and data augmentation on the image generalization of Mask R-CNN. In the fourth experiment, we calculated the costs and the benefits of a selective broccoli harvesting robot equipped with Mask R-CNN.

2.4.1 | Experiment 1

The objective of Experiment 1 was to determine the effect of network simplification and data augmentation on the cultivar-specific segmentation. Cultivar-specific means that Mask R-CNN was trained and tested on images of the same cultivar. In Experiment 1, Mask R-CNN was trained with eight combinations of the two Resnet backbones (Resnet50 and Resnet101) and the four data augmentations (No, G, P, and GP; Figure 8).

2.4.2 | Experiment 2

The objective of Experiment 2 was to determine if network simplification and data augmentation could help to generalize Mask R-CNN on images of broccoli cultivars that were not trained. We used the trained networks of Experiment 1 and tested them on the images of the other two broccoli cultivars that were not incorporated in the training (Figure 8). We called this cultivar-generic segmentation. We determined that image generalization was reached when the segmentation performance did not deviate more than 5% from the cultivar-specific segmentation of Experiment 1. We chose the 5% threshold, because this value allowed us to improve the generalization error by more than 50% compared to the current broccoli head detection algorithms of Blok et al. (2016) and Kusumam et al. (2017) (who both experienced a performance loss of 10% when testing their algorithms on two different broccoli cultivars).

2.4.3 | Experiment 3

The objective of Experiment 3 was to investigate how many cultivar-specific training images were needed to generalize Mask R-CNN on images of that cultivar. This is useful when Mask R-CNN has to be applied on a new cultivar. In Experiment 3, the training was done on ten mixed data sets of the three cultivars (Table 3). These 10 data sets represented seven different percentages of cultivar-specific images (for each cultivar): 5%, 10%, 25%, 33.3%, 50%, 80%, and 90%. For the validation sets, we used the same percentages. The remaining parts of the data sets contained the images of the other two cultivars (Table 3). Because Experiments 1 and 2 delivered data when training on 0% cultivar-specific images (Experiment 2) and 100% cultivar-specific images (Experiment 1), we were able to investigate nine different percentages of cultivar-specific images (0%, 5%, 10%, 25%, 33.3%, 50%, 80%, 90%, and 100%). Like Experiment 2,

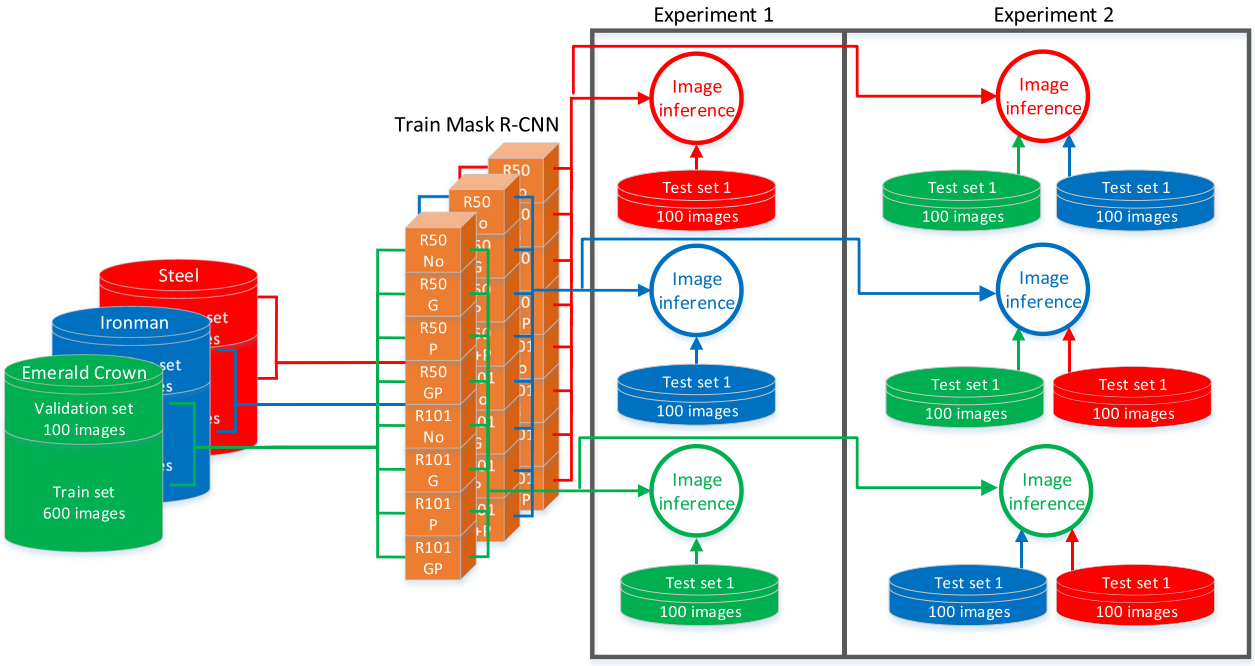

FIGURE 8 In Experiment 1, Mask R-CNN was trained and tested on images of the same cultivar (cultivar-specific). In Experiment 2, the trained networks of Experiment 1 were tested on the images of the two cultivars that were not trained (cultivar-generic). In total, Mask R-CNN was trained with eight different combinations of the two Resnet backbones (Resnet50 and Resnet101) and the four data augmentations (No, G, P, and GP). G, geometric transformations; GP, geometric and photometric transformations; Mask R-CNN, Mask Region-based convolutional neural network; No, no data augmentation; P, photometric transformations [Color figure can be viewed at wileyonlinelibrary.com]

TABLE 3 In Experiment 3, Mask R-CNN was trained on 10 mixed data sets of the three broccoli cultivars (Emerald-Crown, Ironman, and Steel)

	Number and percentage of cultivar-specific training images			
Mixed data set	Emerald-Crown	Ironman	Steel	Total images
1	30 (5%)	30 (5%)	540 (90%)	600
2	60 (10%)	60 (10%)	480 (80%)	600
3	150 (25%)	150 (25%)	300 (50%)	600
4	200 (33.3%)	200 (33.3%)	200 (33.3%)	600
5	300 (50%)	150 (25%)	150 (25%)	600
6	480 (80%)	60 (10%)	60 (10%)	600
7	540 (90%)	30 (5%)	30 (5%)	600
8	30 (5%)	540 (90%)	30 (5%)	600
9	60 (10%)	480 (80%)	60 (10%)	600
10	150 (25%)	300 (50%)	150 (25%)	600

we determined that image generalization was reached when the segmentation performance did not deviate more than 5% from the cultivar-specific performance of Experiment 1. In Experiment 3, Mask R-CNN was solely trained with the method that had the highest performance in the first two experiments.

2.4.4 | Experiment 4

The objective of Experiment 4 was to calculate the costs and the benefits of a robot equipped with Mask R-CNN. This experiment was done with harvestable broccoli heads only, because these heads need to be picked by the robot. The selection criteria for harvestable broccoli heads was obtained from Seminis (2019) that determined that a broccoli head is saleable and thus harvestable when its diameter is between 10 and 15 cm. In this experiment, the Mask R-CNN algorithm was used that had the highest performance in Experiment 3.

2.5 | Evaluation

In the first three experiments, the segmentation performance of Mask R-CNN was evaluated on all broccoli heads, regardless of their size (Section 2.5.1). In the fourth experiment, the Mask R-CNN detection performance was evaluated on the harvestable broccoli heads only (Section 2.5.2). The detection metrics of experiment 4 were used in the cost-benefit analysis (Section 2.5.3).

In all experiments, we used an NMS threshold of 10^{-3}. This threshold removed all ROIs that overlapped with a more confident ROI. We chose this threshold, because broccoli heads do not normally overlap as they grow solitary.

2.5.1 | Evaluation metrics for the broccoli head segmentation

In the first three experiments, the segmentation performance was obtained via two metrics: the algorithm's confidence level and the intersection over union (IoU). A threshold on the confidence level determined whether there was a segmentation (confidence ≥ threshold) or not (confidence < threshold). A threshold on the IoU determined whether a segmentation was broccoli (IoU ≥ threshold) or background (IoU < threshold). The IoU is a measure for the pixel overlap between the ground truth mask and the predicted mask and varies between zero (no overlap) and one (full overlap). With both thresholds on the confidence level and the IoU, we determined the number of true-positives (TPs; confidence ≥ threshold and IoU ≥ threshold), false-positives (FPs; confidence ≥ threshold and IoU < threshold) and false-negatives (FNs; confidence < threshold or IoU < threshold). A TP was a broccoli that was segmented as broccoli. A FP was background that was segmented as broccoli. A FN was a broccoli that was not segmented.

The ratio of TPs, FPs, and FNs determined the precision (Equation 2) and the recall (Equation 3). The precision was the percentage of correct segmentations. The recall measured how well Mask R-CNN was able to detect and segment all object pixels. Both the precision and the recall originated from one threshold on the confidence level and the IoU. With this single set of thresholds, the precision and recall did not express whether the segmentation was precisely located or not. Therefore, we used the mAP, which was calculated by averaging the precision over 101 recall values (0.0–1.0, in 0.01 steps) and 10 IoU values (0.5–0.95, in 0.05 steps; Coco, 2019). The mAP resulted a value close to zero when the segmentation was not precisely located, and a value close to one when the segmentation was precisely located.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \tag{2}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}. \tag{3}$$

In Experiment 1, the mAP was calculated for each test image of the cultivar that was previously trained. Each cultivar had 100 test images in the first test set, thus 300 cultivar-specific mAPs were calculated (Figure 8). A pairwise Wilcoxon's test (Wilcoxon, 1992) with a significance level of 5% was employed for these 300 mAPs to test whether there were statistical differences between the eight training methods. We used the Wilcoxon's test, because it can deal with non-normal distributed data, like the mAP.

In Experiment 2, the mAP was calculated for each test image of the two cultivars that were not trained upon. Again, the test images of first test set were used, resulting in 600 cultivar-generic mAPs (Figure 8). A pairwise Wilcoxon's test with a significance level of 5% was employed for these 600 mAPs to check for significant differences.

In Experiment 3, we used the images of the second test set that were independent of the images of the first test set. This was done, because the results on the first test set determined the choice for the training method in Experiment 3. We calculated the mAPs for each percentage of cultivar-specific images. The cultivar-specific images between 5% and 25% resulted in twice as many mAPs, because these percentages had two differently trained Mask R-CNNs. For example, the 5% mix of cultivar X had one Mask R-CNN trained with the 5%–90%–5% mix and one with the 5%–5%–90% mix (Table 3). These two Mask R-CNNs resulted in 200 mAPs when tested on the test images of cultivar X. For the percentages higher than 25%, only one Mask R-CNN was trained (Table 3), resulting in 100 mAPs per cultivar. Because each cultivar was tested, there were 600 mAPs for the cultivar-specific images between 0% and 25%, and 300 mAPs for the cultivar-specific images between 33.3% and 100%. Due to the difference in calculated mAPs, we employed a summary statistics analysis instead of a Wilcoxon's test.

2.5.2 | Evaluation metrics for the detection of harvestable broccoli heads

In Experiment 4, two different data sets were used to test the Mask R-CNN detection performance on the harvestable broccoli heads. The first data set consisted of 300 images from the third test set. These images were independent of the images of the second test set, because the results on the second test set determined the choice for the Mask R-CNN algorithm in Experiment 4. The second data set consisted of the 300 images that were taken from the online data set of Bender et al. (2019).

From both data sets, we selected the harvestable broccoli heads based on their estimated size. The size was estimated from the world coordinates of the depth images that were obtained from the stereo-vision images (both our data set and Bender's data set contained image pairs from stereo-vision cameras). Because 54% of the broccoli heads (482 from the 892) were partially occluded by leaves, we used the biggest side of the annotated bounding-box (in world coordinates) as a measure for the diameter. Broccoli heads with a diameter between 100 and 150 mm were classified as harvestable.

Mask R-CNN was evaluated on its ability to detect the 408 harvestable broccoli heads that were found in both datasets. The number of TPs, FPs, and FNs were obtained by using a confidence threshold of 0.99 and an IoU threshold of 0.5. The IoU threshold value was considered as the minimum pixel overlap to allow the end-effector to successfully cut a harvestable broccoli head. With the number of TPs, FPs, and FNs, we calculated the precision (Equation 2) and the recall (Equation 3).

2.5.3 | Cost-benefit analysis

With the detection metrics on the 408 harvestable broccoli heads (Experiment 4), we estimated the tentative costs and benefits of the robot, using Equation (A1) till Equation (A6) (appendix). We performed the cost-benefit analysis with the assumption that the robot could harvest four single-line rows of broccoli in one pass (under Dutch growing conditions). The cost parameters are summarized in Table A1 (appendix). We extracted the cost parameters from three cost studies (AgriConnect, 2019; Edwards, 2019; Kwin, 2018) and an analogous research on a lettuce harvesting robot (Birrell, Hughes, Cai, & Iida, 2019). The cost parameters that could not be found in the literature were extracted from an informal panel interview with five broccoli growers from the Netherlands and the United States.

3 | RESULTS

3.1 | The effect of network simplification and data augmentation on the cultivar-specific segmentation (Experiment 1)

Figure 9a,b summarize the effects of network simplification and data augmentation on the cultivar-specific mAP. Network simplification from Resnet101 to Resnet50 resulted in a decrease in mAP for all data augmentations (vertical comparison Figure 9a). For both Resnet backbones, the three types of data augmentation resulted in an increase in mAP compared to no data augmentation (horizontal comparison Figure 9a). For both Resnet50 and Resnet101, the highest increase in mAP was reached with the geometric data augmentation (G). The overall highest mAP of 0.77 was reached with Resnet101 and geometric data augmentation (R101/G). Figure 9b summarizes the p values of the pairwise Wilcoxon test in a mirrored matrix. The completely green-colored horizontal cells of R101/G indicate that the mAP of R101/G was significantly higher than the mAP of the other seven training methods.

3.2 | The effect of network simplification and data augmentation on the cultivar-generic segmentation (Experiment 2)

Figure 10a,b show the results of Experiment 2. The cultivar-generic mAP decreased when the network was simplified from Resnet101 to Resnet50 (vertical comparison Figure 10a). All data augmentations resulted in an increase in mAP compared to no data augmentation (horizontal comparison Figure 10a). The highest mAP of 0.71 was reached with R101/G and R101/GP. The percentages in Figure 10a show that the cultivar-generic mAPs of all training methods deviated more than 5% from the cultivar-specific mAP of Experiment 1. This indicates that none of the cultivar-generic training methods reached image generalization on the untrained cultivars. In Figure 10b, the six green-colored horizontal cells of R101/G and R101/GP indicate that these training methods had an mAP that was significantly higher than the six other training methods.

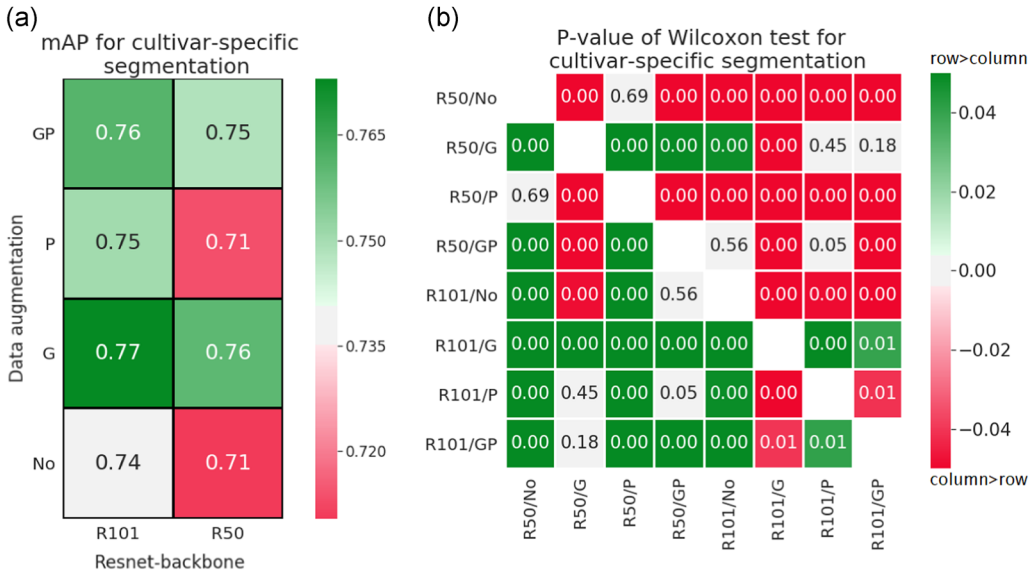

FIGURE 9 (a) The mAP for the cultivar-specific segmentation is summarized for the three broccoli cultivars. The green-colored cells indicate an increase in mAP compared to R101/No, which was the training method without any network simplification and data augmentation. The red-colored cells indicate a decrease in mAP compared to R101/No. (b) The mirrored matrix summarizes the *p* values of the pairwise Wilcoxon test. The green-colored cells indicate that the mAP of the training method in the row is significantly higher than the mAP of the training method in the column ($p \leq .05$). When the cell is red, then the mAP of the training method in the column is significantly higher than the one in the row ($p \leq .05$). mAP, mean average precision [Color figure can be viewed at wileyonlinelibrary.com]

3.3 | Number of cultivar-specific training images to generalize Mask R-CNN on a broccoli cultivar (Experiment 3)

In Experiment 3, the training was solely done with R101/G, because this training method had the highest mAP in both Experiment 1 and 2. Figure 11 summarizes the mAP as a function of the number of cultivar-specific images added to the training set of 600 images. With zero cultivar-specific training images, the mAP was 0.71 (this was the mAP of R101/G in Experiment 2). This mAP was 9% lower than the mAP of 600 cultivar-specific training images (the mAP of R101/G in Experiment 1). With 30 cultivar-specific training images (5%) the mAP was 0.77, which was 1% lower than the mAP of 600 cultivar-specific training images (Figure 11). Thus, training on 5% cultivar-specific

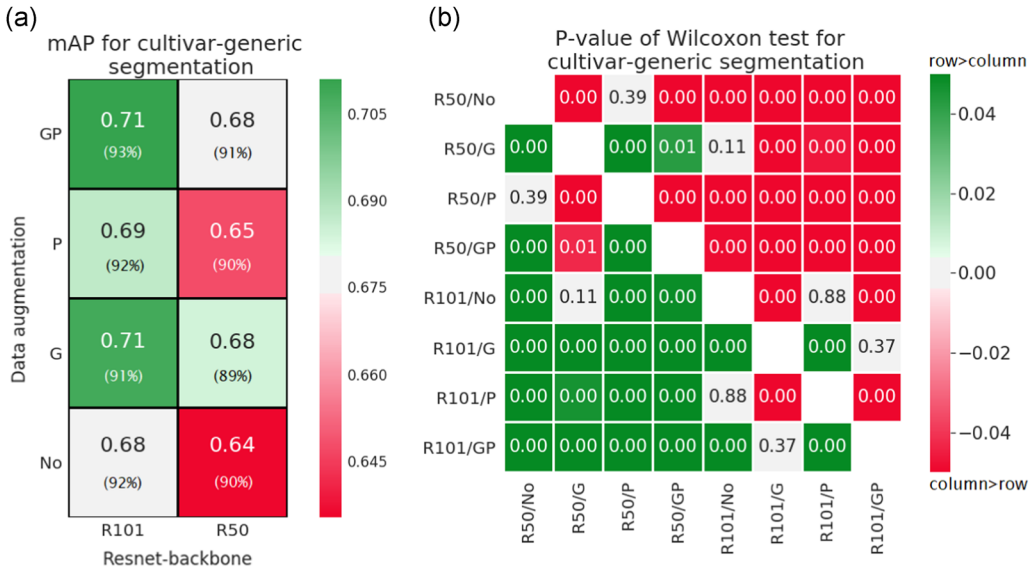

FIGURE 10 (a) The mAP for the cultivar-generic segmentation is summarized for the three broccoli cultivars. The percentages indicate the generalization performance compared to the cultivar-specific segmentation of Experiment 1. Like Figure 9a, the green-colored cells indicate an increase in mAP compared to R101/No and the red-colored cells indicate a decrease in mAP compared to R101/No. (b) Like Figure 9b, the mirrored matrix summarizes the *p* values of the pairwise Wilcoxon test ($p \leq .05$). mAP, mean average precision [Color figure can be viewed at wileyonlinelibrary.com]

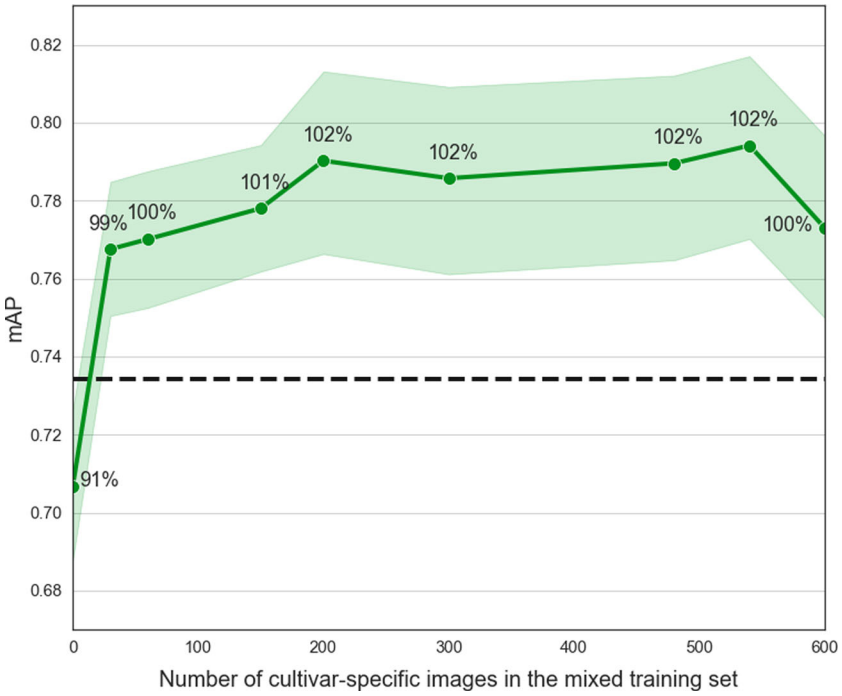

FIGURE 11 The green solid line depicts the mAP as a function of the cultivar-specific images in the mixed training set of 600 images. The green area represents the 95% confidence interval around the mean. The percentages indicate the generalization performance compared to the mAP of R101/G in Experiment 1 (which is the mAP of 600 cultivar-specific training images at the far right of the graph). The black-dashed line indicates the minimum mAP for image generalization. The mAP of zero cultivar-specific training images is equal to the mAP of R101/G in Experiment 2. mAP, mean average precision [Color figure can be viewed at wileyonlinelibrary.com]

images resulted in image generalization. With 200 (33.3%), 300 (50%), 480 (80%), and 540 (90%) cultivar-specific training images, the mAP was 2% higher than the mAP of 600 cultivar-specific training images (100%).

3.4 | Detection of harvestable broccoli heads (Experiment 4)

In Experiment 3, the highest mAP was reached when Mask R-CNN was trained on 200, 300, 480, and 540 cultivar-specific training images. In Experiment 4, we tested the Mask R-CNN that was trained on 200 cultivar-specific images, meaning that we used the Mask R-CNN that was optimized on the lowest number of cultivar-specific training images in this algorithm subset.

On our data set (the third test set), Mask R-CNN detected 229 of the 232 harvestable broccoli heads (see Figure 12a–c for some successful detections). The recall was 98.7% (Table 4). Three FNs were observed. One FN was observed on a broccoli head that was in the shadow of a big leaf (Figure 13a). Two FNs were found on broccoli heads that were (heavily) occluded by a leaf (Figure 13b,c). The leaf occlusion caused that the broccoli head was split into two distant parts, of which one part was detected as an individual (smaller) broccoli head, causing the IoU

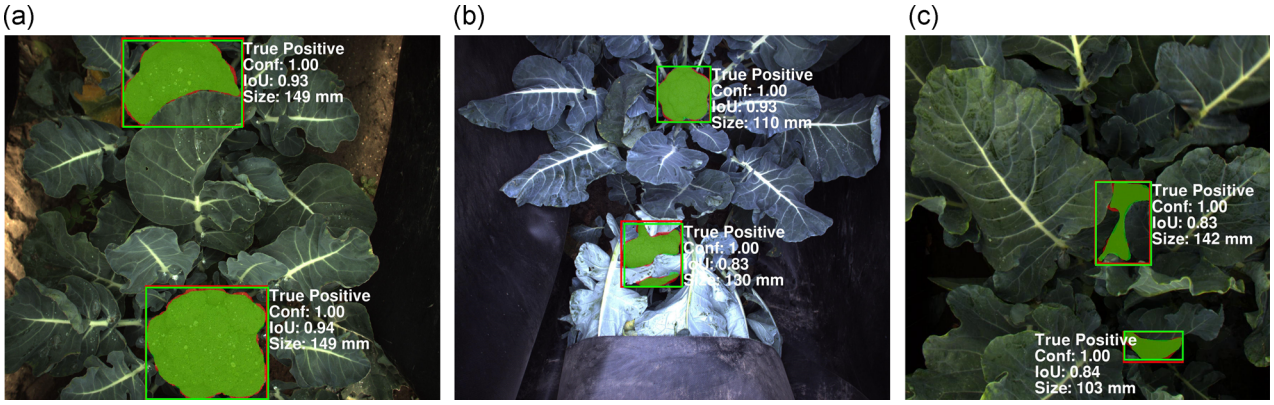

FIGURE 12 (a) True-positive detections on harvestable broccoli heads of the Ironman cultivar. (b) True-positive detections on harvestable broccoli heads of the Steel cultivar. (c) True-positive detections on harvestable broccoli heads of the Emerald-Crown cultivar. The red rectangle is the bounding box from the ground truth and the red pixels visualize the ground truth mask. The green rectangle is the bounding box prediction of Mask R-CNN and the green pixels visualize the predicted mask of Mask R-CNN. All detections in the images (a)–(c) were true-positives, because the values for the confidence level and the IoU exceeded the thresholds (Conf ≥ 0.99 and IoU ≥ 0.5). Size indicates the ground truth size. IoU, IoU, intersection over union; Mask R-CNN, Mask Region-based convolutional neural network [Color figure can be viewed at wileyonlinelibrary.com]

TABLE 4 Mask R-CNN detection performance on the broccoli heads of our own data set and the data set of Bender et al. (2019)

Data set	Harvest specification	Detection metrics				
		#TP	#FN	#FP	Recall (%)	Precision (%)
Own (third test set)	Harvestable (10–15 cm)	229	3	2	98.7	99.1
	Too small (<10 cm)	183	46	6	79.9	96.8
	Too big (>15 cm)	27	0	0	100.0	100.0
	All sizes	439	49	8	90.0	98.2
Bender et al. (2019)	Harvestable (10–15 cm)	175	1	1	99.4	99.4
	Too small (<10 cm)	108	4	2	96.4	98.2
	Too big (>15 cm)	116	0	0	100.0	100.0
	All sizes	399	5	3	98.8	99.3
Both data sets	Harvestable (10–15 cm)	404	4	3	99.0	99.3
	Too small (<10 cm)	291	50	8	85.3	97.3
	Too big (>15 cm)	143	0	0	100.0	100.0
	All sizes	838	54	11	93.9	98.7

Abbreviations: #FN, the number of false negatives; #FP, the number of false positives; #TP, the number of true positives.

to be lower than the threshold (Figure 13b,c). As a consequence, the two detections on the smaller broccoli parts were also false positives (because the IoU was lower than the threshold). There were two false positives in total (Figure 13b,c), resulting in a precision of 99.1% (Table 4).

On the images of the data set of Bender et al. (2019), Mask R-CNN detected 175 of the 176 harvestable broccoli heads (see Figure 14a–c, for some successful detections). There was one FN on a broccoli head that was only partially in the image (Figure 15a). There was one FP on a yellow leaf (Figure 15b). Both the recall and the precision were 99.4% (Table 4).

Mask R-CNN was also tested on the remaining broccoli heads that were either too small (<10 cm) or too big (>15 cm; Table 4). When detecting small-sized broccoli heads in both data sets, there were 50 FNs and 8 FPs (Table 4). There were no errors on the big-sized broccoli heads. When evaluating the object detection on both data sets on broccoli heads of all sizes, 838 of the 892 broccoli heads were successfully detected, resulting in a recall of 93.9%. On both

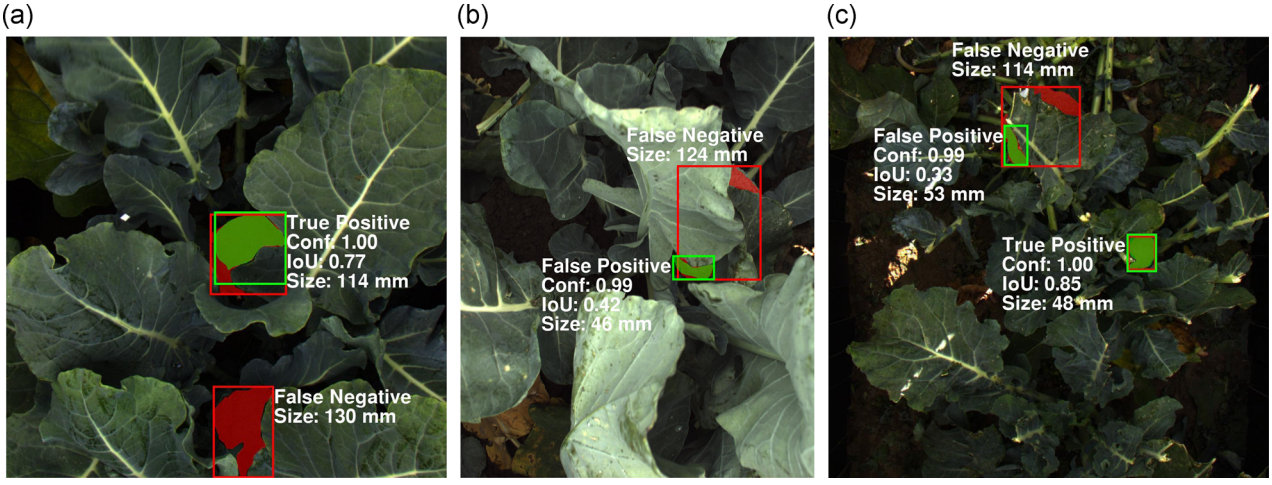

FIGURE 13 (a) On our own data set (the third test set), Mask R-CNN had one false-negative on a broccoli head (Emerald-Crown cultivar) that was in the shadow of a big leaf. (b) One false-positive (green bounding box) and one false-negative (red bounding box) were observed on a leaf-occluded broccoli head (Ironman cultivar). (c) One false-positive (green bounding box on the left) and one false-negative (red bounding box) were observed on a leaf-occluded broccoli head (Emerald-Crown cultivar). The leaf separated the broccoli head into two distant parts of which one part was detected as one (smaller) individual broccoli head instead of the complete broccoli head, causing the IoU to be lower than the threshold. Mask R-CNN, Mask Region-based convolutional neural network [Color figure can be viewed at wileyonlinelibrary.com]

FIGURE 14 (a) True-positive detections on two harvestable broccoli heads on a test image taken from Bender et al. (2019). (b) A true-positive detection on a harvestable broccoli head on another test image taken from Bender et al. (2019). (c) A true-positive detection on a harvestable broccoli head and the absence of a detection on an overripe broccoli head (this head should not be harvested) [Color figure can be viewed at wileyonlinelibrary.com]

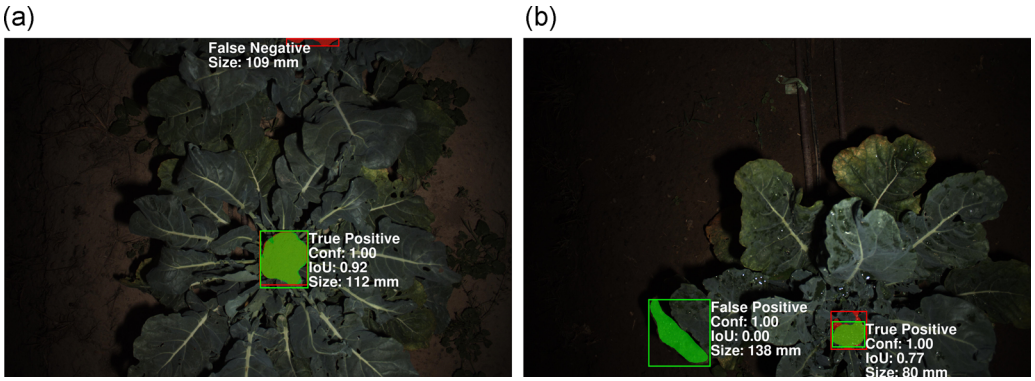

FIGURE 15 (a) On the test set taken from Bender et al. (2019), Mask R-CNN had one false-negative (red bounding box) on a broccoli head that was partially in the image. (b) Mask R-CNN had one false-positive (green bounding box on the left) on a yellow leaf. Mask R-CNN, Mask Region-based convolutional neural network [Color figure can be viewed at wileyonlinelibrary.com]

data sets, there were 11 FPs on the 849 broccoli detections, resulting in a precision of 98.7%. For all broccoli segmentations, the average IoU with the ground-truth mask was 0.87. The median was 0.90 (Figure 16).

3.5 | Cost-benefit analysis

On both data sets, the detection recall on the harvestable broccoli heads was 99.0% (Table 4). This recall was used to predict the harvest performance of a selective harvesting robot. With the prognosed harvest performance, we estimated the tentative benefits of the robot at €16,059 per hectare (using Equation A2 and Table A1 in the appendix).

The fixed costs of the robot were €385 per hectare (Equation A3 and Table A1 in the appendix). To calculate the variable costs of the robot, we first had to calculate the operating speed of the robot. We found that the robot's operating speed was not limited by the image analysis time (the maximum time was 0.27 s; Figure 17) but by the time that was needed to cut broccoli (2.0 s, Table A1 in the appendix). The cycle time of 2.0 s resulted in an operating speed of 0.17 m/s (Equation A6 in the appendix), which was, according to the

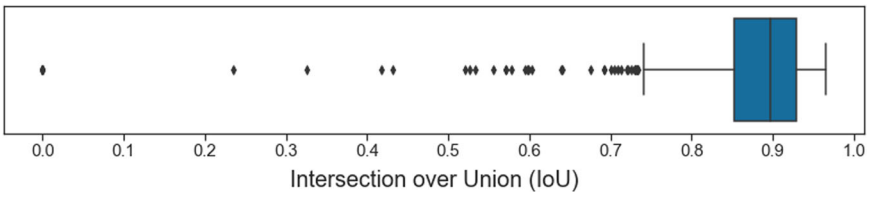

FIGURE 16 The boxplot visualizes the intersection over union between the ground-truth mask and the predicted mask for all 849 broccoli detections (of all sizes) on both data sets [Color figure can be viewed at wileyonlinelibrary.com]

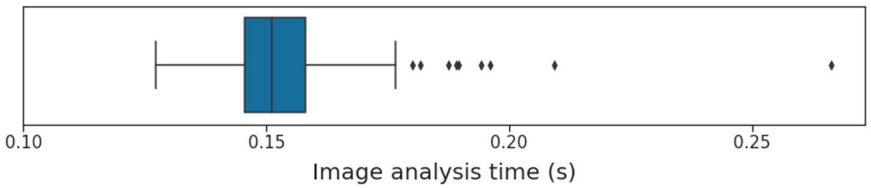

FIGURE 17 The boxplot visualizes the image analysis time of Mask Region-based convolutional neural network on all 600 images from both data sets. The maximum image analysis time was 0.27 s [Color figure can be viewed at wileyonlinelibrary.com]

informal panel interview, comparable to the speed of a human harvest crew. With the machine speed of 0.17 m/s, a robot operator would need 19.2 h to harvest one hectare of broccoli (Equation A5 in the appendix). The variable costs of the robot were €4310 per hectare (Equation A4 in the appendix). With the robot, the income per hectare was €11,365 (Equation A1 in the appendix).

For the hand-harvest, we calculated a benefit of €16,387 per hectare (Equation A2 and Table A1 in the appendix). This benefit was €328 higher than the benefit of the robot. The fixed costs of the hand-harvest were €77 per hectare (Equation A3 and Table A1 in the appendix), which was €308 lower than the fixed costs of the robot. The variable costs of the hand-harvest were €5654 per hectare (Equation A4 and Table A1 in the appendix). These variable costs were €1344 higher than the variable costs of the robot. The higher variable costs of the hand-harvest were caused by the additional 87.8 manhours that were needed to hand-harvest one hectare of broccoli. With the hand-harvest, the income per hectare was €10,657, which was €708 lower than the income of the robot.

4 | DISCUSSION

In Experiments 1 and 2, we observed a decrease in mAP when the network was simplified from Resnet101 to Resnet50. This result is consistent with Yu, Zhang, Yang, and Zhang (2019), who found that Mask R-CNN had better performance with Resnet101 compared to Resnet50 when detecting strawberry fruits. Resnet is a neural network that is designed to limit the loss of information during backpropagation (it solves the vanishing gradient problem). With this feature, a deeper residual network can learn more and potentially better image features without losing information, which can eventually increase the performance (He et al., 2016). Our results support the idea of training a deeper residual network to boost performance.

In Experiments 1 and 2, training with any type of data augmentation resulted in a higher mAP than training without data augmentation. Data augmentation with geometric transformations led to the largest increase in mAP. This finding is consistent with Taylor and Nitschke (2018), although they used different transformations (flipping instead of scaling and principal component analysis instead of light transformations). With the geometric transformations, the network was trained on images that had a transformed orientation, position, and scale compared to the original images. These geometrically transformed images are likely to resemble broccoli heads from our test set because the test images also contained broccoli heads of different sizes, scales, and positions. As a result, the geometrically transformed images allowed the neural network to learn robust features to detect the broccoli heads in the test images, resulting in a higher mAP. With the photometric transformations, we expect that the transformed images were less similar to the broccoli heads of the test set. For example, the light transformations could have transformed the broccoli pixels into unrealistically dark or bright pixels, especially when the input images were already dark or bright. Also, the texture transformations had the risk of changing the textural pattern of the broccoli head so that the network has learned textural patterns that did not resemble to the textural pattern of the broccoli heads in the test set, resulting in a lower mAP.

In Experiment 3, we found that Mask R-CNN reached image generalization on a broccoli cultivar when 5% of the training data set consisted of images of that cultivar. This means that a Mask R-CNN algorithm can be applied on a new crop cultivar, when it is retrained on only a few images of that cultivar.

In Experiment 4, we observed that Mask R-CNN had a higher detection performance on the Bender et al. (2019) data set, which our algorithm was not previously trained on. In total, Mask R-CNN detected 404 of the 408 harvestable broccoli heads. These broccoli heads were from three cultivars, five growing seasons and 11 broccoli fields that were located in three different countries. Our results imply that Mask R-CNN was successfully generalized on the images of multiple broccoli cultivars that differed in color and texture. This is an improvement compared to the algorithms of Blok et al. (2016) and Kusumam et al. (2017) that could not generalize sufficiently on images of two broccoli cultivars.

The higher number of FNs and FPs on the small-sized broccoli heads, indicates that Mask R-CNN can still be improved, especially for the purpose of measuring the size of the broccoli head. For the selective harvest, these FNs and FPs do not affect the performance, because the FNs correspond to small-sized broccoli heads that do not need to be harvested and the FPs will be filtered out by size. Moreover, the selective harvesting robot offers several opportunities to detect the previously FNs, because these small-sized broccoli heads will outgrow to a harvestable size when the robot returns in another field pass.

The maximum image analysis time of Mask R-CNN was only one-seventh of the cycle time of the robotic arm. This has two positive consequences. First, a maximum of seven image frames can be

processed for each broccoli head that must be cut. This multiple image analysis increases the chance of detecting the broccoli head. Second, there can be downgrade of the computing hardware without affecting the robot's operating speed. This hardware downgrade decreases the fixed costs of the robot.

A robot equipped with Mask R-CNN had higher benefits than costs. The robot was also more profitable than hand-harvest. The cost-benefit analysis was performed with two costs assumptions. The first assumption was a complete financial loss of a broccoli head when there was a FN. This financial loss is perhaps too pessimistic as the selective harvesting robot offers several opportunities to detect and harvest the previously FNs in another field pass. The second assumption was that humans could detect and cut broccoli with a success rate of 0.99 (both metrics were obtained from the informal panel interview). For an equal comparison with the robot, we also need to obtain the detection recall and the cut success of humans. In future research, we want to evaluate the robot in the field.

5 | CONCLUSIONS

Network simplification did not improve the image generalization of Mask R-CNN on multiple broccoli cultivars. Data augmentation did improve the image generalization of Mask R-CNN. In data augmentation, the geometric transformations led to a better image generalization than the photometric transformations. Furthermore, Mask R-CNN was generalized on a broccoli cultivar when only 5% of the training data set consisted of images of that cultivar. Our algorithm successfully detected 229 of the 232 harvestable broccoli heads from three cultivars that differed in texture and color. Additionally, our algorithm detected 175 of the 176 harvestable broccoli heads from an online data set, which our algorithm was not previously trained on. We conclude that our Mask R-CNN algorithm achieved better image generalization on multiple broccoli cultivars than existing broccoli detection algorithms from the literature. A robot equipped with Mask R-CNN had higher benefits than costs. Also, the robot was more profitable than human harvest. We conclude that Mask R-CNN provides sufficient basis for the commercialization of a selective broccoli harvesting robot.

ACKNOWLEDGMENTS

We thank Tony Wisdom from Skagit Valley Farm and Wiebe Goodijk from Firma Goodijk for their expert knowledge and financial support. We would like to thank all people who helped us in this study: Jan Zijlstra, Ian Mintz, Paul Goedhart, Wim van den Berg, Janne Kool, Koen van Boheemen, Ard Nieuwenhuizen, Paul Penders, Renee Spierings, Lodewijk Voorhoeve, Bernardo Mendonca, Joris IJsselmuiden, and Hyejeong Kim.

ORCID

Pieter M. Blok https://orcid.org/0000-0001-9535-5354
Frits K. van Evert https://orcid.org/0000-0003-0302-668X
Antonius P. M. Tielen https://orcid.org/0000-0003-3991-3942
Eldert J. van Henten https://orcid.org/0000-0002-1623-9855
Gert Kootstra https://orcid.org/0000-0002-2579-4324

REFERENCES

Abdulla, W. (2017). Mask R-CNN for object detection and instance segmentation on Keras and TensorFlow. Retrieved from https://github.com/matterport/Mask_RCNN

AgriConnect. (2019). Vormen van arbeid en loonkosten. Retrieved from https://agriconnect.nl/thema/vormen-van-arbeid-en-loonkosten

Bender, A., Whelan, B., & Sukkarieh, S. (2019). A high-resolution, multimodal data set for agricultural robotics: A Ladybird's-eye view of Brassica. *Journal of Field Robotics*, *121*, 800. https://doi.org/10.1002/rob.21877

Birrell, S., Hughes, J., Cai, J. Y., & Iida, F. (2019). A field-tested robotic harvesting system for iceberg lettuce. *Journal of Field Robotics*, *37*, 225–245. https://doi.org/10.1002/rob.21888

Blok, P. M., Barth, R., & van den Berg, W. (2016). Machine vision for a selective broccoli harvesting robot. *IFAC-PapersOnLine*, *49*(16), 66–71. https://doi.org/10.1016/j.ifacol.2016.10.013

COCO. (2019). Detection evaluation. Retrieved from http://cocodataset.org/#detection-eval

Edwards, W. (2019). Estimating farm machinery costs. Retrieved from https://www.extension.iastate.edu/agdm/crops/html/a3-29.html

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. Cambridge, MA: MIT press.

Haralick, R. M., Shanmugam, K., & Dinstein, Ih. (1973). Textural features for image classification. *IEEE Transactions on Systems, Man and Cybernetics*, *SMC-3*(6), 610–621

He, K., Gkioxari, G., Doll, P., & Girshick, R. B. (2017). Mask R-CNN. *CoRR*, *abs/1703.06870*.

He, K., Gkioxari, G., Dollár, P., & Girshick, R. (2017). Mask R-CNN. *IEEE International Conference on Computer Vision (ICCV)*, Venice, 2980–2988. https://doi.org/10.1109/ICCV.2017.322

Hernández-García, A., & König, P. (2018). Data augmentation instead of explicit regularization. *arXiv preprint arXiv*, *1806*, 03852.

Jiang, Y., Shuang, L., Li, C., Paterson, A. H., & Robertson, J. (2018). *Deep learning for thermal image segmentation to measure canopy temperature of Brassica oleracea in the field*. Paper presented at the 2018 ASABE Annual International Meeting, St. Joseph, MI. http://elibrary.asabe.org/abstract.asp?aid=49404&t=5

Jung, A. (2019). Imgaug. Retrieved from https://github.com/aleju/imgaug

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.2018.02016

Keras. (2019). Keras callbacks. Retrieved from https://keras.io/callbacks/

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

KWIN. (2018). *Kwantitatieve Informatie Akkerbouw en Vollegrondsgroenteteelt 2018*. Wageningen: Wageningen University and Research.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*, 436–444. https://doi.org/10.1038/nature14539

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., … Zitnick, C. L. (2014). *Microsoft coco: Common objects in context*. Paper presented at the European conference on computer vision.

Lin, T. (2019). labelImg. Retrieved from https://github.com/tzutalin/labelImg

Perez, L., & Wang, J. (2017). The effectiveness of data augmentation in image classification using deep learning. *arXiv preprint arXiv*, *1712*, 04621.

Ramirez, R. A. (2006). *Computer vision based analysis of broccoli for application in a selective autonomous harvester* (Master of Science Master's). Blacksburg, Virginia: Virginia Polytechnic Institute and State University.

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 & Machine Intelligence*, *6*, 1137–1149.

Romera-Paredes, B., & Torr, P. H. S (2016). Recurrent instance segmentation. *arXiv preprint arXiv*.

Rosebrock, A. (2018). *Deep learning for computer vision with python: Starter bundle*. PyImageSearch.

Seminis. (2019). Best management practices for broccoli. *Agronomic Spotlight*. Retrieved from http://www.seminis-us.com/resources/agronomic-spotlights/best-management-practices-for-broccoli/

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

Shijie, J., Ping, W., Peiyi, J., & Siping, H. (2017). *Research on data augmentation for image classification based on convolution neural networks*. Paper presented at the 2017 Chinese Automation Congress (CAC).

Springenberg, J. T., Dosovitskiy, A., Brox, T., & Riedmiller, M. (2014). Striving for simplicity: The all convolutional net. *arXiv preprint arXiv*, *1412*, 6806.

Taylor, L., & Nitschke, G. (2018). *Improving Deep Learning with Generic Data Augmentation*. Paper presented at the 2018 IEEE Symposium Series on Computational Intelligence (SSCI).

Wilcoxon, F. (1992). Individual comparisons by ranking methods. In S, Kotz, NL & Johnson (Eds.), *Breakthroughs in statistics: Methodology and distribution* (pp. 196–202). New York, NY: Springer.

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

Zagoruyko, S., & Komodakis, N. (2016). Wide residual networks. *arXiv preprint arXiv*, *1605*. 07146-2221

Zhu, Y., Aoun, M., Krijn, M., Vanschoren, J., & Campus, H. T. (2018). *Data augmentation using conditional generative adversarial networks for leaf counting in arabidopsis plants*. In: CVPPP.

How to cite this article: Blok PM, van Evert FK, Tielen APM, van Henten EJ, Kootstra G. The effect of data augmentation and network simplification on the image-based detection of broccoli heads with Mask R-CNN. *J Field Robotics*. 2020;1–20. https://doi.org/10.1002/rob.21975

APPENDIX A: COST-BENEFIT ANALYSIS

The following section summarizes the equations and the cost parameters that were used in the cost-benefit analysis. Refer to Table A1 for the values and the description of the cost parameters that were used in Equation (A1) till Equation (A6).

With Equation (A1), we calculated the income per hectare of broccoli when using a robot equipped with Mask R-CNN (i_r). The income per hectare was the difference between the benefits (b_r) and the costs, which was the sum of the fixed costs (c_{fr}) and the variable costs (c_{vr}).

$$i_r = b_r - c_{fr} - c_{vr}. \tag{A1}$$

The benefits of the robot (b_r) depended on the revenue of the broccoli heads that were successfully harvested per hectare (Equation A2). The revenue was derived from the number of broccoli heads per hectare (bh_{ha}), the sale value per broccoli head (s_v) and the harvest success of the robot. The harvest success depended on the recall of the detection system (r; using Table 4), and the broccoli cut success of the end-effector (s).

$$b_r = bh_{ha} \cdot s_v \cdot (r \cdot s). \tag{A2}$$

The fixed costs of the robot per hectare (c_{fr}) were calculated from the robot's price (p_r), its salvage value (v_r), its economic life (t_r) and the hectares that are harvested per year (ha_y) (Equation A3).

$$c_{fr} = \frac{p_r - v_r}{t_r \cdot ha_y}. \tag{A3}$$

The variable costs of the robot (c_{vr}) were the sum of the costs for crop production (c_c) and the costs for labor (Equation A4). The costs for labor were derived from the hourly wage (c_l) and the total labor requirement per hectare. The total labor requirement was the sum of the labor for crop care (l_c), harvest (l_h) and postharvest (l_{ph}). The labor for harvest (l_h; Equation A5) depended on the number of people needed to operate the robot (p), the number of broccoli cuts per year (cu_y) and the harvest capacity of the robot. The harvest capacity depended on the robot's operating width (w_r) and its operating speed (v). The operating speed was influenced by the intra-row spacing between the broccoli heads (d_{bh}) and the maximum time that was needed to either analyze an image (t_i) or cut a broccoli (t_a; Equation A6).

We also calculated the labor that was needed for the headland maneuver (assuming a reversed turn). This labor depended on the field width (w_f), the robot's turning radius (r_t) and the distance between the camera and the robotic arm (d_{ca}). This distance is the distance that the robot had to travel to complete the harvest operation before it could start the turning procedure. In Equation (A5) we accounted for the conversion between m^2 and hectare (10,000 m^2 = 1 ha) and the conversion between seconds and hour (3600 s = 1 h).

$$c_{vr} = c_c + c_l \cdot (l_c + l_h + l_{ph}), \tag{A4}$$

$$l_h = \frac{p \cdot cu_y \cdot \left(\frac{10,000}{w_r \cdot v} + \left(\frac{w_f}{w_r} - 1\right) \cdot \frac{2 \cdot d_{ca} + (\pi + 1) \cdot r_t}{v}\right)}{3600}, \tag{A5}$$

$$v = \frac{d_{bh}}{\max\ (t_i, t_a)}, \tag{A6}$$

To determine the profitability of the robot compared to the hand-harvest, we also calculated the costs and the benefits for the hand-harvest, using Equation (A1) to Equation (A4). All cost parameters can be found in Table A1.

TABLE A1 The cost parameters that were used in the cost-benefit analysis for the robotic harvest ("Robot") and the hand-harvest ("Hand"). Source refers to the reference that was used to extract the cost parameter. The cost parameters from p to t_a (used to calculate the labor for harvest) were only obtained for the robotic harvest and not for the hand-harvest, because the labor for the hand-harvest was already estimated at 107 hours by KWIN (2018)

Para-meter	Description	Unit	Robot	Source	Hand	Source
bh_{ha}	Broccoli heads per hectare	–	30,400	KWIN (2018)	30,400	KWIN (2018)
s_v	Sale value of one broccoli	€	0.55	KWIN (2018)	0.55	KWIN (2018)
r	Detection recall on the harvestable broccoli heads	–	0.99	Section 3.4	0.99	Panel interview
s	Broccoli cut success	–	0.97	Birrell et al. (2019)[a]	0.99	Panel interview
p_r	Machine purchase price	€	500,000	Panel interview	150,000	Panel interview
v_r	Salvage value of the machine	€	115,000	23% of the purchase price according to Edwards (2019)	34,500	23% of the purchase price according to Edwards (2019)
t_r	Economic life of the machine	y	10	Panel interview	15	Panel interview
ha_y	Hectares harvested per year	ha	100	Panel interview	100	Panel interview
c_c	Crop costs per hectare	€	2607	KWIN (2018)	2607	KWIN (2018)
c_l	Hourly labor wage in The Netherlands	€	15.31	AgriConnect (2019); KWIN (2018)	15.31	AgriConnect (2019); KWIN (2018)
l_c	Labor for crop care	h	49	KWIN (2018)	49	KWIN (2018)
l_h	Labor for harvest	h	19.2	Section 3.5	107	KWIN (2018)
l_{ph}	Labor for postharvest	h	43	KWIN (2018)	43	KWIN (2018)
p	People needed to operate the robot	–	1	Panel interview		
cu_y	Broccoli cuts per year	–	3	KWIN (2018)		
w_r	Operating width of the robot	m	3	Panel interview		
v	Operating speed	m/s	0.17	Section 3.5		
w_f	Field width	m	60	Panel interview		

(Continues)

TABLE A1 (Continued)

Para-meter	Description	Unit	Robot	Source	Hand	Source
d_{ca}	Distance between the camera and the robotic arm	m	2	Panel interview		
r_t	Turning radius robot	m	5	Panel interview		
d_{bh}	Intra-row distance	m	0.33	Section 2.1.2		
t_i	Maximum image analysis time	s	0.27	Section 3.5		
t_a	Cycle time of the robotic arm	s	2.0	The cycle time of the robotic arm is the sum of the time of the forward and backward movement of the arm (1.2 s) and the time for defoliating and cutting the broccoli (0.8 s) (based on the specifications of the robotic arm)		

[a]*Note:* There was no data on the broccoli cut success of the robot, so we extracted the 97% detachment success of Birrell et al. (2019) who researched a comparable use-case (robotic harvest of iceberg lettuce).