Smart Farming: Using End-To-End Deep Learning to Estimate the Size of Broccoli Heads

A.C. Rijken[1]

Supervisor(s): dr.ir. C.C.S. Liem[1]**, dr.ir. J. Wildenbeest**[2]

[1]EEMCS, Delft University of Technology, The Netherlands

[2]Robotics, Inholland University of Applied Sciences, The Netherlands

A thesis submitted to EEMCS Faculty Delft University of Technology,
in partial fulfilment of the requirements
for the Bachelor of Computer Science and Engineering
June 20, 2026

Name of the student: A.C. Rijken
Final project course: CSE3000 Research Project
Thesis committee: dr.ir. C.C.S. Liem, dr.ir. J. Wildenbeest, dr. X. Zhang

Abstract

Broccoli farming requires accurate estimations of the head diameter to determine when crops are ready to be harvested. An existing growth prediction model was trained on data gathered at Verdonk Broccoli. This paper explores an alternative way of gathering the data needed to train the prediction model, namely by using an end-to-end deep learning model to estimate the diameter of a broccoli head based on an image and the camera height.

This new method was first evaluated by comparing the original diameter measurements with the model outputs. The impact of the method was then assessed by examining the accuracy of the prediction model when trained on the data generated using this approach.

The results show that this alternative method can achieve at least the same accuracy as the original approach, but further improvements in accuracy are limited by the dataset used to train the model.

1 Introduction

1.1 Problem Context

Agriculture is a major industry in the Netherlands, and broccoli farming is a substantial part of the sector. In recent decades, the number of broccoli crops cultivated each year has increased significantly. In 2025, a total of 33.9 million kilogrammes of broccoli were produced on 3,028 hectares [4]. Although production has increased, broccoli is still often harvested manually. Workers go into the fields, measure the heads of broccoli to determine whether they are ready for harvest, and cut them if they are. Given that manual workers are difficult to find and this is expected to become even more challenging in the future, there is an interest in automating broccoli production.

Furthermore, significant crop losses occur during harvesting. This is because farmers can perform only a limited number of harvest passes due to logistical constraints. If a broccoli head is too small (less than 500 grammes), farmers cannot sell it on the fresh market. Since broccoli does not grow uniformly, the effectiveness of one-over harvesting is limited [8]. This leads to both food waste and reduced profit for farmers. Predicting broccoli growth helps this problem, as farmers can create optimal harvesting schedules based on predicted growth. Conseqquently, harvest passes will contain more broccoli of the ideal size, reducing food waste and increasing profits.

This paper builds on research conducted in 2025 by N. Mateijsen et al. [7]. They collected data on broccoli head growth and trained a model to estimate future broccoli growth based on this data. This research was done in close collaboration with Verdonk Broccoli, where the training data was collected. This study also primarily used the dataset collected by them during the previous summer.

The model proposed by Mateijsen et al. [7] performed quite well and predicted the diameters of the broccoli heads with a mean absolute error (MAE) of 0.583 cm. However, according to the authors, a significant portion of this inaccuracy can be attributed to inaccuracies in the diameter estimations of the training data. In an effort to determine whether alternative methods for estimating broccoli head size are viable, an end-to-end deep learning approach will be explored. The model will be designed and trained with an emphasis on end-to-end learning and minimal feature engineering. Afterwards, we will evaluate whether this is a feasible method for diameter estimation. The main aim of this paper is to match the accuracy of the original method by looking at both the estimations and their influence on the predictions.

1.2 Research Questions

To structure this research project, we will answer the question: *Can an end-to-end deep learning model provide a reliable alternative to the broccoli head estimation technique as proposed by N. Mateijsen et al. [7]?*. To answer this question effectively, the main research question is broken down into three sub-questions:

1. *How can a deep learning model be designed to estimate the diameter of a broccoli head with precision similar to the method proposed by Mateijsen et al. [7]?*
2. *What effect does replacing the original diameter estimates with deep learning based estimates have on downstream broccoli growth prediction accuracy?*
3. *To what extent does the model's accuracy transfer to the dataset of P. Blok et al. [2]?*

The first sub-question explores how the model is designed and trained. The aim is to develop a deep learning model that can take an image of a broccoli and the camera height, and it outputs the diameter of the head. This is achieved by using end-to-end methods. The model will be evaluated using MAE and looking for trends between the error rate and other factors. The second sub-question further evaluates the model by looking at the impact it has on the prediction accuracy. This is done by assessing how the accuracy changes when the model is used instead of the original method to estimate the head sizes of broccoli when collecting the dataset used to train and validate the prediction model as proposed by Mateijsen et al. [7]. The final sub-question will look at the transferability of the model architecture and training methods to a different dataset, namely that of Blok et al. [2]. Since this dataset provides occlusion ratios for the broccoli images, we will analyse whether the occlusion ratio correlates with the error rate.

In addition to this paper, the code used for data processing and conducting experiments is published at `https://github.com/Rijkaron/broccoli-public`.

1.3 Structure

This paper is structured as follows. In Section 2, two other relevant papers are highlighted that contain details necessary to understand the rest of this paper. Section 3 explains the methods used to answer the research questions. The results of these experiments are described in Section 4 and interpreted in Section 5. Sections 3 to 5 each contain subsections dedicated to one of the research sub-questions of the research. Section 6 contains remarks on the difficulty of replicating this research. Finally, Section 7 summarises the results, discusses their implications, and offers suggestions for future work.

2 Background

A relevant external paper for this research is that of P. Blok et al. [3]. This research focused on the estimation of the diameter of the heads of the occluded broccoli. Their approach for estimating the diameter consisted of three steps: first, fitting a circle to the broccoli head; second, determining the pixel-to-millimetre ratio for the distance to the broccoli; and third, using this ratio to convert the diameter of the fitted circle into a diameter in millimetres. Another contribution of this work is that it provides a dataset consisting of 1,613 images of broccoli heads, annotated with diameters and occlusion ratios [2].

Another relevant paper is, logically, the paper on which this research is based, namely that of Mateijsen et al. [7]. There are two main relevant parts of their paper, namely the data collection and the model training. We will outline these here briefly.

First, the data collection process is described. The data was collected over 7 measuring days between June 11th and June 25th 2025. On these measurement days, two Intel RealSense D415 RGB-D cameras were used while walking through the broccoli field. Each camera captured a separate track of crops. The field was divided into 10 segments, each containing 30 broccoli heads. In total, they tracked $2 \times 10 \times 30 = 600$ broccoli crops. For every day, segment, and camera, the videos are analysed using a YOLOv8n-model to detect broccoli heads. Then, all the detected broccolis are given an ID, so their growth can be tracked on the different measuring days. After this, the diameter of every broccoli could be determined. This was initially done using Equation 2.1, where D is the diameter in metres, D_{px} is the number of pixels on the longest side of the bounding box around the head of the broccoli, H_1 is the distance from the camera to the centre of the crown in metres, and f_x is the focal length of the camera.

$$D = \frac{D_{px} \cdot H_1}{f_x} \tag{2.1}$$

This formula was then fine-tuned using a correction factor, C_b, which was empirically determined to be 0.1, as shown in Equation 2.2.

$$D = \frac{D_{px} \cdot H_1 \cdot (f_x + D_{px} \cdot C_b)}{f_x^2} \tag{2.2}$$

In practice, H_1 was estimated by shrinking the bounding box by 50% around its centre and computing the median depth. To estimate the actual diameter of a broccoli head on a given measurement day, the diameter was calculated for every frame in which the broccoli was detected. To reduce the impact of noise and occlusions, the final diameter estimate was taken as the median of the upper 50th percentile of the measured diameters. Afterwards, the dataset was filtered by removing broccoli heads that had significant measurement errors during the process, such as a locally decreasing growth curve.

To train the model, N. Mateijsen et al. compared multiple model architectures. The models used 15 features derived from environmental conditions, with thermal time being the most influential, followed by the vapour pressure deficit (VPD) and the humidity and temperature of the soil. The best-performing model was a multilayer perceptron model (MLP), which achieved an MAE of 0.583 cm. Given that the standard deviation of the upper 50th percentile was 0.14 cm, the inaccuracies in the measurements contribute significantly to the precision of the predictions.

3 Methods

3.1 SQ1: Training the Deep Learning model

3.1.1 Data Gathering

In order to effectively train a model, it is important that the data is of sufficient quality. Before starting the research, we assumed that the data collected at Verdonk Broccoli by Mateijsen et al. [7] was immediately suitable for training the model. However, this quickly proved not to be the case.

Firstly, the data was divided in such a way that it was complex to pair the image of a broccoli to its assigned ID and its measured diameter. As a result, a considerable amount of time was spent replicating the results and figuring out how we could use the provided data and code.

Secondly, which is more interesting, was that when we composed the dataset, there were several large outliers. For example, we found broccoli detections with an estimated diameter of more than 10 metres. This was due to noise in the depth channel, which caused an impossible high estimation of the camera height. In Figure 3.1, one such instance is shown. In this image, the white box is the bounding box around the broccoli head generated by the YOLO model, and the pixels inside the green box (the region of interest) are used to calculate the distance from the camera. The colours that are blended over the image give an indication of the measured depth at a specific pixel, where blue is closer and red is further away.

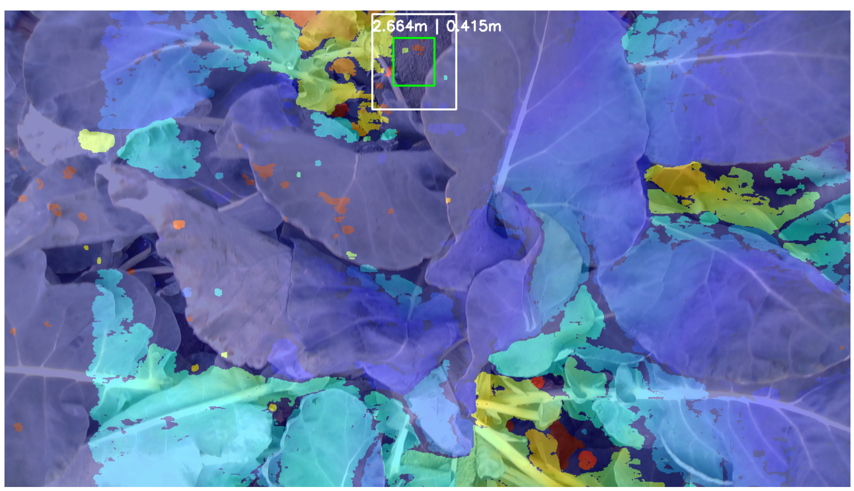

Figure 3.1: Image with the depth channel overlay

To prevent these faulty measurements from ending up in the training set, not only was the median value in the region of interest calculated, but we also kept track of the spread of the depth measurements (defined as the difference between the 95th percentile and the 5th percentile), the standard deviation, and the number of pixels. We required the spread to be less than 0.15 metres, the standard deviation to be less than 0.6 metres, and the number of pixels to be greater than 50. These thresholds were chosen empirically, trying not to include the obvious faulty data while keeping as much of the correct data.

Additionally, since the aim is to design a model that works well for one broccoli head per image, we also excluded all images containing multiple broccoli heads. Lastly, broccoli heads that were partially in the frame were not useful because they do not give accurate measurements. Therefore, we discarded all images where the bounding box of the broccoli head was within 50 pixels from the border of the image.

3.1.2 Model Architecture

We used PyTorch [1] to design and train the model. The architecture combines the images with a depth scalar, feeds them into a regression head, and outputs a single scalar representing the diameter. The exact architecture and hyperparameter configuration were determined by trial-and-error, and we stopped when a satisfactory result was obtained.

The backbone of the image branch was ResNet18 [6], a lightweight CNN model. ImageNet-1K weights were used, as described by Deng et al. [5]. The classification layer was removed, causing the model to output a 512-dimensional feature vector per image. To incorporate the height of the camera, we used a small multilayer perceptron. The depth value was projected onto a 16-dimensional feature vector by the first layer. The 16 features were then passed through another linear layer with 16 dimensions. The resulting depth representation was concatenated with the output of the image branch, producing a 528-dimensional vector.

This vector was passed through another linear layer, resulting in a 128-dimensional vector. Then a dropout of 0.2 was applied. Finally, the vector passed through the last layer to produce the single value representing the diameter.

3.1.3 Training and Evaluating

Before training, the input images were resized to a resolution of 224×224, and normalised using the parameters required by ImageNet-1K[1]. In addition, the following data augmentation techniques were used: random horizontal flip, random rotation of up to 10 degrees, random brightness, contrast, and saturation jitter of 0.2, and hue jitter of 0.1.

Since the dataset often contains multiple images of the same broccoli on the same date, the images were randomly sampled to include N images, where $N = \lfloor \sqrt{m} \rfloor$, and m is the number of images of the same broccoli on the same date. For example, if there were 9 images of the same broccoli on the same date, only 3 were used.

The training was done using 10-fold cross-validation. That is, the dataset was divided into 10 folds and, for 10 iterations, a different fold was used to validate, while the remaining nine folds were used for training. The folds were constructed to contain similar amounts of data and to each represent the full time span over which the data was collected. The model was trained for 20 epochs per fold, with a batch size of 32 and the Adam-optimiser with a learning rate of 0.0001.

The performance of the model was assessed using the mean absolute error (MAE) metric. This metric is defined by Equation 3.1:

$$MAE = \frac{1}{n} \sum_{i=1}^{n} |y_i - x_i| \tag{3.1}$$

[1]These parameters can be found at `https://docs.pytorch.org/vision/main/models/generated/torchvision.models.resnet18.html`

Here, n is the sample size, y_i is the predicted value, and x_i is the true value.

To decide whether the model is accurate, the final MAE will be used, in addition to assessing whether there is a trend between the error of an image and several other factors, namely the height of the camera, the different cameras, different measurement days, and different segments.

Interestingly, while conducting this experiment, we noticed that although Mateijsen et al. [7] claimed to use the corrected formula to estimate the diameter based on the bounding box (see Equation 2.2), the shared results were in line as if they were using the original pinhole formula (see Equation 2.1). Because of this, we assumed that the original research used the original pinhole formula. However, we will also briefly compare the difference in results when using either of the two as ground truth.

3.2 SQ2: The Influence of the Alternative Estimation Method on the Prediction Accuracy

To further validate the accuracy of the model and assess its impact on the accuracy of broccoli growth predictions, we replaced the diameter estimation method used in the pipeline of Mateijsen et al. [7] with our model.

The first step is to obtain the predictions from the model. This is done using a similar method as in the previous experiment, namely by 10-fold cross-validation. We decided on using the original pinhole formula as our ground truth, since that was probably also used by Mateijsen et al. [7].

There is, however, a difference in the data used for the training and validation, as compared to what we used in the first sub-question. The pipeline benefits from having multiple measurements of the same broccoli on the same date, since it can take the median of all the measurements to reduce noise. Therefore, we did not downsample the dataset before dividing it into folds. Instead, we used the entire dataset and made sure that images of the same broccoli on the same date were present in the same fold, while still stratifying on the measurement date. This allowed us to downsample the training set for every fold, while keeping the validation set the same size, and still preventing the model from memorising specific broccoli heads. The validation results of these folds were then merged into one large dataset, which was used to train the prediction model.

When examining the code Mateijsen et al. [7] used to train the prediction model, we noticed that they only used a subset of their full dataset, seemingly by accident. They split the full dataset into a training set and a validation set for a proof-of-concept model, but forgot about this and conducted k-fold cross-validation only on the training set. In order to fairly compare their results to ours, we retrained the model on the full dataset, so both the training and the validation set.

In addition to these two results, we were also interested in assessing the impact of filtering on the data, specifically for two filtering steps. Firstly, the criteria we used to select images to use in our training set as described in Section 3.1.1, which we refer to as pre-filtering from now on. Secondly, the removal of the measurements of the same broccoli on the same date that fall below the median as proposed in the paper by Mateijsen et al. [7], which we refer to as post-filtering. Therefore, we decided to train the model on 5 different configurations, namely: the original dataset without pre-filtering but with post-filtering, corresponding to the original method; the original dataset with pre-filtering and post-filtering; the original dataset with pre-filtering but without post-filtering; the deep learning dataset, which is always pre-filtered, with post-filtering; and the deep learning dataset without post-filtering.

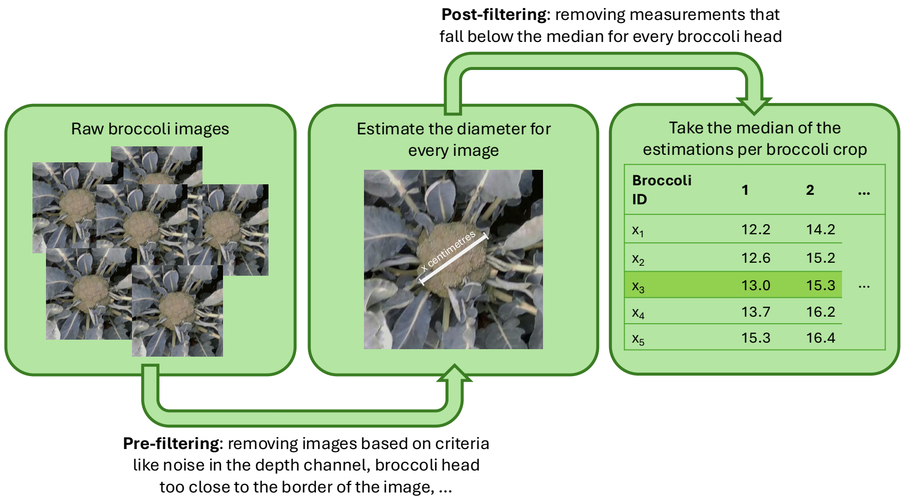

Figure 3.2: Visualization showing pre- and post-filtering.

Because training a model involves some randomness, we ran the training process 100 times for each configuration and sorted the results by the mean MAE over the 10 folds. This resulted in 100 MAE values, each with the corresponding standard deviation across the folds. We then took the best, median, and worst MAE values to compare the results.

3.3 SQ3: Evaluating the Model on an External Dataset

The purpose of this sub-question is twofold. Firstly, we evaluated how well the model generalises to an external dataset, namely the dataset of Blok et al. [2]. Secondly, since this dataset contained not only images of broccoli heads with their estimated diameter but also their corresponding occlusion ratio, we could stratify by occlusion ratio and look for a trend between error rate and occlusion. A limitation of the dataset is that it only contained two measurement days.

We validated the model on this data similarly to sub-question one. We therefore apply the same pre-filtering criteria to the images and performed 10-fold cross-validation. However, we stratified by the occlusion ratio instead of the measurement date. The evaluation metrics are again similar: we used MAE and searched for correlations between the error rate and different factors, namely the distance to the camera and the occlusion ratio.

The dataset from Blok et al. [2] was significantly smaller than the one we used in sub-question one. To fairly compare the MAE values, we randomly downsampled the latter dataset to contain the same amount of training data. It was then trained and evaluated the same way as in sub-question one.

4 Results

4.1 SQ1: Results of Training the Model on Verdonk Data

As stated, we trained the model across 10 folds for 20 epochs each. This was done using PyTorch CUDA on an NVIDIA GeForce RTX 4080 graphics card. The dataset contained, after downsampling, 10,619 images.

We ran the experiment twice, using the standard pinhole formula as ground truth the first time and the altered formula with the correction factor the second time. For the pinhole formula, the mean MAE across the folds was 0.374 centimetres, with a standard deviation of 0.013 across the different MAE values of the fold. For the corrected formula, the mean MAE was 0.381 centimetres, with a standard deviation of 0.008.

For the results of the original pinhole formula, we also looked for correlations between the error rate and the camera height, and looked whether the errors were evenly distributed across the cameras, measurement days, or segments.

The camera heights were evenly divided into 9 bins between heights of 0.39 and 0.84 cm. The MAE and the mean relative error (MRE) were then calculated for the measurements that fell into each bin. The results showed higher values of MAE and MRE for lower camera heights, gradually decreasing for measurements with the camera at a larger distance, before slightly increasing again when the camera was further from the broccoli heads. This trend follows the amount of training data that is available for a specific bin. The more training data for a specific distance, the more accurate the predictions. For the full results, see Table 4.1 and Figure 4.1.

Table 4.1: Mean absolute error and relative error by camera height. Values are reported as mean ± standard error.

Camera height (cm)	MAE (cm)	MRE (%)	Bin size
$[0.39, 0.44)$	1.151 ± 0.252	27.94 ± 3.81	25
$[0.44, 0.49)$	0.775 ± 0.104	18.14 ± 1.96	42
$[0.49, 0.54)$	0.694 ± 0.086	15.67 ± 1.84	56
$[0.54, 0.59)$	0.583 ± 0.047	7.20 ± 0.60	146
$[0.59, 0.64)$	0.522 ± 0.017	5.34 ± 0.21	1085
$[0.64, 0.69)$	0.431 ± 0.008	5.14 ± 0.14	3087
$[0.69, 0.74)$	0.318 ± 0.005	5.73 ± 0.11	4271
$[0.74, 0.79)$	0.274 ± 0.006	6.41 ± 0.17	1838
$[0.79, 0.84)$	0.330 ± 0.039	7.97 ± 0.68	68

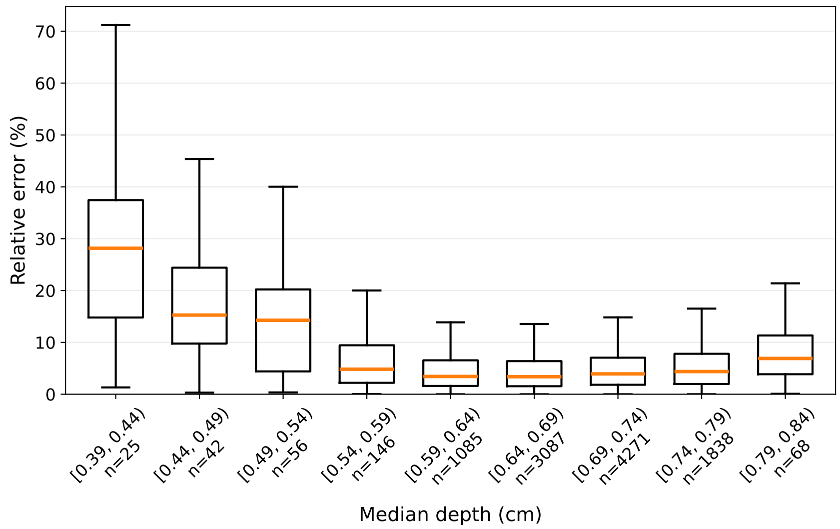

Figure 4.1: Mean relative error by camera height

We also examined whether the errors were evenly distributed between the different cameras, measurement days, and segments. Camera one had an MAE of 0.375 cm and an MRE of 5.6% on 5432 images, while camera two had an MAE of 0.372 cm and an MRE of 6.0% across 5187 images. The MAE values across the different measurement days are generally larger the later the measurement date is. The MRE follows the inverse pattern: it generally decreases when considering the dates in chronological order. The MAE values for the different segments lie between 0.339 cm (segment 3) and 0.432 cm (segment 9). The complete results for the different measurement days can be found in Table 4.2, and the values for the different segments can be found in Table 4.3.

Table 4.2: MAE and MRE by measurement day. Values are reported as mean ± standard error.

Day	MAE (cm)	MRE (%)	Sample size
07-11	0.275 ± 0.009	12.42 ± 0.43	756
07-14	0.243 ± 0.005	6.85 ± 0.21	2006
07-16	0.299 ± 0.007	5.66 ± 0.15	1651
07-18	0.340 ± 0.006	4.58 ± 0.09	2808
07-21	0.474 ± 0.014	4.99 ± 0.22	1339
07-23	0.545 ± 0.014	4.60 ± 0.14	1532
07-25	0.668 ± 0.029	5.30 ± 0.26	527

Table 4.3: MAE and MRE by segment. Values are reported as mean ± standard error.

Segment	MAE (cm)	MRE (%)	Sample size
1	0.372 ± 0.013	5.74 ± 0.29	1148
2	0.372 ± 0.012	5.77 ± 0.23	1149
3	0.339 ± 0.010	5.30 ± 0.18	1142
4	0.345 ± 0.010	5.41 ± 0.20	1074
5	0.349 ± 0.010	5.11 ± 0.18	1143
6	0.372 ± 0.014	5.75 ± 0.23	854
7	0.392 ± 0.012	6.16 ± 0.23	1122
8	0.387 ± 0.014	6.47 ± 0.28	1015
9	0.432 ± 0.017	6.48 ± 0.24	998
10	0.385 ± 0.015	6.25 ± 0.26	974

4.2 SQ2: Results of the Prediction Model Training

The diameter estimation model was trained again using 10-fold cross-validation with 20 epochs per fold. We used the original pinhole formula (see Equation 2.1) as the ground truth. The total dataset contained 51,531 images, but the model was again trained on a downsampled set, as described in Section 3.2. The model achieved an overall MAE of 0.411 cm with a standard deviation of 0.014 cm.

The prediction model was trained 100 times on five different configurations of pre-filtering, post-filtering, and whether the method of Mateijsen et al. [7] or the deep learning model was used to estimate the diameters. For each configuration, this resulted in 100 mean MAE values with their standard deviation, calculated over the MAE values of the 10 folds for each iteration. The best, median, and worst iterations were collected, ordered by MAE value. As shown in Table 4.4, the original dataset with the original filtering configuration (without pre-filtering and with post-filtering) performed the best in all three columns. However, the improvement is only marginal; the MAE values in the best, median, and worst columns are all within 0.02 cm of each other across the different configurations.

Table 4.4: Validation results of the prediction model for the different configurations. MAE values are reported with their standard deviations in parentheses.

Dataset	Pre-filtering	Post-filtering	Best MAE (cm)	Median MAE (cm)	Worst MAE (cm)	Training set size
Original	YES	YES	0.5395 (0.0178)	0.5519 (0.0289)	0.5750 (0.0243)	2399
Original	YES	NO	0.5400 (0.0246)	0.5528 (0.0407)	0.5692 (0.0428)	2383
Original	NO	YES	**0.5266 (0.0241)**	**0.5405 (0.0333)**	**0.5584 (0.0581)**	2449
Deep learning	YES	YES	0.5373 (0.0261)	0.5525 (0.0277)	0.5695 (0.0390)	2374
Deep learning	YES	NO	0.5384 (0.0373)	0.5512 (0.0357)	0.5688 (0.0498)	2341

4.3 SQ3: Results of the Validation of the Model on an External Dataset

The Blok et al. [2] dataset contained 1,613 images, of which 1,443 satisfied the pre-filtering criteria. Therefore, the dataset of Mateijsen et al. [7] was randomly downsampled to 1,443 images to match this size. We then trained and validated both datasets in the same way, using 10 folds and 20 epochs per fold. The difference is that the Blok et al. [2] dataset was stratified by occlusion ratio, whereas the Mateijsen et al. [7] dataset was stratified by measurement date.

The resulting mean MAE over the 10 folds for the external dataset was 0.668 cm with a standard deviation of 0.043 cm between folds. The original downsampled dataset had a mean MAE of 0.512 cm with a standard deviation of 0.036 cm.

We also investigated possible correlations between the error rate and the occlusion ratio. The results are presented in Table 4.5 and Figure 4.2. As broccoli heads became more occluded (broccoli heads with a higher occlusion ratio), both MAE and MRE increased. The lowest MAE was observed in the occlusion ratio bin of <0.1, with a value of 0.467 cm, while the highest MAE occurred in the 0.8–0.9 cm bin, with a value of 0.972 cm.

In addition, we looked for a correlation between the error rate and the camera height, as in the first sub-question. The results showed a similar pattern: the error generally decreased as the distance between the camera and the broccoli heads increased, reached a minimum, and then increased again. These results are shown in Table 4.6 and Figure 4.3.

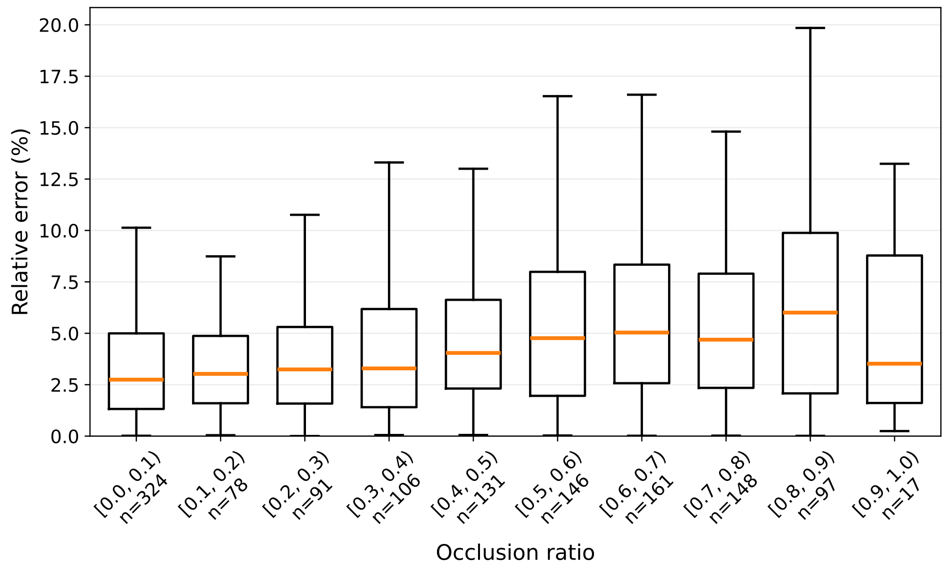

Figure 4.2: Mean relative error by occlusion ratio.

Table 4.5: MAE and MRE by occlusion ratio. Values are reported as mean ± standard error.

Occlusion ratio	MAE (cm)	MRE (%)	Bin size
$[0.0, 0.1)$	0.467 ± 0.021	3.49 ± 0.16	324
$[0.1, 0.2)$	0.509 ± 0.046	3.63 ± 0.34	78
$[0.2, 0.3)$	0.545 ± 0.048	4.00 ± 0.36	91
$[0.3, 0.4)$	0.610 ± 0.061	4.43 ± 0.39	106
$[0.4, 0.5)$	0.699 ± 0.052	5.38 ± 0.42	131
$[0.5, 0.6)$	0.749 ± 0.053	5.92 ± 0.47	146
$[0.6, 0.7)$	0.826 ± 0.054	6.15 ± 0.40	161
$[0.7, 0.8)$	0.814 ± 0.061	6.15 ± 0.46	148
$[0.8, 0.9)$	0.972 ± 0.091	7.20 ± 0.70	97
$[0.9, 1.0)$	0.856 ± 0.271	5.83 ± 1.64	17

Table 4.6: MAE and MRE by camera height. Values are reported as mean ± standard error.

Camera height	MAE (cm)	MRE (%)	Bin size
$[0.44, 0.49)$	1.433 ± 0.568	9.03 ± 2.76	5
$[0.49, 0.54)$	1.126 ± 0.157	7.37 ± 0.85	45
$[0.54, 0.59)$	0.711 ± 0.039	4.65 ± 0.23	293
$[0.59, 0.64)$	0.597 ± 0.019	4.59 ± 0.15	782
$[0.64, 0.69)$	0.767 ± 0.057	6.62 ± 0.55	162
$[0.69, 0.74)$	0.935 ± 0.278	9.43 ± 3.09	12

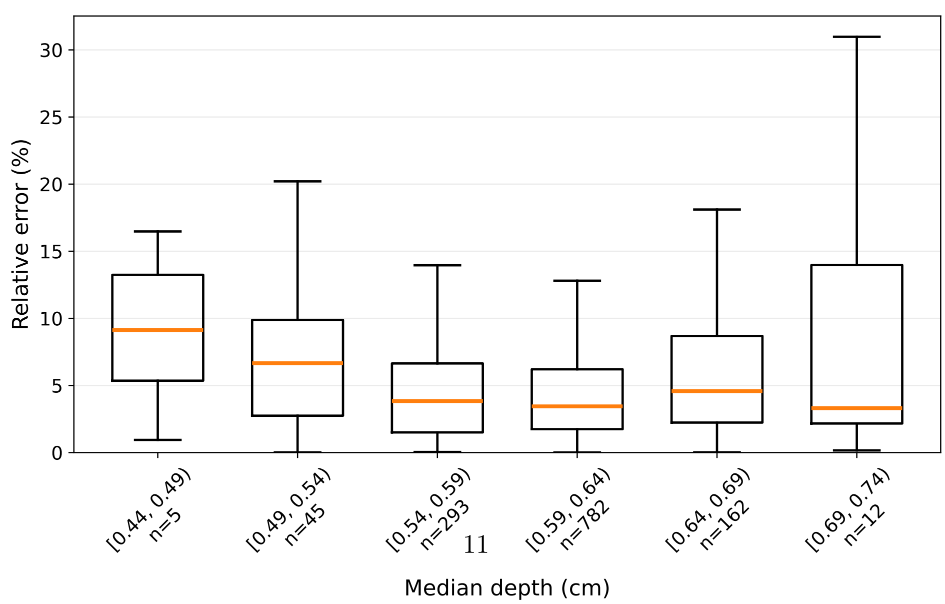

Figure 4.3: Mean relative error by camera height.

5 Discussion

5.1 SQ1: Diameter Estimation Accuracy on the Verdonk Dataset

The mean absolute error of all the predictions compared to the original dataset is 0.374 cm. This means that the diameter estimations of the deep learning model differed, on average, by 0.374 cm from the original dataset of Mateijsen et al. [7]. It is important to note that this dataset does not necessarily reflect the absolute truth of the real world. The ground truth is the original method of estimating the diameter. Because of this, the goal was not necessarily to compare the accuracy of our model against the real world, but rather to aim for precision similar to that of the original paper.

Another factor to keep in mind is the resolution of the input images. The weights we used were trained on images with a resolution of 224×224. Given that the focal length f of the camera was around 923 pixels and the input images originally had a resolution of 550, we can calculate the size of a pixel s at distance D from the camera with Equation 5.1.

$$s = \frac{D}{f * (224/550)} \tag{5.1}$$

Since most of the images of broccoli heads were taken between at a distance of 60–80 cm, the length of a pixel is between 0.16–0.21 cm. This number, divided by our MAE, gives a mean absolute error of around 2 pixels.

The results also show that the model performs best for camera distances that are well represented in the training set. Both absolute and relative errors increase for distance bins with fewer samples. This suggests that the model does not extrapolate well outside the training set. If this model were to be used, one should ensure that the training set contains the full range of camera distances expected during use.

Relative error is generally more informative than absolute error when evaluating the complete training set because the head sizes differ by a factor of 10 between the first and last measurement day. This is why we mainly used MRE to look for trends. The camera and different segments are all very similar in both their MAE and MRE, but for measurement dates this is slightly more complex. Generally, the RLE is stable around 5–6%, but on July 11th, the first measurement date is an outlier. This is because the mean diameter of the broccoli is 2.5 cm on that date, and then an absolute error of 0.3 cm already results in a relative error of 12%. Keeping this in mind, these are not unexpected results.

5.2 SQ2: The Influence of Using the Model on Growth Prediction Accuracy

The replacement of the original diameter estimation method with the deep learning model had only a small effect on the accuracy of downstream growth predictions. The different MAE values of the configurations were very close, differing by at most 0.02 cm. The original method with the original filtering setup still performed best, but the difference was very small. This suggests that the deep learning model can replace the original estimation method, but it does not necessarily improve the accuracy of growth predictions.

The impact of filtering is also interesting. It appears that post-filtering is most useful when no pre-filtering is applied. When pre-filtering is applied, the effect of post-filtering becomes slightly smaller. However, the original method without pre-filtering but with post-filtering still performs best. This indicates that the pre-filtering we used might also be removing useful data.

5.3 SQ3: Validating the Model with the External Dataset

The model performed worse on the Blok et al. [2] dataset than on the Verdonk dataset, even when downsampling the latter to the same size. This shows that the model architecture can be applied to another dataset to some extent, but the performance does not necessarily transfer. It is difficult to pinpoint why this is, but it could be due to the method of diameter estimation used by Blok et al. [3] being slightly more complex, with our current architecture not being able to replicate this. Another cause might be that the external dataset contains more occluded broccoli heads, and the results show an increase in the error rate when broccoli heads are increasingly occluded.

Another use of the external dataset was to look at the trends between the occlusion ratio and the error rate. This appeared to be a factor. The relative error doubles when we validate the model on broccoli heads that are more occluded than not. However, the MRE still only maxes out at around 7%, which could still be usable depending on the purpose.

6 Responsible Research

The research described in this paper is difficult to replicate because the original dataset is not publicly available. In order to fully replicate the results of this research, one must collect the data themselves using techniques similar to those described in this paper. The code used for data processing and experiments was developed to be as modular as possible, so that a person could process the data they collected themselves up to a certain 'checkpoint', from which point on they can use the published code from this paper to further process the data or run the experiments. The code can be found at `https://github.com/Rijkaron/broccoli-public`.

Furthermore, since we were working with closed source data, we had to pay close attention to how we handled and shared data. We fully processed the data locally. The data was shared using the approved Azure Blob Storage, SharePoint platform, or password protected zip files.

7 Conclusions

To recap, the main research question was *Can an end-to-end deep learning model provide a reliable alternative to the broccoli head estimation technique as proposed by N. Mateijsen et al. [7]?*. It was divided into the following three sub-questions:

1. *How can a deep learning model be designed to estimate the diameter of a broccoli head with precision similar to the method proposed by Mateijsen et al. [7]?*
2. *What effect does replacing the original diameter estimates with deep learning based estimates have on downstream broccoli growth prediction accuracy?*
3. *To what extent does the model's accuracy transfer to the dataset of P. Blok et al. [2]?*

To answer the first sub-question, this was done by using a model with two input branches: one for the image of a broccoli head and the other for a scalar denoting the distance from the broccoli head to the camera. The image was passed through a ResNet18 CNN model, the depth scalar through a small multilayer perceptron model. The outputs of these branches were processed in a regression head with one hidden layer, before outputting the single scalar, which was the estimated diameter. The model matched the results of the original method with a mean absolute error of 0.374 cm, or about 2 pixels. This is close enough to say that we have matched the accuracy of the method proposed by Mateijsen et al. [7]. Looking at the trends between the error rate and the camera height, it appears that the model is generally poor at extrapolating results. So, if this model were to be used in a production environment, it should be trained on the full range of camera heights that are expected during use.

To answer the second sub-question, although the original method of collecting data was slightly more accurate when looking at the MAE, the differences were marginal, being less than 0.02 cm. The different configurations of filtering techniques on the datasets used to train and evaluate the prediction model also had little impact on the final results. Again, this shows that the end-to-end deep learning model provides very similar results compared to the original approach of estimating diameters.

To answer the third sub-question, the model performed worse on the Blok et al. [2] dataset. Compared to the dataset of Mateijsen et al. [7], which was downsampled to match the size of the former dataset, it performed worse with an MAE of 0.668 cm with a standard deviation of 0.043 cm across folds, compared to an MAE of 0.512 cm with a standard deviation of 0.036. This is explained by the Blok et al. [2] dataset containing more occluded broccoli heads. Additionally, the method of calculating the ground truth for this dataset (the algorithm used to estimate the broccoli head diameters) was slightly more complex and therefore more difficult to fit with our minimal architecture. We also noticed that the error rate would generally increase as the broccoli head became more occluded. This is expected but, depending on the purpose of the data, could still be accurate enough.

In short, we designed and trained an end-to-end deep learning model which could be used instead of the original method of estimating diameters, while maintaining a similar level of (downstream) accuracy. So, the answer to the main research question is yes.

The implications of this research are not very strong for now. Although this method of diameter estimation can be used in the original data processing pipeline, it has only been shown to achieve similar accuracy, while being more complex and using more resources. This is not only because training the model requires substantial computational resources, but also because running the model on an input image and camera height value uses more resources than evaluating simple formulas. However, if the accuracy of this model can be shown to be better than the original method, it could be a viable replacement.

7.1 Future Work

As stated in the previous paragraph, when the accuracy of the end-to-end deep learning model increases, it could become useful. There is, however, one problem with training an end-to-end deep learning model on a generated dataset that we touched on briefly before: a model will only ever be as good as its dataset. Therefore, to make meaningful claims about the accuracy of a model on real world data, it is important that the training set matches the real world ground truth. In this case, this can be done by manually measuring broccoli heads, or it may be possible to show that a generated dataset matches real world data in some other way.

If a better dataset is found, there should be a simple method to further improve the model, namely by doing hyperparameter tuning. This was something we did not have time for, so we settled on an architecture and training parameters that worked well enough. However, there is probably still quite a lot to gain here.

References

[1] Jason Ansel, Edward Yang, Horace He, et al. Pytorch 2: Faster machine learning through dynamic python bytecode transformation and graph compilation. In *Proceedings of the 29th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 2 (ASPLOS '24)*. ACM, 2024.

[2] Pieter M. Blok, Eldert J. van Henten, Frits K. van Evert, and Gert Kootstra. Data underlying the publication: Image-based size estimation of broccoli heads under varying degrees of occlusion. 4TU.ResearchData. `https://doi.org/10.4121/13603787`, 2021.

[3] Pieter M. Blok, Eldert J. van Henten, Frits K. van Evert, and Gert Kootstra. Image-based size estimation of broccoli heads under varying degrees of occlusion. *Biosystems Engineering*, 208:213–233, 2021.

[4] Centraal Bureau voor de Statistiek. Groenteteelt; oogst en teeltoppervlakte per groentesoort, 2025. `https://opendata.cbs.nl/#/CBS/nl/dataset/37738/table` (accessed June 2, 2026).

[5] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 248–255. IEEE, 2009.

[6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[7] N. Mateijsen. Field-based predictive growth modeling of broccoli (Brassica oleracea var. italica) within precision agriculture. Master's thesis, Delft University of Technology, 2025. `https://repository.tudelft.nl/record/uuid:2bb6b9b4-c552-4286-ac5d-676bb0db9122`.

[8] L. R. Walton and J. H. Casada. Evaluation of broccoli varieties for mechanical harvesting. *Applied Engineering in Agriculture*, 4(1):5–7, 1988.

A Usage of Generative AI

During the course of this research project, different OpenAI models were used to assist in writing the code to conduct the experiments. In addition, large language models were sometimes used to help with spelling and improve readability. All changes proposed by these models were validated by the authors.