agronomy

MDPI

Article

A Comparative Dataset of Annotated Broccoli Heads Recorded with Depth Cameras from a Moving Vehicle

Oliver Hardy *, Karthik Seemakurthy and Elizabeth I. Sklar

Lincoln Institute for Agri-Food Technology, University of Lincoln, Lincoln LN6 7TS, UK; karthikvjit@gmail.com (K.S.); esklar@lincoln.ac.uk (E.I.S.)
* Correspondence: ohardy@lincoln.ac.uk

Abstract: An extensive, publicly available dataset is presented—the LAR Broccoli dataset—which contains 20,000 manually annotated images of broccoli heads captured from a moving tractor at an organic farm in the UK. The dataset contains images of the same row of broccoli heads recorded at 30 frames per second (fps) with three different cameras. Two off-the-shelf, relatively low-cost depth-sensing cameras were used, with the tractor moving at a speed of around 1 km/h, in addition to a webcam, with the tractor moving twice as fast. The utility of the dataset is demonstrated in four ways. First, three different state-of-the-art detector models were trained on the dataset, achieving an overall mean Average Precision (mAP) score of over 95% for the best-performing detector. The results validate the utility of the dataset for the standard task of in-field broccoli head recognition. Second, experiments with transfer learning were conducted, initialised with a smaller pre-trained broccoli detection model, and refined with the LAR Broccoli dataset. Third, we assessed the advantages of transfer learning not only using mAP but also according to time and space requirements for training models, which provides a proxy metric for energy efficiency, a practical consideration for real-world model training. Fourth, the cross-camera generalisation among the three camera systems was compared. The results highlight that testing and training detector models using different camera systems can lead to reduced performance, unless the training set also includes some images captured in the same manner as those in the test set.

Keywords: broccoli; selective harvesting; transfer learning; annotated dataset; annotations

Citation: Hardy, O.; Seemakurthy, K.; Sklar, E.I. A Comparative Dataset of Annotated Broccoli Heads Recorded with Depth Cameras from a Moving Vehicle. *Agronomy* **2024**, *14*, 964. https://doi.org/10.3390/agronomy14050964

Academic Editor: Yanbo Huang

Received: 1 April 2024
Revised: 25 April 2024
Accepted: 29 April 2024
Published: 3 May 2024

1. Introduction

Broccoli is a high-value open-field crop that requires labour-intensive manual harvesting in order to maximise yield. Manual harvesting is physical, back-breaking work with a machete, often in uncomfortable working conditions (e.g., heat, humidity, direct sunlight), extending for long hours in the peak of the season. The limiting factor for mechanising the harvesting process is that the broccoli heads do not mature uniformly; thus, a process of selective harvesting is undertaken where only ripe heads are cut in order to maximise yield. This process requires an ability to discern which broccoli heads are ripe and which are not. With selective harvesting by hand, a balance must be struck between allowing broccoli heads to grow to the optimal size and the availability of human work crews to send out for harvesting, often repeatedly to the same field. Further, considering the exacting standards from retailers as to the size and specification of broccoli heads that they will purchase (e.g., colour, shape), this can lead to high levels of in-field waste crop. As the need for food production grows and the availability of labour decreases, there have been pushes towards developing AI-enabled autonomous harvesting solutions. As part of such a solution, an AI-guided vision model would generally be required to perform the task of identifying which broccoli heads are ready to be harvested.

While AI-powered solutions for identifying broccoli heads are an attractive option, they require robust, annotated image data on which to train models that, at a minimum,

detect broccoli heads and, ideally, also estimate head size. It is important that these datasets represent, as much as possible, the complete landscape of image variations, including features such as occlusions (e.g., leaves growing such that they obscure the complete broccoli head in the camera frame), different crop varieties (e.g., ironman vs. steel, purple sprouting vs. tenderstem), different growing environments (e.g., traditional farming vs. organic farming, where weeds are typically more prevalent within the crop), and varied weather and lighting conditions (e.g., dry vs. wet, sunny vs. cloudy), as well as variations in data collection methodologies (e.g., RGB vs. RGB-D vs. point-cloud images, single images vs. streaming video).

Here, we present the LAR (Lincoln Agri-Robotics) broccoli dataset, a publicly available, manually annotated dataset of broccoli heads captured from a tractor in motion. The data set has frame-by-frame annotations of video from three different off-the-shelf cameras facing down towards the same row of broccoli. An example image of the same head of broccoli taken by each of the three cameras, with annotations, is provided in Figure 1.

Figure 1. The same head of broccoli from each of the three cameras with polygonal annotations. **Top left**: Logitech, **Top Right**: RealSense, **Bottom**: Left and Right ZED-2 Stereo image. The left hand ZED-2 annotation is in a different colour (yellow) to differentiate between left and right.

The frame-by-frame annotations capture image effects related to the vibrations of a vehicle in motion driving across uneven farmland terrain, creating a dataset that is representative of a practical automated harvesting scenario. In support of the manual annotations, we provide benchmarking data of the annotated images compared to three state-of-the-art image detectors from two different pre-trained models. Whilst the key contribution of this paper is introducing the dataset itself, we have analysed subsets attributed to each of the three different cameras to highlight the challenges of testing and training broccoli detection models using image data from different camera sources or capture methods, as well as the benefits of transfer learning to address these challenges. The dataset's extensive annotations, use of off-the-shelf cameras, and benchmarking results against current state-of-the-art detectors make this a resource well-suited to those developing a selective harvesting system for broccoli. The dataset is available on Kaggle (https://www.kaggle.com/datasets/oliverhardy/the-lar-broccoli-dataset (accessed on 12 February 2024)).

2. Related Work

Our new dataset contributes to the existing broccoli-specific datasets by providing a high number of annotated frames captured using three different forms of relatively low-cost camera systems, all applied to the same crop on the same day. As a point of comparison, we highlight three previous studies of identifying broccoli with computer vision and depth-sensing camera systems that have contributed datasets publicly, as well as three additional datasets that used visual grading methods on RGB images to identify harvestable broccoli heads. Table 1 provides a summary of the six related datasets, including our new dataset, listing parameters such as the camera system, location, cultivar, frames per second, and total images in the dataset.

Kusumam et al. [1] provide RGB-D data from three sites in Lincolnshire, UK for the Ironman cultivar and one site in Spain with the Titanium cultivar. Across the three sites, a range of weather conditions and plant maturity levels are represented. The Kusumam dataset was collected from a tractor driving slowly over the crop and imaged with an RGB-D camera (RealSense D435 [2]) and a Kinect 2 [3] for creating point-cloud images. Within the dataset, there are three sets of annotations provided: two annotated datasets from a site in the UK and one from a site in Spain. The datasets were collected using framerates of 7.5 frames per second (fps) and 3.3 fps for the UK sites and 6.4 fps for the Spanish site. The total numbers of annotated frames are 2201, 1580, and 1518, respectively, with each dataset being extensively annotated with centroids covering each broccoli head and cluster points, both of which are derived from ground-truth measurements.

The next substantial dataset of broccoli images was that of Bender et al. [4]. Bender et al. created a multimodal 10 week time-series dataset of broccoli and cauliflower images captured using an autonomous robot in New South Wales, Australia. The plants were grown in beds subject to different growing conditions, such as levels of irrigation, to create variation in growth. The robot was equipped with a stereo imaging system constructed from two 2.3 Megapixel (MP) CMOS cameras [5], as well as thermal and hyper-spectral imaging capabilities, and a number of environmental factors were recorded with sensors that were not on the robot platform. Images were captured at an initial rate of 3–4 fps in the early growing stages, and the rate was then doubled in the second half of the growing season. The total number of annotated RGB images in this dataset is 1248 images, with a mix of broccoli and cauliflower at various stages of growth during the 10 week period. The annotations are for the left-hand-side stereo camera and are annotated using bounding boxes.

Blok et al. [6] created a large dataset that addressed a feature that was not prominently highlighted in the previous datasets. The major contribution from the work of Blok et al. is capturing broccoli heads with systematic levels of occlusion. Occlusion is the issue of a broccoli head being partially or fully obscured from the camera, typically by the large leaves of the broccoli plant, which makes identifying and correctly sizing broccoli heads more challenging. Blok et al.provide depth image data on two broccoli fields: one in Santa Monica, California, US and the other in the Netherlands. The Santa Monica dataset was created with a combined RGB camera (IDS UI-5280FA-C-HQ [7]) and monochromatic stereo-vision camera (IDS Ensenso N35 stereo [8]). In the Santa Monica dataset, the capture process was repeated with and without natural occlusion of the broccoli from surrounding leaves at a speed of 0.14 m/s, and the images were captured from a tractor. The second dataset, collected in the Netherlands, also used the RealSense D435 RGB-D camera. The camera was mounted on a static metal frame placed over the broccoli heads with differing levels of occlusion created by placing broccoli leaves over portions of the broccoli heads. This dataset provides four to ten frames per broccoli head in each location. The total number of annotated images is 2560 across both locations. Each image was rescaled and zero-padded to 1280×1280 resolution, and the annotations are a polygonal mask of the visible region of the broccoli head and a circular mask that encompasses both the visible and occluded regions of the total broccoli head.

In addition to these three available datasets, there have been a number of studies on broccoli detection that do not use depth cameras, instead training computer vision models

to classify RGB images as harvestable or not. Zhou et al. [9] created a dataset of 506 images of broccoli heads in Zhejiang province with three varieties: Zhenqing96, Taliv1, and Taliv2i. The images were captured using a semi-autonomous platform travelling at a speed of 1 m/s. The platform utilised a Canon EOS 90D positioned at a height of 1.5 m with an LED lighting system on the camera to illuminate the crop. The broccoli heads were graded for maturity based on the number of yellow buds visible on the head of the broccoli, and yield estimates were calculated by weighing the heads within 2 h of harvesting. This dataset is available on request from the corresponding author of [9].

Garcia-Manso et al. [10] created a dataset (which is not publicly available) of 6139 images from broccoli plantations around Badajoz, Spain. In this dataset, the position of the broccoli was annotated, and each head was classified by an expert as "harvestable", "immature", or "wasted". There was a small initial dataset collected with a Sony-DSC-S950 with a 1/2.3 inch CCD sensor and a resolution of 2592 × 1944 pixels (618 images) and a Nikon-E4600 with a 1/2.5 inch CCD sensor and a resolution of 2288 × 1712 pixels (57 images); the images were recorded from static positions at an unspecified height. The majority of the dataset was produced with a Sony IMX179 image sensor camera (1/3.2 inch sensor) placed on a tractor, automatically capturing over 10 fps while the tractor was moving at a speed of approximately 1 km/h (5464 images), simulating a possible real-world scenario in which images need to be processed at this rate. This work is the one that most closely resembles our own data collection process, except for our focus on using depth cameras to determine if a broccoli head should be harvested according to size rather than the visual grading methods applied to RGB images used by Garcia-Manso.

Most recently, Kang et al. [11] created a dataset of 770 images captured from a drone flying 1 m above the ground at a speed of 1 m/s (0.06 km/h). The drone (DJI Mavic Air 2) recorded video at 30 fps with a resolution of 1920 × 1080, which was filtered and resized to 512 × 512 to reduce the computational load. The broccoli are of the Youxiu variety grown in Zhejiang province and were classified by experts into one of four maturity categories. A key feature of this dataset is that images were captured at multiple times of the day and with varying lighting conditions. This dataset is available upon request from Kang et al.

Table 1. A summary of the datasets containing broccoli heads.

Dataset	Camera(s)	Loc.	Cultivar(s)	fps	Images
Kusumam et al. [1]	RealSense D435 and Kinect 2	UK UK ES	Ironman Ironman Titanium	7.5 3.3 6.4	2201 1580 1518
Bender et al. [4]	2.3 MP CMOS, thermal and hyper-spectral	AU	Broccoli and Cauliflower	3–8 per plant	1248
Blok et al. [6] (Wageningen dataset)	IDS (RGB, stereo) RealSense D435	US NL	Avenger Ironman	4–10 per plant	2560
Zhou et al. [9]	Canon RGB	CN	Zhenqing96, Taliv1, Taliv2i	N/A	506
Garcia-Manso et al. [10]	Sony, Nikon RGB	ES	Parthenon	10	6139
Kang et al. [11]	DJI Mavic 2 camera	CN	Youxiu	30	770
Hardy, Seemakurthy, & Sklar (LAR Broccoli)	RealSense D435, Zed-2, Logitech RGB	UK	Ironman	30	27,646

3. Data Collection Methodology

This section describes the methodology that we employed for collecting the LAR Broccoli dataset and annotating the data. The overall concept underlying the need for such a dataset is that of an automated harvesting machine, either powered on its own or pulled by a farm vehicle (e.g., tractor), passing over the crop with a camera facing down. Detection, cutting, and retrieval of broccoli heads occur continuously whilst the vehicle is in motion. A similar concept is shown by Blok et al. [6], and this a common configuration for an autonomous harvesting system [12].

3.1. Camera Selection

For the creation of the LAR Broccoli dataset, three different imaging systems were utilised: a Logitech Brio camera [13], which captures RGB images; a RealSense D435 depth camera, which creates an RGB-D image using a combination of structured light and stereoscopic imaging technology; and a ZED-2 [14], which is an off-the-shelf stereo camera system. The RealSense D435 camera has previously been used in agricultural contexts including the creation of a dataset by Kusumam et al. [1] and the Wageningen [6] dataset (both discussed in Section 2). Both Blok et al. [6] and Bender et al. [4] have previously used stereo imaging with two independent cameras, but the ZED-2 is a single unit that is comparable in price to the RealSense camera. The performance of the RealSense and ZED camera systems has previously been compared for agricultural and robotics uses by Condotta et al. [15] and Tadic et al. [16]. Both studies recognise that the two different systems are capable of the task of capturing images with depth, but Condotta et al. point out that there are advantages to using stereoscopic technology in natural light, and Tadic et al. notes higher precision and accuracy for the ZED camera for sizing tasks.

3.2. Data Collection

The LAR Broccoli data was collected in October 2022 towards the very end of the harvest season from a row of the Ironman cultivar at an organic farm in Nottinghamshire, UK. The average height of the broccoli plants was approximately 55 cm to the top of the head, and this crop had had at least one previous manual harvest performed on it before our data capture. The row of broccoli captured features a range of broccoli head sizes, several cut stems, and weeds. On the day of data capture, the weather was overcast, providing even lighting conditions. The three different imaging systems were mounted on a custom-built frame attached to the rear of a standard agricultural tractor and raised to a height of 1 m above ground level. The position of the camera on the rear of the tractor is highlighted in Figure 2.

The speed of the tractor was determined by the conditions of the field and requirements set by industry collaborators who were developing a harvester that can operate between speeds from a minimum of 1 km/h up to a speed of 3 km/h. Variations in speed are an important consideration for a detection system, as a faster-moving tractor can cover more ground but will create blurrier images. As an indication of the blurriness of each image set, for each image, a Laplacian filter was applied to a greyscale version of the image to identify edges, and then the variance was measured across the filtered image. The lower the variance, the less well-defined the edges are in the image, which is associated with a blurrier image. A visual comparison of two images with different blur values from the Logitech dataset is presented in Figure 3. The variance was calculated for each filtered image, and the total values for each camera dataset were averaged to provide an indication of the relative blurriness of each image set. These values, alongside image sizes and tractor speeds, are recorded in Table 2. The dataset in which the tractor was travelling the fastest (Logitech dataset) was the blurriest, and all three datasets had broad standard deviations, showing that a vehicle in motion will have irregular blurring patterns unless some form of dampening or filtering is applied.

Figure 2. The data capture process using cameras mounted to the rear of a tractor. The position of the camera is highlighted with a red box.

Figure 3. An example of two images within the Logitech dataset with different levels of blur; the left image has a variance of 40, and the right image has a variance of 6.

Table 2. Image capture parameters related to image quality and tractor speed.

Camera	Tractor Speed (km/h)	FPS (sec)	Resolution (pixels)	Blur Average	Blur Standard Deviation
ZED 2	1.5	30	1920 × 1080	416	89
RealSense	1.2	30	1280 × 720	260	127
Logitech 4k	3.4	30	1920 × 1080	27	73

As the three cameras were used on the same crop, there is an accompanying chart with the dataset that shows which annotated object (i.e., broccoli head) in each dataset corresponds to another. The cameras were used on individual passes of the tractor over the crop and have different fields of view; therefore, there are differences in the number of broccoli heads observed in each camera's dataset. In addition, for the RealSense dataset, there was a loss in image capture for a period where the camera twisted from its downward orientation facing away from the crop; therefore, the dataset is split into two separate files, A and B (before and after the period of lost data, respectively). Due to the loss of light at the end of the day, after the three datasets were collected, we elected not to repeat the experiment and instead annotate the data that we collected. Within the RealSense B dataset, there are several broccoli heads that were captured at a tilted angle before the issue with the camera swivel was rectified; these heads have been annotated at the tilted angle. The

images that contained tilted broccoli heads are indicated in the dataset's accompanying datasheet, should a user choose to exclude these images. We found that the inclusion of these images created no significant issues within our analyses.

3.3. Annotation

Each image in the LAR Broccoli dataset was annotated by the same person using CVAT.AI (Computer Vision Annotation Tool), an online annotation tool [17]. For each frame in which a broccoli head is visible in the central row of the crop, a single polygon was carefully drawn around the perimeter of the largest visible area of a broccoli head. Where a broccoli head is completely exposed, the polygon will cover the total area of the head, and where, for example, a broccoli leaf bisects the head, creating two distinct visible regions of the same broccoli, only the larger of the two regions will be annotated. For the ZED-2 dataset, the left- and right-hand images of the stereo camera were independently annotated as separate items. It is important to note that we have annotated the Broccoli heads only along the central region of the image. The reason behind this choice of annotation was that the prototype harvester for which this dataset was created had the capability of collecting broccoli from a single row. There are multiple instances in the images where neighbouring unannotated broccoli heads are visible; these were excluded from our detector models' field of view by applying a mask to the side of the images. The mask was created to show the central 50% of the image, excluding the left- and rightmost 25% of the image The full unmasked images are provided, but using the images without further annotation or masking will lower the accuracy score of the models (see Section 4.2). The number of broccoli heads observed in each annotated dataset varied from 130 to 147 depending on the run due to changes in the field of view between cameras or levels of occlusion. The total number of images for each dataset, the number of images that were annotated (i.e., containing a broccoli head), and the total number of observed broccoli heads are presented in Table 3.

Table 3. Content of the LAR Broccoli dataset. RealSense A and B refer to data collected before and after the camera was physically re-mounted, respectively (see narrative). ZED-2 L and R refer to the left and right stereo cameras, respectively.

	Logitech	**RealSense A**	**RealSense B**	**ZED-2 L**	**ZED-2 R**
Total Frames	4421	2888	7340	8799	8799
Annotated Frames	2693	2888	5887	7993	8185
Total Broccoli Heads	130	23	120	147	147

3.4. Limitations

Our dataset is suitable for testing and training for the segmentation of broccoli heads with a relatively high framerate in the field under realistic harvesting conditions but is not suitable for evaluating the classification of the broccoli heads as harvestable or not. The major limitation of our dataset is the lack of ground-truth data for the size of the heads of broccoli, such as the dimensions provided by Blok et al. [6] or maturity classifications provided by Garcia-Mansos et al. [10]. Both the RealSense and ZED-2 cameras can produce a depth map aligned to an RGB image; with ground-truth measurements of the broccoli head dimensions, it would have been possible to compare the sizing capabilities of the two depth cameras. A reasonable scenario for using this dataset would be to train or test a system using these cameras to draw a bounding box around an RGB image where a head of broccoli is and then use the aligned depth map image to evaluate the size of the bounding box in metric units.

Both the RealSense and ZED-2 cameras have the ability to record not only image data but also the accompanying depth map and motion data. On our day of data capture, we were unable to record in this format for the ZED-2 camera. Without these data, it is difficult to compare the processing speeds and real-time detection abilities of the two cameras.

To compensate for the loss of the aligned depth map image of the ZED-2, both the left- and right-hand cameras of the stereo camera were independently annotated. There are methods for reconstructing a 3D depth map from a pair of stereo images, such as matching keypoints using the SIFT algorithm [18], but without the ground-truth size data, this is of limited use. The depth map data, including IMU (Inertial Measurement Unit) data from the RealSense camera, are not included in the public dataset due to the large file size, but they are available upon request by contacting the corresponding author.

4. Dataset Evaluation

In order to evaluate the utility of the LAR Broccoli dataset, three different sets of experiments were conducted utilising transfer learning to pre-train the detector models on two different datasets, one general and one broccoli-specific. First, three different state-of-the-art detector models were trained using a large general dataset and refined with the LAR Broccoli dataset, achieving an overall mean Average Precision (mAP) score of over 95%. Second, experiments were initialised with a smaller pre-trained broccoli detection model and refined with the LAR Broccoli dataset. Third, the cross-camera generalisation among the three camera systems was compared. Each set of experiments is described in this section, and the results are presented. For each of the experiments in our dataset evaluation, the images for each camera were combined into three datasets, which are the Logitech dataset, the RealSense dataset, which comprises RealSense A and RealSense B datasets, and the ZED 2 dataset, which comprises the left- and right-hand stereo images combined. These three datasets are abbreviated as LB, RS, and Z2 respectively. Each of these datasets is randomly divided into training and validation sets with a 80:20 split, which we use for all the experiments described in the following sub-sections.

4.1. Standard Detection Experiments Using Pre-Trained Weights

The annotated dataset was used to compare three different detector models, YOLOv8, FasterRCNN, and DETR, for the task of detecting bounding boxes around the heads of broccoli in the images. Note that the aim of this work is to highlight the utility of the curated dataset and not to focus on fine-tuning the example models. Hence, we do not optimise the default hyperparameters that are given in the original implementation of these respective detectors. For each model, we compared two sets of pre-trained weights, which are the outputs of the models trained on previous related datasets used as the starting point for training on our new dataset. In this way, the models will have some previous knowledge of broccoli heads but in a different domain; for example, differences in camera angle, lighting conditions, or the cultivar observed. From the initial weights, our models were fine-tuned using the 80:20 training and testing split for each camera image set in the LAR Broccoli dataset. The use of pre-trained weights is an example of transfer learning and was previously used by Garcia-Manso [10] in the context of broccoli; they used the ImageNet dataset to pre-train their broccoli detection model. The first dataset used to pre-trained models in this study is COCO (Common Objects in Context), a large dataset of over 300,000 annotated images across 91 different categories, including several thousand images of broccoli heads [19]. The second dataset used to pre-train the models was the Wageningen dataset created by Blok et al., which was previously described in Section 2. COCO is a large but very general dataset and Wageningen dataset has fewer images, but the images are much closer in appearance to those in the LAR Broccoli dataset, which is to say that there is a smaller domain shift in the Wageningen dataset.

The first of the three types of detector compared is FasterRCNN [20]. FasterRCNN is an adaptation of a regional convolution neural network detector, a two-stage detector that first identifies regions of interest within an image using a Region Proposal Network (RPN). The output of the first stage is passed onto the region of interest (ROI) pooling layer, where the features corresponding to different proposals are normalised to equal length and are further fed as input into classification and bounding box regression heads where the final classification label and the coordinates of the objects of interest will be predicted.

Blok et al. [6] used the related Mask RCNN method for their work on broccoli occlusion. YOLOv8 (You only look once) [21] is the eighth iteration of a CNN detector first developed in 2015 [22]. As the name suggests, this is a single-stage detector and is generally faster and of a lighter weight than RCNN-based detectors. Yolov8 has five different versions (n, s, m, l, and x) and is named after the number of parameters used in its corresponding architecture (nano, small, medium, large, and extra large, respectively). The larger the network, the more time and energy are needed for training. Larger networks may lead to better accuracy, but this is dependent on having a sufficiently large training set of images. While YOLO- and RCNN-based models are well established, DETR (Detection Transformer) [23] is a newer form of detector using transformer architectures to make predictions. This method uses elements of CNNs but with a streamlined training pipeline. Two different forms of DETR (R50 and R101) were used, and they refer to the number of layers used in the CNN backbone: 50 or 101 layers, respectively. These three model families—RCNN, YOLO, and DETR—represent the current state of the art for image detectors.

To compare the detectors' performance when evaluated with the manually annotated dataset, the positions of the annotated broccoli heads were converted into bounding boxes. The bounding box is a rectangle that encloses the area of the broccoli heads. Bounding boxes drawn by the detectors were compared to the annotated positions using the common metric known as the mean Average Precision (mAP). This metric employs the concept of the intersection over union (IoU), encompassing the numbers of true positives, true negatives, false positives, and false negatives, as well as the discrepancy between where the model draws the bounding box and where the bounding boxes are drawn manually [24].

4.2. Benchmarking with Detectors Pre-Trained on the COCO Dataset

In Table 4, we compare the performance of the three different detectors described above (FasterRCNN, YOLOv8, and DETR) after being pre-trained on the COCO dataset and adjusted using the LAR Broccoli dataset captured with the Logitech, RealSense, and ZED-2 cameras. Each row represents the detectors that have been pre-trained on the COCO dataset, and the columns present the mAP scores for a randomly selected 80:20 training and testing split for each camera's image set, using a single combined image set for each camera. In addition to the mAP values, the precision and recall values for this experiment are listed, respectively, in Tables A1 and A2, which are included in Appendix A.

Table 4. mAP values used to compare the performance of the state-of-the-art detectors pre-trained on the COCO dataset for the LAR Broccoli dataset. The highest score for each dataset is in bold.

Detectors Pre-Trained on the COCO Dataset	Logitech	RealSense	ZED-2
FasterRCNN	91.91	95.37	93.76
YoloV8n	96.6	98.4	97.3
YoloV8s	**97.0**	98.5	**97.5**
YoloV8m	96.7	**98.8**	96.7
YoloV8l	96.5	98.5	96.4
YoloV8x	94.3	98.7	96.2
DETR-R50	75.4	80.4	80.0
DETR-R101	74.9	76.8	78.2

The first conclusion that we can draw from these results is that the accuracy scores across the cameras for the CNN detectors (Yolo and FasterRCNN) are in excess of 90, which indicates that these are good-quality annotations. The DETR results were significantly lower; this is somewhat surprising due to the transformer-based architecture, but it is likely the case that the COCO dataset, whilst large, is not sufficient to train a very accurate model using DETR. It is important to consider not only the accuracy but also the steps to achieve

this accuracy and if it is feasible to create a large enough dataset of broccoli heads for training the DETR model. The best-performing camera was the RealSense camera, and the worst-performing camera was the Logitech camera, but this was only a slight difference that was most apparent in the FasterRCNN results. The best-performing detector was the different forms of YOLOv8, which worked very well with all three image subsets. There was limited difference among the different versions of YOLOv8, so we can conclude that it would be advantageous to use the smallest version of YOLOv8 when time and space are practical considerations.

4.3. Benchmarking with Detectors Pre-Trained on a Broccoli-Specific Dataset

In this experiment, we repeated the training and testing process for each of the three cameras in the LAR Broccoli dataset using detectors that had previously been trained on the Wageningen dataset rather than the COCO dataset. The same 80:20 testing and training split as in Table 4 was used. The performance of these detectors as measured with the mAP is presented in Table 5. For completeness, the precision and recall values for this experiment are included in Tables A3 and A4, which can be found in Appendix A.

Table 5. Performance comparison of the state-of-the-art detectors pre-trained on the Wageningen dataset for the LAR Broccoli dataset. mAP values are shown. The highest score for each dataset is in bold.

Detectors Pre-Trained on the Wageningen Dataset	Logitech	RealSense	ZED-2
FasterRCNN	91.42	95.71	93.07
YoloV8n	95.9	97.8	96.5
YoloV8s	96.3	**98.5**	97.2
YoloV8m	96.2	98.4	**97.4**
YoloV8l	96.3	98.4	96.5
YoloV8x	**96.4**	98.2	96.3
DETR-R50	73.6	79.2	82.3
DETR-R101	72.5	78.5	80.1

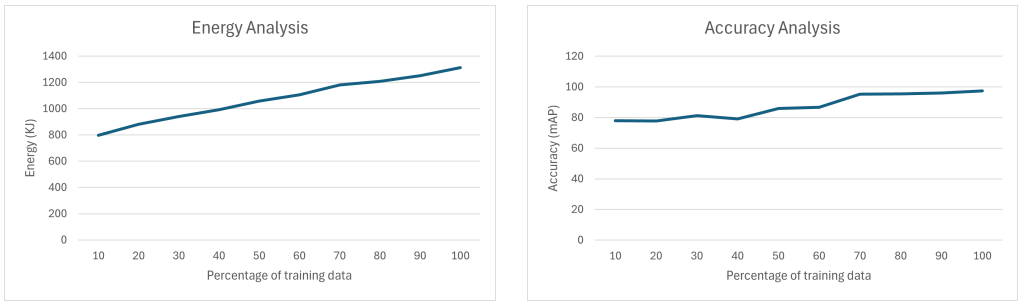
In comparing Table 5 with Table 4, the mAP scores of both the detectors pre-trained on the larger COCO dataset and the detectors pre-trained on the more closely related Wageningen dataset are similar, with a slight advantage for the COCO-trained dataset in terms of mAP. In terms of the mAP, this result shows that using a model pre-trained on COCO is the better option. However, the value of transfer learning needs to be evaluated in terms of reductions in training time, reduced energy consumption, and the need for smaller annotated datasets. In a variable environment such as an organic broccoli farm, there will often be domain shifts between the appearance of broccoli in the field ready to be harvested and the example images that a detector model has been trained on—for example, different weather or lighting conditions. A broccoli producer needing to re-train their model might be willing to sacrifice a small amount of precision for a more simplified training process to prepare the new model. In this respect, the Wageningen dataset converged to the final training outputs faster than the COCO dataset on account of the smaller domain shift than that of the COCO dataset for the LAR Broccoli dataset. An evaluation of the energy savings that can be achieved with a reduced dataset, as well as a method of more energy-efficient training of models using transfer learning, is described in Section 4.4.

4.4. Advantages of Transfer Learning

In this subsection, we demonstrate an experiment using the YOLOv8s detector, the best-performing detector identified in Section 4.2, and we propose a strategy to save energy consumed during the model training process by reducing the number of images for training.

By using the pre-trained weights from the Wageningen dataset, we evaluated the mAP score for each image in the training split of all the images across the three cameras combined into a single partition with no additional training. Each image in the partition was then sorted from the lowest to the highest mAP score. While both the LAR and Wageningen dataset are images of broccoli from a similar angle, there are differences in the capture methods and location between the LAR dataset and the Wageningen dataset, and there will be instances where the pre-trained model will fail to identify broccoli heads correctly in the new data. By ranking the new training data by the mAP, the lowest-scoring images that most diverge from the original dataset are those that are more valuable to train a model with. The highest-scoring images provide little or no new features to the model, and their inclusion may not improve a new model beyond the pre-trained weights.

Figure 4a indicates the energy needed for various proportions of sorted training data. Energy values for training the model were computed used Nvidia utilities (nvidia-smi). Additionally, the larger the training set, the greater amount of time that the model needs to converge, as the time taken for each epoch increases, which also impacts the total GPU energy consumption. From Figure 4, it can be seen that the energy increases almost linearly with the increase in the size of the training set, regardless of if the images improve the detector's accuracy or not.



(**a**) (**b**)

Figure 4. Advantage of transfer learning. (**a**) Energy analysis; (**b**) accuracy analysis.

Figure 4b indicates the quantitative performance with respect to the percentage of sorted training data that were used for the model. From the graph, the mAP score converges at 70% of training data, with the additional 30% providing little to no improvement but still consuming energy and time. The accuracy improvement is not linear, and the growers and producers using such a system may need to consider if they want to expend the time and resources to train a model beyond the base performance. Although we demonstrated the importance of transfer learning using the combined dataset (LB + RS + Z2), the savings in energy and time remain the same across any of the dataset combinations. The testing data achieved an mAP score of 80 before the additional training data, which, depending on the time of harvest and how easy or difficult it is to deploy the harvesting system, may be of sufficiently high accuracy. These results indicate that even if a large training dataset is collected, only a subset of the data need to be annotated and used to train the model, which can potentially save significant time and energy, paving the way for building sustainable AI systems.

4.5. Cross-Camera Generalisation

Our final analysis compared mixing the testing and training data between cameras to show how well the model trained on the images from one camera responded to new images from another camera or combinations of different cameras. This is an important consideration when selecting an off-the-shelf camera as part of developing a harvesting system. Does the selection of a camera early in development lock the user into using this camera should the camera be discontinued, a new camera becomes available, or prevent cross-use of data collected from other projects? To assess this question, the 80:20 training

and testing splits that were partitioned for the analyses in Tables 4 and 5 were combined into various combinations of the the three cameras, and the model was then run using the best-performing detector in terms of accuracy, which was YOLOv8s pre-trained on the COCO dataset. The results of this analysis are presented in Table 6. As an example, the fourth column in the first row shows the mAP score for the YOLOv8s model using the COCO pre-trained weights when trained on the training split of the Logitech data and tested on the combined validation split of the Logitech and RealSense data. The data in Table 6 have been colour-coded by the range of mAP scores from red (lower) to green (higher) to more easily visualise the distribution of the mAP score among the different combinations of testing and training datasets.

Table 6. Cross-camera generalisation performance comparison using YOLOv8s. Rows represent the training dataset, and the columns represent the testing dataset. Each data point is an mAP value; values in excess of 95 are coloured green, those between 85 and 95 are coloured in yellow, those between 80 and 85 are coloured in orange, and those below 80 are depicted in red.

mAP Scores	LB	RS	Z2	LB + RS	LB + Z2	RS + Z2	LB + RS + Z2
Logitech (LB)	96.3	83.3	52.2	87.6	77.3	74.5	80.5
RealSense (RS)	84.2	98.3	64.9	94.3	76.0	91.2	89.4
ZED-2 (Z2)	38.4	70.0	96.6	59.0	61.8	71.4	63.3
LB + RS	96.5	98.3	65.55	97.8	84.4	91.1	92.4
LB + Z2	96.3	87.0	96.6	89.8	96.4	89.8	90.8
RS + Z2	85.9	98.5	96.6	95.0	90.1	98.1	95.2
LB + RS + Z2	96.6	98.1	95.6	97.6	96.2	97.7	97.3

The results from the cross-camera generalisation experiment indicate that the RealSense is the camera that generalises the best and that the ZED-2 camera has the lowest ability to generalise to the other cameras. The ZED-2 dataset was the lowest performing as a single camera when used as both the testing camera and the training camera. Possible reasons for this may be that the ZED-2 dataset has a wider field of view, the inclusion of part of the tractor frame in the image, and a difference in the positions of the cameras (left and right stereo vs. single camera). This is an interesting result, as it suggests that the camera position and field of view are bigger influencing factors in generalisation than different vehicle speeds, though this needs to be investigated further. Although the drop in the mAP score was greatest going from or to ZED-2, the other instances of changing from one camera type to another also saw a significant drop in performance. There is a clear penalty from adapting from one camera system to another, but this can be offset by including training data collected in the same manner as the testing set. Another point of interest is that the mAP scores for any combination of two cameras used in the training set when tested on a third camera not included in the training set (e.g., training using LB + RS and testing on Z2) produced a slightly better score than that of either of the single datasets (e.g., LB testing and Z2 training). The best results are those that include the same camera in the testing and training data; but if using a new camera system with no available training data, then a diversity of camera systems may provide a small improvement in the mAP score.

5. Conclusions

The LAR Broccoli dataset is an available resource for those looking to utilise off-the-shelf depth cameras in an AI-enabled broccoli harvesting system. The dataset is manually annotated with polygons drawn around broccoli heads in the image for 30 fps footage recorded from a tractor using three different cameras. This dataset follows from the work of annotated datasets created using depth-sensing cameras from Kusuman [1], Bender [4], and Blok [6] by focusing on capturing image data in motion with different camera types.

The LAR Broccoli dataset was evaluated using state-of-the-art detector models (YOLOv8, FasterRCNN, and DETR) pre-trained on two different datasets, showing YOLOv8 to be

the best performing for all three detector architectures; FasterRCNN had a similar level, and DETR was substantially less accurate with this dataset. The current state-of-the-art detectors perform well in the task of identifying the position of broccoli heads in our dataset, though for such tasks, the mAP should not be the only factor considered. For example, using the smallest version of YOLOv8 (nano) with transfer learning to pre-train the model is not the highest-scoring detector, but it is an easier-to-train model without sacrificing a substantial loss in precision. The metrics that are used to evaluate AI systems in agriculture need to reflect the needs of the farmers and resources required, not just the ability to detect crops. Transfer learning has many benefits in improving the practicality of an AI-based selective harvesting—for example, the benefit that we demonstrated in transfer learning saving energy consumption during the training process for a broccoli detection model.

Comparing three camera systems, we found that in the task of detecting broccoli heads, the dataset created with the RealSense D435 camera was the best performing on both the models trained on the COCO dataset and those trained on the dataset from Blok [6]. The RealSense dataset was also able to cross-generalise the best from or to the other camera datasets. These results alone are not enough to unequivocally recommend one camera over the other, as the mAP scores were high for all three cameras. The determining factors will need to be how the cameras fit into the pipeline of a complete harvesting system. Although either of the depth cameras is suitable for such a system, there is an associated drop in performance in mixing the camera type for testing and training. Therefore, the recommendation is that current off-the-shelf cameras and detector systems work well, but it is better to select one camera and aim to standardise the method of image capture as much as possible. The imaging systems and detector architecture are currently at a high level, and there now needs to be a push for more diverse datasets; in particular, there should be efforts to create and distribute datasets with variable lighting conditions, such as in the work of Kang et al. [11]. The final note is that image detectors and the availability of high-quality off-the-shelf cameras have matured, but there needs to be a greater consideration of the aspects of an automated broccoli harvester system after the broccoli head has been identified. If there are established methods of mechanically separating and gathering the broccoli head from the plant, then it will be easier to define the parameters of a broccoli detection system. With the release of this dataset using off-the-shelf cameras, we encourage its use to further the development of a complete broccoli harvesting system.

Author Contributions: O.H. and K.S. collected the data and conducted the machine learning experiments. O.H. annotated the dataset and took the lead on writing this paper. E.I.S. supervised the work and contributed to writing this paper. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Innovate UK grant number 10028115.

Data Availability Statement: The dataset for this paper is available at https://www.kaggle.com/datasets/oliverhardy/the-lar-broccoli-dataset accessed on 12 February 2024.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A

In this supplementary section, we present some additional quantitative results to complement our results in Tables 4 and 5. The aim is to enhance the broccoli detection performance of the machine vision system onto which our trained models are finally deployed. This will enable the system to accurately identify the broccoli heads. In addition to maximising the broccoli detections, it is important to note that precise localisation is very important, as it has a direct impact on harvesting the fully grown broccoli heads, as the diameter estimation relies on accurate localisation. Hence, the choice of our model was completely based on the best-performing detectors reported in Tables 4 and 5, namely, the yolov8s and yolov8m models. Between these two detectors, it has to be noted that yolov8s has a smaller model size than that of yolov8m, which impacts the real-time detection perfor-

mance and is the reason for why we reported our cross-camera generalisation performance on yolov8s in Table 6.

Table A1. Precision values used to compare the performance of the state-of-the-art detectors pre-trained on the COCO dataset for the LAR Broccoli dataset. The highest score for each dataset is in bold.

Detectors Pre-Trained on the COCO Dataset	Logitech	RealSense	ZED-2
FasterRCNN	91.2	95.1	93.1
YoloV8n	96.1	97.1	95.9
YoloV8s	96.3	98.1	**98.1**
YoloV8m	**96.5**	**98.4**	96.8
YoloV8l	96.7	97.3	96.3
YoloV8x	96.2	97.8	97.2
DETR-R50	74.9	81.1	79.5
DETR-R101	74.1	77.1	78.5

Table A2. Recall values used to compare performance of the state-of-the-art detectors pre-trained on the COCO dataset for the LAR Broccoli dataset. The highest score for each dataset is in bold.

Detectors Pre-Trained on the COCO Dataset	Logitech	RealSense	ZED-2
FasterRCNN	89.2	93.1	91.2
YoloV8n	91.3	95.0	93.4
YoloV8s	**92.6**	95.8	92.7
YoloV8m	91.9	95.5	**94.3**
YoloV8l	92.5	**96.6**	91.7
YoloV8x	90.7	95.8	92.0
DETR-R50	74.9	79.1	79.2
DETR-R101	73.5	75.5	77.5

Table A3. Precision values used to compare the performance of the state-of-the-art detectors pre-trained for the Wageningen dataset. The highest score for each dataset is in bold.

Detectors Pre-Trained on the Wageningen Dataset	Logitech	RealSense	ZED-2
FasterRCNN	92.5	96.5	94.1
YoloV8n	97.3	98.4	97.1
YoloV8s	96.7	**98.8**	98.2
YoloV8m	98.2	97.9	**98.9**
YoloV8l	**99.1**	98.0	96.3
YoloV8x	97.7	97.2	97.8
DETR-R50	73.9	80.2	83.4
DETR-R101	73.1	79.5	80.4

Table A4. Recall values used to compare the performance of the state-of-the-art detectors pre-trained on the Wageningen dataset. The highest score for each dataset is in bold.

Detectors Pre-Trained on the Wageningen Dataset	Logitech	RealSense	ZED-2
FasterRCNN	89.9	92.1	91.7
YoloV8n	91.1	94.0	93.4
YoloV8s	92.3	94.6	93.7
YoloV8m	91.3	**96.4**	91.0
YoloV8l	91.3	95.3	93.4
YoloV8x	**93.2**	95.4	**94.0**
DETR-R50	71.2	80.4	80.1
DETR-R101	70.3	77.5	76.7

References

1. Kusumam, K.; Krajník, T.; Pearson, S.; Duckett, T.; Cielniak, G. 3D-vision based detection, localization, and sizing of broccoli heads in the field. *J. Field Robot.* **2017**, *34*, 1505–1518. [CrossRef]
2. Depth Camera D435. Available online: https://www.intelrealsense.com/depth-camera-d435/ (accessed on 12 February 2024).
3. Caruso, L.; Russo, R.; Savino, S. Microsoft Kinect V2 vision system in a manufacturing application. *Robot. Comput.-Integr. Manuf.* **2017**, *48*, 174–181. [CrossRef]
4. Bender, A.; Whelan, B.; Sukkarieh, S. A high-resolution, multimodal data set for agricultural robotics: A Ladybird's-eye view of Brassica. *J. Field Robot.* **2020**, *37*, 73–96. [CrossRef]
5. Grasshopper 3. Available online: https://www.flir.co.uk/products/grasshopper3-usb3/ (accessed on 12 February 2024).
6. Blok, P.M.; van Henten, E.J.; van Evert, F.K.; Kootstra, G. Image-based size estimation of broccoli heads under varying degrees of occlusion. *Biosyst. Eng.* **2021**, *208*, 213–233. [CrossRef]
7. UI-5280FA-C-HQ IDS Camera. Available online: https://www.loretech.io/products/ui-5280fa-c-hq (accessed on 12 February 2024).
8. Ensenso N-Series. Available online: https://en.ids-imaging.com/ensenso-n35.html (accessed on 12 February 2024).
9. Zhou, C.; Hu, J.; Xu, Z.; Yue, J.; Ye, H.; Yang, G. A Monitoring System for the Segmentation and Grading of Broccoli Head Based on Deep Learning and Neural Networks. *Front. Plant Sci.* **2020**, *11*, 402. [CrossRef]
10. García-Manso, A.; Gallardo-Caballero, R.; García-Orellana, C.J.; González-Velasco, H.M.; Macías-Macías, M. Towards selective and automatic harvesting of broccoli for agri-food industry. *Comput. Electron. Agric.* **2021**, *188*, 106263. [CrossRef]
11. Kang, S.; Li, D.; Li, B.; Zhu, J.; Long, S.; Wang, J. Maturity identification and category determination method of broccoli based on semantic segmentation models. *Comput. Electron. Agric.* **2024**, *217*, 108633. [CrossRef]
12. Bac, C.W.; Van Henten, E.J.; Hemming, J.; Edan, Y. Harvesting robots for high-value crops: State-of-the-art review and challenges ahead. *J. Field Robot.* **2014**, *31*, 888–911. [CrossRef]
13. Logitech Brio Datasheet. Available online: https://www.logitech.com/content/dam/logitech/en_gb/video-collaboration/pdf/brio-datasheet.pdf (accessed on 12 February 2024).
14. ZED 2 Datasheet. Available online: https://store.stereolabs.com/cdn/shop/files/ZED_2_Datasheet.pdf (accessed on 12 February 2024).
15. Condotta, I.C.; Brown-Brandl, T.M.; Pitla, S.K.; Stinn, J.P.; Silva-Miranda, K.O. Evaluation of low-cost depth cameras for agricultural applications. *Comput. Electron. Agric.* **2020**, *173*, 105394. [CrossRef]
16. Tadic, V.; Toth, A.; Vizvari, Z.; Klincsik, M.; Sari, Z.; Sarcevic, P.; Sarosi, J.; Biro, I. Perspectives of RealSense and ZED Depth Sensors for Robotic Vision Applications. *Machines* **2022**, *10*, 183. [CrossRef]
17. CVAT.AI Homepage. Available online: https://www.cvat.ai/ (accessed on 12 February 2024).
18. Lowe, D. Object recognition from local scale-invariant features. In Proceedings of the Seventh IEEE International Conference on Computer Vision (ICCV), Kerkyra, Greece, 20–27 September 1999; Volume 2, pp. 1150–1157. [CrossRef]
19. Lin, T.Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft coco: Common objects in context. In *Proceedings of the Computer Vision—ECCV 2014: 13th European Conference, Zurich, Switzerland, 6–12 September 2014, Proceedings, Part V 13*; Springer: Cham, Switzerland, 2014; pp. 740–755.
20. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster r-cnn: Towards real-time object detection with region proposal networks. *Adv. Neural Inf. Process. Syst.* **2015**, *28*, 1–9. [CrossRef] [PubMed]
21. Jocher, G. 2023. Available online: https://github.com/ultralytics/ultralytics/tree/main (accessed on 12 February 2024).
22. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You Only Look Once: Unified, Real-Time Object Detection. *arXiv* **2016**, arXiv:cs.CV/1506.02640. [CrossRef]

23. Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; Zagoruyko, S. End-to-end object detection with transformers. In *Proceedings of the European Conference on Computer Vision*; Springer: Cham, Switzerland, 2020; pp. 213–229.
24. Beitzel, S.M.; Jensen, E.C.; Frieder, O. MAP. In *Encyclopedia of Database Systems*; Liu, L., Özsu, M.T., Eds.; Springer: Boston, MA, USA, 2009; pp. 1691–1692. [CrossRef]