Advances in visual perception for agricultural robotics

Gert Kootstra, Wageningen University & Research, The Netherlands

1 Introduction

Like any autonomous system, an agricultural robot is in continuous interaction with the environment. There are two aspects to this interaction: the robot senses the environment and it acts on the environment. To decide on appropriate actions to fulfill its task, the robot needs to understand the state of the environment. That is, it needs to know about the relevant aspects of the environment that influence the decision-making process. Some of this knowledge might be expert knowledge describing facts or regularities that are generally true. However, as the agricultural environment is uncertain, variable, cluttered, and constantly changing, a fundamental part of the knowledge about the state of the environment needs to be extracted from the sensor data. The process to extract relevant information from sensor data is called perception.

Modern robots are equipped with numerous sensors, which can be divided into proprioceptive and exteroceptive sensors (Siegwart et al., 2011). Proprioceptive sensors measure quantities related to the internal state of the robot, such as the temperature of its electronics, battery charge, motor speed, and joint angles. Exteroceptive sensors measure quantities of the robot's environment, such as light intensity, distance, sound amplitude, and the earth's magnetic field. Although all these sensors are relevant to develop an autonomous agricultural robot, this chapter focuses on the use of vision-based

http://dx.doi.org/10.19103/AS.2023.0124.01
© The Authors 2024. This is an open access chapter distributed under a Creative Commons Attribution 4.0 License (CC BY).

Chapter taken from: van Henten, E. and Edan, Y. (ed.), *Advances in agri-food robotics*, pp. 3–36,
Burleigh Dodds Science Publishing, Cambridge, UK, 2024, (ISBN: 978 1 80146 277 8; www.bdspublishing.com)

sensors and range sensors providing two-dimensional (2D) and three-dimensional (3D) images of the environment, as these are the predominant exteroceptive sensors used in current robotics. Using such sensors, notably vision cameras and LiDAR sensors, a wealth of information about the spatial arrangement and spectral properties of the environment can be extracted through the process of visual perception.

Robotic perception in agriculture faces two major challenges: the challenge of dealing with variations in the environment and the challenge of operating with incomplete information. This chapter discusses advances in visual perception for agricultural robotics with respect to these two challenges and it provides an outlook on future research to better deal with the challenges. The aim of the chapter is to give a broad and comprehensive overview of current academic literature on this topic, to provide insights into the main research directions, and to propose directions for future research. Examples given in this chapter are mainly for arable, greenhouse, orchard, and livestock farming, as the author has personal expertise in these areas, but it must be noted that forest robotics is facing similar challenges. For an overview of forest robotics, the reader can refer to Billingsley et al. (2008) and Oliveira et al. (2021).

In Section 2, a short introduction to visual perception will be given, including a description of different sensors and sensor systems with their advantages and disadvantages and a set of typical perception tasks. Section 3 describes the main challenges for visual perception in the agricultural domain. Approaches to deal with these challenges presented in the literature are then discussed in subsequent sections. Section 4 discusses methods to deal with variation with a focus on the use of deep neural networks, and Section 5 discusses methods to deal with incomplete information, motivating the need for active perception. Section 6 proposes future research directions and the chapter is concluded in Section 7.

2 A short introduction to visual perception in agriculture

Visual perception is defined as the process to extract relevant information about the environment for the operation of the robot based on data from imaging sensors. This section covers different imaging sensors often used in agriculture (Section 2.1), image acquisition on robotic platforms (Section 2.2), and the visual information extracted from the sensors (Section 2.3).

2.1 Imaging sensors

Imaging sensors detect electromagnetic radiation and form an image to represent that information. An image is a 2D array with pixels representing the intensity of the radiation. Imaging sensors exist for different parts of

Published by Burleigh Dodds Science Publishing Limited, 2024.

the electromagnetic spectrum, such as radar, X-ray scanners, and Teraherz cameras. However, for agrirobotic applications, typically, cameras are used for wavelengths in the visual and (near) infrared (NIR) spectrum to get information about natural objects in the robot's surroundings under natural or artificial lighting. Hence, the remainder of this chapter discusses these camera systems. The camera measures the light reflected by the objects.

Red-green-blue (RGB) color cameras are popular for robotic systems. The quality, price, size, and energy efficiency of these sensors have improved drastically in the past decades with the introduction of mobile phones. They allow to get color information about the surroundings of the robot in high spatial resolution. This provides sufficient information to detect objects in the agricultural environment and to get geometrical and color information about these objects needed to make decisions for navigation and operation. In environments with natural illumination, such as arable fields or greenhouse, there can be high differences in the intensity in reflectance of sun-lit and shadow parts of the scene, which can cause over- or under-exposure of camera images. To deal with the large intensity difference between sun and shadow, cameras with a high dynamic range can be used, for instance, to segment vegetation and soil in the images (Suh et al., 2018a). Such high-intensity differences can also be prevented by using covers and artificial lighting (e.g. Arad et al., 2019).

To get more information about the status of the crop or its produce, spectral imaging systems (also called imaging spectroscopy) can be used. Where RGB cameras use three spectral bands (corresponding to red, green, and blue), spectral cameras use multiple bands in the electromagnetic spectrum, in some cases ranging up to hundreds of spectral bands (Polder and Gowen, 2021). In the past, this technique would be referred to as multi- or hyper-spectral imaging, but nowadays the term 'spectral imaging' is preferred (Polder and Gowen, 2021). The cameras provide high spectral resolution, combining the benefits of imaging systems with spectroscopy, allowing chemometrics to assess the chemical composition. Information about pigments in plants or fruits, such as chlorophyll or carotenoids, can be obtained from the visible part (400–700 nm) and information about, for instance, sugars, proteins, and water can be obtained from the NIR part (700–2500 nm). This can, for instance, be used for disease detection (Mishra et al., 2020; Polder et al., 2019) to establish food safety and quality (Qin et al., 2013) or to quantify variations in fish feeding behavior (Zhou et al., 2017). The downside of spectral cameras is the higher price and the lower spatial resolution.

For robots to be able to operate in the agricultural environment, it is important that they get 3D information about objects in the environment. To this end, depth cameras or light-detection-and-ranging (LiDAR) sensors are used (Horaud et al., 2016). Depth cameras, often in combination with color (so-called RGB-D cameras) get depth information typically from triangulation

using stereo vision (comparing left and right camera images) or using project light (comparing an emitted pattern with the reflected pattern observed in a camera image). LiDAR is based on the time-of-flight principle, where the time it takes for an emitted laser beam to reflect back in the sensor is measured to get depth. Although a LiDAR is strictly speaking not an imaging sensor, it can be used as such in combination with an RGB camera. Both RGB-D cameras and LiDAR can provide 3D point clouds, which are a set of points with their 3D position coordinates. Active 3D camera systems, like LiDAR and projected-light cameras, emit light onto the environment. These systems sometimes have trouble in outdoor situations with strong sun light interfering with the projected light or rain and snow disturbing the distance measurements, although the quality of these sensors is rapidly improving, making outdoor applications possible. Stereo-vision cameras are passive sensors that have no trouble dealing with outdoor situations. However, the downside of stereo vision is the lack of 3D information on non- or low-textured surfaces. Using 3D information allows, for instance, fruit detection and localization in orchards (Fu et al., 2020) or 3D localization and size estimation for robotic harvesting of broccoli heads in arable fields (Blok et al., 2021).

2.2 Image-acquisition systems

The acquisition of imaging data can be done with different robotic devices that can generally be divided into mobile robots – unmanned aerial vehicles (UAVs) and unmanned ground vehicles (UGVs) – and robotic arms (Corke, 2017). Unmanned aerial vehicles are often used for monitoring and phenotyping applications on arable fields (Tsouros et al., 2019; Xie and Yang, 2020; Chlingaryan et al., 2018). They have the advantage that they can easily cover large areas in a short time. Despite the high flying altitudes, modern high-resolution cameras still allow sub-centimeter pixel resolution, sufficient to monitor individual plant growth of, for instance, cabbage and pumpkin (Jamil et al., 2022). The downside of UAVs, however, is that they need relatively stable weather conditions and have a limited flight time. Furthermore, due to limited payload and the disturbance of the crop when flying at low altitudes, UAVs are not very suitable to perform operations.

Unmanned ground vehicles are able to get much closer to the crops, providing higher resolution imaging data and allowing robotic operations. Autonomous weeding robots, for instance, typically classify plants and detect their location based on high-resolution color and/or depth data to determine spraying actions (Ruigrok et al., 2020) or mechanical weeding actions (Pretto et al., 2021). Other examples are mobile robots in orchards for yield prediction (Bargoti and Underwood, 2017), and autonomous harvesting (Silwal et al., 2017; Williams et al., 2019). Another advantage of a UGV for image acquisition

Published by Burleigh Dodds Science Publishing Limited, 2024.

is the potential to use artificial light and a cover to shield off ambient light. The resulting controlled illumination conditions improve perception, for instance, to detect the subtle visual changes caused by potato diseases (Polder et al., 2019). In other work, a strong short-range flashlight is used to minimize the effect of ambient light, for instance, Arad et al. (2019).

Eye-in-hand systems, where a camera is mounted on a robot arm, make it possible to easily reposition the camera to observe a target object from different viewpoints, providing more complete information and dealing with occlusions (Hemming et al., 2014). In Barth et al. (2016), an eye-in-hand setup was developed to actively search for bell peppers to harvest in a greenhouse. Using visual servoing, the robot could subsequently find an optimal pose for harvesting (Arad et al., 2020). Using an array of nine cameras in the robot's hand, the next viewpoint that provided the most unoccluded view on the object of interest could be determined, which iteratively resulted in the best local view on the object (Lehnert et al., 2019). The topics of multi-view perception and active vision are discussed in more detail in Section 5.

Besides the use of the above-mentioned mobile robots, fixed image-acquisition systems are often used for plant phenotyping or monitoring tasks in confined spaces. Examples are Gantry systems that transport cameras and other sensors over the crop, a so-called sensor-to-plant system, for instance, Chaudhury et al. (2019), sometimes used in combination with a mobile robot (Nicolas et al., 2016), and plant-to-sensor systems, where plants are transported to a sensor station for detailed scanning, for instance, (Golbach et al., 2016; York, 2019). It must also be mentioned that the robotic imaging systems do not necessarily need to be fully autonomous, but can, for instance, also be mounted on implements that are still navigated with human assistance.

2.3 Visual information

The purpose of the image-acquisition system is to provide relevant information to the robot about the objects in its environment. On the one hand, this concerns information that is relevant to control the robot's actions, for instance, the detection and locations of plants or trees to navigate through a row (Bonadies and Gadsden, 2019; Blok et al., 2019), or the location of a ripe fruit to perform harvesting actions (Zhao et al., 2016). On the other hand, it concerns information that needs to be provided to the farmer or a management system, such as the count of apple blossoms to determine thinning actions and predict future yield (Bulanon et al., 2020) or the plumage condition of chickens to prevent welfare issues (Lamping et al., 2022).

Three types of information are distinguished that are relevant to extract from image data: (1) spatial information about objects, (2) information about

object properties, and (3) temporal information about objects. These types of information are discussed in the next paragraphs.

Object pose is an example of spatial information. Typical, object pose is defined as the 3D position and 3D orientation (six-dimensional pose) of an object. Methods have been developed, for instance, to predict the pose of fruits to provide information for grasping (Wagner et al., 2021; Guo et al., 2020). When there is rotational symmetry in the fruits, the dimensionality of the pose can be reduced to, for instance, 5D for apples (Kang et al., 2020). For animals, often another definition of pose is used. Animal pose is typically defined by a set of key points representing different body parts, providing information about the posture, for instance, Mathis et al. (2018) and Russello et al. (2022). Also, the object shape can be relevant, for instance, to grade kiwifruits (Fu et al., 2016), to determine the size and dimensions of orchard trees for variable-rate spraying (Zhang et al., 2018; Abbas et al., 2020), to estimate heifer height and body mass of young cows (Nir et al., 2018), or to estimate feed intake of individual cows (Saar et al., 2022). Furthermore, the spatial arrangement of objects in the environment is important information for robot operation, for instance, information about the location of trees for navigation in orchards (Blok et al., 2019), information about crop rows for localization and navigation in arable fields (Winterhalter et al., 2021), and the spatial arrangements of fruits to plan harvesting actions (Kurtser and Edan, 2020).

Object properties are visual and geometrical properties that provide information about the objects themselves. Some of these properties are low-level features that can directly be observed, such as color properties and texture, and others are more high-level features that can be derived from low-level features, such as ripeness and the detection of diseases. In Halstead et al. (2018), for instance, the ripeness of bell peppers on the plant is estimated from color images. Other examples are the detection and classification of defects of broccoli heads in the field based on color images (Blok et al., 2022), the detection of the potato virus Y on potato plants based on spectral images (Polder et al., 2019), and the automatic scoring of the body condition of dairy cows (Spoliansky et al., 2016).

Temporal information is crucial when analyzing the gait of animals. Based on video frames, the animal pose can be determined, for instance, for cows (Russello et al., 2022) and poultry (Doornweerd et al., 2021). By tracking the animal over time, the changes of the pose provide information about the gait, which can be used as a welfare indicator (Nasiri et al., 2022). Also for robot perception, it can be important to track objects over time, for instance, to prevent double counts when assessing fruit yield in a greenhouse (Smitt et al., 2021; Halstead et al., 2018; Halstead et al., 2021).

Published by Burleigh Dodds Science Publishing Limited, 2024.

3 Challenges in visual perception for agricultural robotics

In the previous section, many examples of agricultural robotic systems have been discussed. However, despite the active research field, there are not many commercially successful agricultural robots on the market yet. This is a sharp contrast with, for instance, the automotive or electronics industry, where many robots are operational in the production line to date. The main difference between the agricultural and automotive environment is the level of control over the environment. The environment for robots working in an automobile factory is completely under control, with controlled lighting, clean rooms, identical objects, and full knowledge about the location and shape of all the objects. This provides the robot with complete knowledge of the environment, which eliminates any uncertainties, allowing even pre-programmed behaviors. The agricultural environment, on the other hand, is far less structured, giving rise to uncertainty about the state of the environment. These uncertainties stem from variations in the objects and environmental conditions, and from incomplete information about the environment (Kootstra et al., 2020; Kootstra et al., 2021). In Kurtser and Edan (2018), a statistical model was developed to model the detectability of bell peppers depending on factors like the viewing distance and viewing angle. Their analyses clearly illustrate both the challenge of incomplete information and the variation over different greenhouses and cultivars of bell-pepper plants. Both challenges will be discussed in the next subsections.

3.1 The challenge of variation

The agricultural task environment contains a lot of variation. Four different types of variation are distinguished (Kootstra et al., 2020; Kootstra et al., 2021): (1) object variation, (2) environmental variation, (3) variation in cultivation systems, and (4) task variation.

1. The object variation is a consequence of natural variation. Every instance of a plant, fruit, vegetable, or animal is unique in appearance, geometry, mechanical properties, and behavior. Already within one species, the variation can be substantial and it is even more apparent between species. Also, non-living matter, such as soil comes in many different variations.
2. The environmental variations are, for instance, differences in weather conditions, causing differences in illumination (brightness and color) and humidity, and also deliberate changes made to the physical environment throughout production cycles, such as plowing and irrigation.
3. Different farmers use different cultivation systems. Apple trees, for instance, can be grown in orchards using the tall spindle system, v-trellis

system, or espalier support system, and animals can be raised outdoors in a meadow or indoors in a barn.

4 Task variation is caused by the need for farmers to use a robot not for one single task, but for a variety of different tasks. A robot that can be used for harvesting, deleafing, and rehanging of tomato plants, for instance, has more economical value to a grower than a robot that is specialized in harvesting only.

All these different sources of variation cause a lot of uncertainty for the robots operating in agricultural environments. Specifically, for visual perception, it results in a lot of variation in the appearance of the objects in the images and point-cloud data. This requires that perception methods can deal with this variation and that they can generalize to new environments. In Ruigrok et al. (2023), the generalization of a deep neural network in the agricultural context was tested using imaging data from many different sugar-beet fields, containing, among others, variation in cultivar, soil type, growth stage, and weed pressure. It was shown that even state-of-the-art deep neural networks suffer from a loss in performance when tested on a new field, the so-called generalization gap. Such a gap was also found when applying a trained neural network on a new variety of broccoli (Blok et al., 2020). In Section 4, approaches to deal with variation are discussed.

3.2 The challenge of incomplete information

An agricultural robot has only partial information about the environment due to (1) sensor limitations and (2) occlusions.

1 Apart from the limits in which sensors operate, such as the spectral wave lengths and field of view, imaging sensors occasionally fail in agricultural environments, especially under uncontrolled illumination. Examples are overexposure due to direct sun light, under exposure due to strong shadows, and lens flares. Image data are also often noisy, which is particularly apparent in depth data, which is incomplete at untextured surfaces and near object boundaries.
2 When objects are occluded, they are hidden from the camera view by other objects or by the object itself (self-occlusion). Especially in cluttered environments, such as orchards and high-wire greenhouses, occlusions form a substantial challenge for robots to perform their tasks. A study of the visibility of bell peppers in a greenhouse showed that even with a requirement of only 50% visibility, at best 69% of the fruits were visible in the images (Hemming et al., 2014). Combining five viewpoints could raise this to a 90% detection rate. Similarly, in Boogaard et al. (2020), 36

Published by Burleigh Dodds Science Publishing Limited, 2024.

viewpoints per plant were needed to reliably observe the leaf and fruit nodes in cucumber plants. Bulanon et al. (2009) analyzed the visibility of fruits on orange trees and reported that for a single viewpoint, only 19–47% of the fruits were visible, while combining 9 viewpoints resulted in 91% visibility.

In Section 5, approaches to deal with incomplete information, specifically with occlusion, are discussed.

4 Dealing with the challenge of variation

Deep neural networks boosted the ability of agricultural robots to deal with variation, as will be discussed in Section 4.1. However, even with the use of deep learning, variation in agricultural environments remains a problem as the generalization of perception methods is not sufficient. Section 4.2 discusses different approaches in the literature to minimize the generalization gap.

4.1 The era of deep learning

Generally speaking, a machine vision algorithm consists of two parts; a feature-detection part to extract relevant image features from the camera images and a decision-making part, where the features are used to perform perception tasks, such as segmentation, classification, or object detection. For many decades, developments in both parts were focused around hand-crafted algorithms. Many powerful feature extractors were developed, such as the scale-invariant feature transform (SIFT) (Lowe, 2004) and pyramids of histograms-of-gradients (Bosch et al., 2007), which could be used, for instance, to match the current camera image with features of objects stored in a database. Later, the decision-making part was replaced with machine learning algorithms. In (Suh et al., 2018b), for instance, a support-vector machine, random forest and neural network were used to classify sugar-beet and potato plants based on a bag-of-visual-words model using SIFT features. A downside of these methods, however, is that the quality of the machine vision algorithm is limited by the quality of the image features. And it turned out to be very challenging to develop feature algorithms that are invariant to various variations in the appearance of objects in agricultural environments.

With the success of AlexNet (Krizhevsky et al., 2012) in the ImageNet competition (Russakovsky et al., 2015), a new class of methods started to dominate the field of machine vision; those based on deep learning, specifically on convolutional neural networks (CNNs). With a combination of convolutional layers, pooling layers, and fully collected layers, these networks can jointly learn feature extraction and decision-making. With end-to-end learning, machine

vision tasks can be optimized based on a large training set with images and corresponding desired output values (Goodfellow et al., 2016).

These techniques also revolutionized machine vision for robotic applications in agriculture. In a survey of 40 studies on the use of deep learning for machine vision tasks in agriculture and food production, it was concluded that deep neural networks outperform other existing techniques (Kamilaris and Prenafeta-Boldu´, 2018). Zhu et al. (2018) provided an overview of deep learning for the use of smart agriculture. Deep neural networks were shown to be able to detect seven different types of fruits in Sa et al. (2016) using a multi-modal version of Faster R-CNN based on RGB and NIR images. Also when dealing with 3D point clouds, deep neural networks have shown their benefits (Boogaard et al., 2021; Shi et al., 2019) although in the 3D domain, more classical feature-based methods are still competitive (Dutagaci et al., 2020).

However, even with the use of deep neural networks, a generalization gap is often observed, limiting the extent to which agricultural robots can deal with variations in the environment. The difference in the distribution of the images in the training set and that of the new images during execution is causing a drop in performance. This was, for instance, observed when dealing with a new cultivar of broccoli (Blok et al., 2020) and when confronted with a new sugar-beet field with, among others, a variation in cultivar, soil types, growth stages, and illumination (Ruigrok et al., 2023). This limitation to deal with variation is a showstopper for commercial applications, as these need a guaranteed performance in any situation occurring in practice. Falling back on non-deep-learning approaches with limited but more predictable behavior can then be preferred. This calls for methods to minimize the generalization gap.

4.2 Minimizing the generalization gap

This subsection discusses a few approaches to minimize the generalization gap.

4.2.1 Transfer learning and image augmentation

To allow deep neural networks to generalize better to new situations – that is, to let them deal better with the variation present in the agricultural environment – it is important to apply regularization techniques, especially when there is limited available training data. Transfer learning, the process of first training a neural network based on a large general image dataset, such as ImageNet, followed by fine-tuning on a specific dataset in the relevant domain, was shown to also be beneficial for agriculture tasks (Chen et al., 2020; Thenmozhi and Reddy, 2019). Another common method is to apply image augmentation to increase the variation of the training set. This was shown, for instance, to increase the level of generalization to a new species of broccoli in Blok et al. (2020) and

improve the identification of sheep (Hitelman et al., 2022). Smart sampling techniques during learning can improve performance in the presence of class imbalance (Boogaard et al., 2022).

4.2.2 Composition of the training set

The generalization gap arises when the distribution of images in the training set is not covering the distribution of images that are encountered during execution (or in the test set). Figure 1 illustrates a two-class classification problem with the total distribution covering all possible variations in the shaded green and red areas. On the left, you see a typical situation, where training data is taken from only a limited set of different conditions (fields, cultivars, etc.). In such situations, there is a high change that the trained perception method is confronted with new situations that cannot be handled. One way to improve the quality of the training set is to take more training samples from the training conditions (bottom of Fig. 1). This does cover a somewhat larger part of the total distribution. However, large parts are still not covered. Getting training data from more different conditions is more effective to cover a larger area of the total distribution (top of Fig. 1). However, both are needed to properly deal with all variations that can be encountered during execution. This illustration matches with the results found in Ruigrok et al. (2023), where data from 25 different fields with sugarbeet and potato plants were collected, with variation in cultivars, spoil types, growth stages, and weed pressure, to train a deep neural network for plant detection. Keeping the number of training images fixed to 500, the performance on the five independent test sets was shown to increase significantly when training data were used from more different fields.

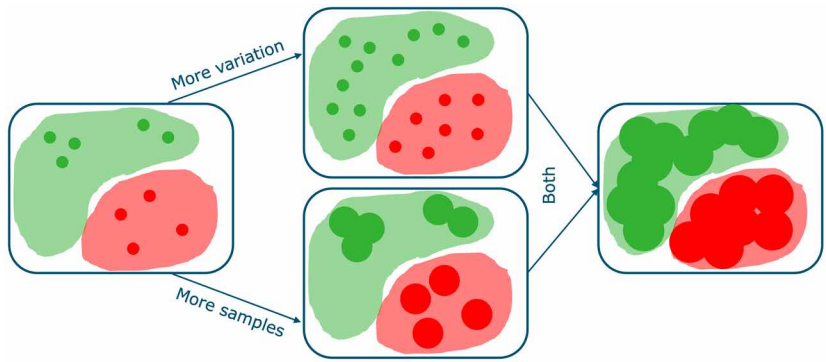

Figure 1 An illustration of the two ways to improve the quality of the training set. The shaded green and red areas symbolize the total distribution of all possible images of plants A and B. The darker areas illustrate the distribution of the training data. The training data can better cover the total distribution by adding more samples of the same fields, but more importantly, by adding more variation by adding data from new fields.

The performance was further improved by including more training images. This shows the importance to have sufficient variation and samples in the training set in order to minimize the generalization gap.

4.2.3 Incremental learning

However, even when using all training data from 20 different fields, a significant generalization gap was shown for some of the test fields that still contained something unique not found in the training fields (Ruigrok et al., 2023). This generalization gap could effectively be mitigated using incremental learning, where the neural network trained on the large and varied data set was fine-tuning on only a small number of training images from the specific field. The same was observed in Blok et al. (2020), where only a handful of images were needed to successfully fine-tune an instance-segmentation network to detect broccoli heads of a new cultivar.

4.2.4 Active learning

To obtain a high-quality dataset, it is important to add the right samples to the training data, that is, to add training samples that better cover the variation in possible future encounters. Randomly adding new training samples runs the risk of adding redundant samples containing information that was already covered in the training set. Instead, a different strategy is needed to select the most relevant new training samples, allowing the perception methods to achieve greater accuracy with fewer labeled training samples. This is important, especially in the field of agricultural robotics, where there is plenty of unlabeled data, but where it is difficult and time-consuming to get high-quality annotated data.

Active learning refers to a set of methods that allow machine learning methods themselves to choose the training data from which to learn (Settles, 2012). The general concept is illustrated in Fig. 2. Based on an initial training set, an active-learning model is trained. This model is then applied to all the data in a large pool of unlabeled images. For each unlabeled image, the method provides a prediction-uncertainty value. These are then used to select a number of new images that the method is most uncertain about, which are then labeled by a human (the oracle) to add as new data to the training set. Doing this in an iterative manner, allows the active learner to establish a high-quality training set.

The key in active learning is to estimate the prediction uncertainty of the model, also called the epistemic uncertainty (Kendall and Gal, 2017). In Bayesian neural networks, the epistemic uncertainty is modeled by including a distribution over the weights of the network, basically capturing the set

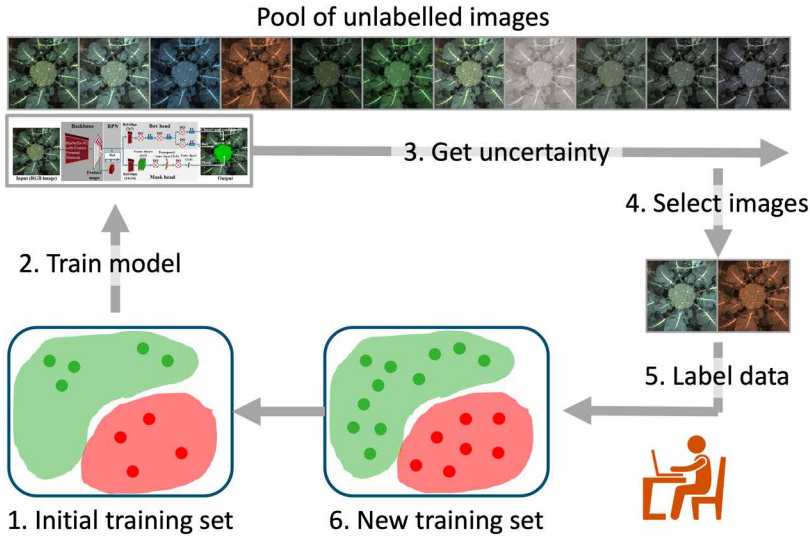

Figure 2 An illustration of the active-learning cycle for a classification task (green/red). The initial training set covers only part of the total distribution that can occur in reality (shaded green and red parts). Starting from this initial training set, the perception model is trained. This model is then applied to all available unlabeled images or during runtime. The images that the model is most uncertain about are then selected for human annotation to create an improved new training set. This process is continuously repeated.

of all plausible models given the training data. With more training data, the estimation of the model weights typically becomes more certain. However, to keep things tractable, usually an approximation of this Bayesian formulation is used by applying Monte-Carlo dropout (Gal and Ghahramani, 2016). In this procedure, multiple stochastic forward passes are applied to the same input. In each inference run, random dropout is applied, which basically disables part of the network connections at random. Prediction uncertainty can then be obtained by measuring the variation in the predictions made in the multiple stochastic forward passes, for instance, using the variance of the predictions for regression or using entropy for classification. Building upon Morrison et al. (2019), in Blok et al. (2022) Monte-Carlo dropout was used to estimate model uncertainty for an instance segmentation task of the detection of broccoli heads and the classification of different defects. Here, dropout was applied solely in the different heads of the Mask R-CNN network, not in the feature extraction part, speeding up computation time. The overall epistemic uncertainty consisted of semantic uncertainty (class), spatial uncertainty (bounding-box and mask), and uncertainty in object presence.

Using the epistemic uncertainty for active learning resulted in significantly better performance with the same amount of training data compared to the random selection of training data (Blok et al., 2022), indicating a better selection

of training data using epistemic uncertainty. This was further supported by the fact that the method selected more images containing underrepresented classes, resulting in a more balanced training set. A similar active-learning approach was taken in Chandra et al. (2020) to improve the performance of a deep neural network for panicle detection in cereal crops. Older work applied active learning to classical machine learning methods to classify crops and weeds using spectral imaging data (Pantazi et al., 2016). However, it must be noted that active learning is not always successful. In Rawat et al. (2022), active learning outperformed random sampling in only one of three semantic segmentation tasks for fruit detection. This might be caused by the composition of the datasets and the specific uncertainty metric used. More research is needed to ensure effective active learning in all cases.

Active learning is a human-in-the-loop machine learning approach (Monarch, 2021). By combining the expertise of human annotators with the ability of the machine to process large amounts of data, this hybrid approach allows the efficient selection of training data. Moreover, it allows live-long or continuous learning, where the agricultural robot asks for human assistance whenever it is uncertain about its own perception, becoming more and more capable to deal with the variation in agricultural environments.

4.2.5 Unsupervised and self-supervised

Where labeled data is costly to acquire, the costs of getting unlabeled data are typically quite low, for instance, by using UAVs to acquire images of fields. Unsupervised learning methods can exploit large unlabeled datasets by finding patterns in the data in order to learn good representations of the data. In dos Santos Ferreira et al. (2019), for instance, unsupervised deep clustering methods were used on two large image datasets with weed plants, which allowed the unsupervised training of visual features that were representative of the data. This was then combined in a semi-supervised approach, where human annotators annotated the clusters, greatly reducing the amount of human annotations needed, while achieving high classification results.

Another use of unsupervised methods is dimensionality reduction. Auto encoders are a popular way to reduce the dimensionality of image data and to capture the most important patterns in the data. This was applied to disease detection on peach leaves in Bedi and Gole (2021). They used a convolutional auto-encoder to learn a latent/embedded representation of the images, which was then used as input for a simple CNN to classify images as healthy or diseased, combining the power of unsupervised and supervised learning.

Self-supervised learning is another brand of unsupervised methods that gaining a lot of momentum and was promoted by LeCun and Misra (2022). Like supervised learning, these methods also learn from input–output pairs, but

different from supervised learning, these outputs are not provided by human annotation, but collected by the system itself from the unlabeled data. This can be framed as predictive learning, for instance, by learning to predict one part of the image based on another part (image completion), or by predicting the future based on the past using temporal data.

Contrastive learning is a self-supervised learning approach where a network is trained to learn an embedding that minimizes the distance between images from the same domain (positive samples) while maximizing the distance between images from different domains (negative samples). These positive and negative samples can be taken, for instance, from datasets of different crops, or a positive pair can be made with an image and a heavily distorted/augmented version of that image, while the negative pair are two different images. This way, contrastive learning results in a meaningful feature representation of the images capturing essential patterns in the unlabeled data. This principle was used in the agricultural domain, for instance, in Güldenring and Nalpantidis (2021), such an approach was applied for unsupervised pre-training of a CNN, which was then fine-tuned for a downstream task (plant classification) on a small amount of labeled data. This was shown to outperform conventional methods that use pre-training based on ImageNet, especially when limited amounts of labeled data are available.

4.2.6 Synthetic data and generative learning

The lack of training data to get robust perception models that generalize well can also be solved by enriching the training set using synthetic data. In Turgut et al. (2022), 3D synthetic rosebush models were successfully used to pre-train 3D point-based deep neural networks for plant-part segmentation. Using realistic 3D structural plant models, Barth et al. (2018) and Barth et al. (2019) rendered a large set of synthetic images to bootstrap training of a CNN for the segmentation of plant parts for a bell-pepper harvesting robot. They further increased the realism of these images using a generative adversarial network (GAN), more specifically CycleGAN (Barth et al., 2020). A conditional GAN was used in (Abbas et al., 2021) to generate additional synthetic images of tomato leaves, boosting the performance of a CNN for disease classification. A similar approach was taken in Arsenovic et al. (2019) using StyleGAN to generate additional synthetic images. Although several successes of the use of GANs to generate additional training data are reported in the literature, van Marrewijk et al. (2022) noted that in some studies, the evaluation was not sound as the GANs themselves were trained on data outside of the original training set, giving an unfair advantage over training on only the original set. They also concluded that the advantage of using synthetic data from a GAN is limited and time spent on setting up the GAN and optimizing the hyperparameters might be better spent on labeling some additional training data. Hence, the

effectiveness of using generative learning to improve the generalization of neural networks needs to be further investigated.

5 Dealing with the challenge of incomplete information

Section 5.1 discusses some passive approaches to deal with incomplete information. However, as discussed in Section 5.2, multi-view and active-vision approaches have been proposed that allow agricultural robots to actively gather more information about the environment to minimize the information gap.

5.1 Passive visual perception

Despite the fact that robots have the ability to actively gather new information, most of the work on machine vision for agricultural applications have focused on passive perception, that is, trying to get as much information out of a single image. Whenever the objects of interest are well visible in the image, that approach is perfectly valid. However, when the objects are not well visible in the image, for instance, due to sensor noise, flares, or because the objects are partially or completely occluded, a passive approach is inadequate.

To some extent, deep neural networks can deal with partial occlusions using a single image by developing and training them to be robust to occlusion and even to predict the occluded parts. In Li et al. (2022), for instance, a method was developed for the localization of partially occluded apples. Blok et al. (2021) trained a neural network to predict the complete 2D mask of partially occluded broccoli heads, improving the size estimation. Similarly, the full 3D point cloud, including the non-observable parts, could be predicted from monocular images (Zeng et al., 2018). Others used a conditional GAN to learn to generate non-occluded versions of images with occluded grape bunches (Kierdorf et al., 2022).

Although these methods gain some improvement in performance, they can only predict the occluded parts based on patterns observed in the training data. If there is sufficient regularity or symmetry in the shape of the objects, this can work well. However, many agricultural objects are quite irregular, making it much more challenging to accurately predict the occluded parts, which is often needed for robotic operation. Furthermore, in agricultural production, anomalies and deviations from the norm are often relevant features, which cannot be predicted from partial information at all. Moreover, in cluttered environments, many objects are heavily or even fully occluded. In a bell-pepper greenhouse, for instance, in 31% of the cases, less than half of

the fruit surface was visible in the images (Hemming et al., 2014), many were even completely hidden from view. To properly deal with such situations, the information gap needs to be minimized by allowing robots to gather more information.

5.2 Minimizing the information gap

This subsection discusses a few approaches to minimize the information gap.

5.2.1 Multi-view visual perception

Hemming et al. (2014) showed that in a cluttered environment, the detection of bell peppers was significantly improved when combining multiple viewpoints with an eye-in-hand system; going from 69% for a single viewpoint to 90% for five viewpoints. In Sa et al. (2017), a fixed scanning path was applied to acquire a more complete 3D point-cloud reconstruction of a bell pepper from multiple viewpoints to be used to guide the harvesting operation. A similar approach was taken in Barth et al. (2016), where apart from a 3D reconstruction, the scan was used to find the best view on the to-be-harvested bell pepper.

In (Halstead et al., 2018), a vision system was developed to estimate the yield in a bell-pepper greenhouse. They developed a multi-view approach with the robot taking images while traversing the crop row. To prevent double counting and to deal with fruits that are (partially) occluded in some frames, a multi-object tracker was implemented, based on the overlap in object detections in consecutive frames. The tracking method was improved in Smitt et al. (2021) using the reprojection of the object mask in the current frame to the expected location in the next frame using depth data and wheel odometer. This method was further refined and applied in two different use-cases in Halstead et al. (2021). Kurtser and Edan (2018) analyzed the detectability of bell pepper depending on viewing distance and angle, which can be used to select more profitable sets of viewpoints to increase detectability. Not only can multiple views increase the chance of having the objects of interest in view, they can also be used to improve the detection. Observing plants from ten different camera viewpoints, for instance, did not only result in more complete 3D reconstructions but it also improved the plant-part segmentation (Shi et al., 2019). Van Essen et al. (2022) showed that fruit maturity classification could be improved by dynamically selecting additional viewpoints. In van Essen et al. (2021), the species classification of fish on a conveyor belt was improved by tracking the fish over multiple frames and combining the predictions of a neural network probabilistically. Russello et al. (2021) showed that the pose estimation

of partially occluded walking cows could be improved with a temporal network including multiple video frames.

It is clear that multi-view perception has many advantages over passive perception. Still, the viewpoints in the above-mentioned studies are set in advance by the developer and not tuned to the current situation, which does not guaranty that the required information is acquired. To further improve perception, agricultural robots would benefit from actively viewing the world.

5.2.2 Active visual perception

Using a matrix of nine cameras connected to the end effector of a robot, Lehnert et al. (2019) developed a visual servoing method that determines the next viewpoint to get a less occluded view of a bell pepper. This system was used to train a deep neural network to predict the next viewpoint from just one camera image (Zapotezny-Anderson and Lehnert, 2019).

The downside of visual-servoing approaches is that they require a non-occluded starting viewpoint and can get stuck in local optima. Next-best-view (NBV) planners based on occupancy grids can deal with this, as they are based on a calculation of information gain that guides the robot to explore unseen parts of space. As illustrated in Fig. 3, based on a current reconstruction of the scene, the robot evaluates a few candidate viewpoints to then choose the viewpoint with the highest expected information gain. In Zaenker et al. (2021), this was, for instance, used to detect fruits on a number of plants in a small simulated greenhouse. It showed beneficial to combine the global NBV

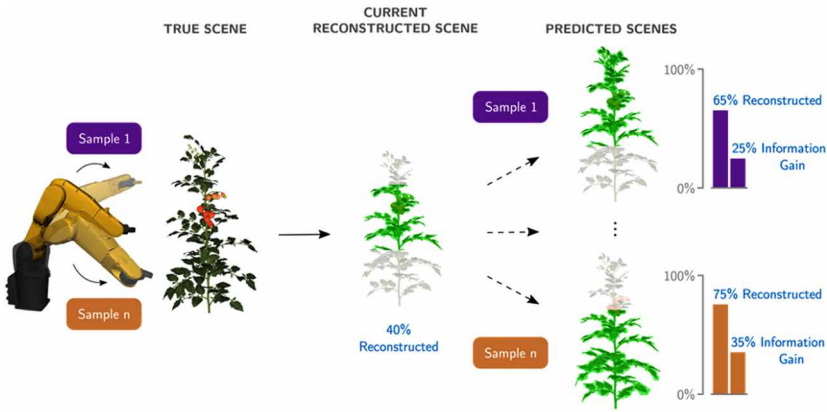

Figure 3 An illustration of the active-perception next-best-view paradigm. The robot has a current reconstruction of the world (a tomato plant in this case), which represents the seen and unseen parts of space. Based on this, a prediction can be made of the expected information gain from different new viewpoints, in order to select the best next view.

planning with local view planning based on visual servoing (Zaenker et al., 2021). Burusa et al. (under review) added an attention mechanism to efficiently reconstruct specific plant parts and showed that it outperformed random and fixed multi-view approaches.

The NBV planning could be further optimized and accelerated using deep reinforcement learning, providing additional coverage of unexplored space and new information about the relevant parts as a reward to train the system (Zeng et al., 2022). Alternatively, in Han et al. (2022), the information gain for new viewpoints was learned by a network in a supervised manner, acquiring the training data by comparing the reconstruction to a known 3D point cloud of the object.

5.2.3 Sensor fusion

Instead of redirecting the same sensor, the challenge of incomplete information can also be approached by combining different complementary sensor modalities. The detection of immature citrus fruit, for instance, was boosted by combining color and thermal images (Gan et al., 2018). Semantic segmentation of apple trees (Kang and Wang, 2023) and cucumber plants (Boogaard et al., 2021) was improved with a fusion of 3D and color data. For the automatic identification of plant diseases using UAVs, the fusion of thermal and spectral imaging was often successful (Neupane and Baysal-Gurel, 2021).

6 Directions for future research

Sections 4 and 5 discuss approaches to deal with the challenges of variation and incomplete information. However, the challenges are far from being resolved and future research needs to further progress along these lines. There are some suggestions for future research on the short term to better deal with the challenges, and some suggestions for future research on the long term to get to a higher level of autonomy in agricultural robotics.

6.1 On the short term: dealing better with variation and incomplete information

From this review, it is clear that the perception for agricultural robotics has greatly improved in the past decade. Building upon the progress in the field of deep learning, the performance of machine vision algorithms improved significantly and methods became more robust to variations present in the agricultural environment. However, as discussed in Section 3, the challenge of variation is not yet solved. Similarly, progress in the general field of robotics provided methods for multi-view and active perception, yet, the challenge of

incomplete information remains for robots to effectively deal with cluttered agricultural environments. In this subsection, a few specific directions for future research are being discussed.

6.1.1 Quality of training set

It was shown in Ruigrok et al. (2023) that the generalization of deep neural networks to new situations greatly benefits from a rich and varied training set. This calls for better sharing of labeled datasets. Even research institutes are cautious with sharing their data, but especially companies consider data the new gold. Although this is possibly true, keeping the data private hampers progress in the field. Moreover, many of the small startups that currently enter the field are too small to collect rich-enough datasets to deal with all the variance present in agricultural environments. Alternatively, datasets can be enriched using image augmentation. Currently, mainly general augmentation methods are applied, which are suboptimal or sometimes even harmful in a specific domain (Balestriero et al., 2022). Instead, domain-specific augmentation can improve the generalization of the perception methods (e.g. Su et al., 2021; Ratner et al., 2017).

6.1.2 Active learning and uncertainty

Perception models need to become better aware of the uncertainty in the prediction. This is important to improve the safety and reliability of agricultural robots. When uncertain, actions can be postponed, new observations can be acquired or human assistance can be called to prevent errors. The latter might sound like a weak option, and fully autonomous systems are indeed the aim, but for the coming decade, robots need the assistance of humans to operate over long periods of time and in many different environments. More research on human-in-the-loop systems therefore needs to be done. Not only on the topic of active learning but also on effective interfaces to minimize the human effort to assist the robot (Monarch, 2021).

6.1.3 Unsupervised learning

In his cake analogy, Yann LeCun views machine intelligence as a cake, with reinforcement learning as the cherry on top, supervised learning as the icing, and unsupervised learning as the bulk of the cake (LeCun, 2016). In other words, unsupervised learning is essential to create intelligent systems. It allows to exploit large sets of unlabeled data, typically available also in agriculture, to learn powerful low-dimensional representations of the data, which allows efficient supervised learning of downstream tasks based on small labeled datasets. To unlock the power of unsupervised learning, more research is

needed in the development and application of unsupervised techniques like predictive learning, contrastive learning, deep clustering, domain adaptation, and self-supervised learning.

6.1.4 Use of domain knowledge

The completely data-driven deep-learning approaches ignore the available domain knowledge to solve agricultural tasks. Combining data-based and model-based methods would allow to put constraints on the learning process to achieve better performance especially when training data are limited. This can be done, for instance, by adding domain knowledge to create synthetic data based on models, for instance, using plant models (Turgut et al., 2022; Barth et al., 2019), to include specific image features as additional input (Liu et al., 2021; Milioto et al., 2018), or to use domain knowledge in the loss function to enforce consistency of the network predictions, such as geometrical and temporal constraints for pose detection (Dabral et al., 2018). Domain knowledge can also be used to deal with missing information, for instance, by using leaf templates to reconstruct plant leaves in occluded scenarios (Marks et al., 2022).

6.1.5 Active perception

Robots have the ability to control their own sensory input. This fact is hardly explored in current agricultural-robotic research, which instead focuses predominantly on the processing of single images. As discussed in Section 5, multi-view and active NBV planning improves the performance of perception methods, especially in occluded environments. Further research is needed to make these methods even more effective, for instance using attention mechanisms (Burusa et al., under review), adding prior knowledge for effective object search, or allowing robots to learn optimal view planning from their own experience (Zeng et al., 2022; Han et al., 2022). The robot's active capabilities can also be used in an active learning setting, where the robot actively plans a path to collect useful training data to be labeled by a human (Rückin et al., 2022). The next step is to develop interactive object learning, where a robot actively explores objects in the environment to learn and improve its perception models without or with minimal human intervention (Lyubova et al., 2016).

6.2 On the long term: autonomous agricultural robotics

With the research directions described earlier, in the next decade, the field can develop perception methods for agricultural robots to robustly perform

specific tasks, such as harvesting fruits or removing weeds. These robots will probably work semi-autonomously, needing occasional input from a human operator. On the long term, however, robots are needed that can perform multiple tasks and operate fully autonomously. This requires the next level of artificial intelligence, going beyond the understanding of a single domain to mastering multiple domains. Some first steps in generalizing over different tasks have recently been taken with Gato, a generalist agent (Reed et al., 2022). However, it also requires a true understanding of the world, going beyond patterns extracted from training data to grounding the external world in the robot's own sensorimotor system (Harnad, 1990). This research challenge is enormous, and I definitely do not want to claim to know the solution. However, there are a few principles that I believe are important to reach this goal.

Most importantly, perception and action need to be tighter integrated to achieve sensorimotor coordination. When a robot takes an action, it immediately results in a change of the sensory observations. When the robot changes its position, it gets a new view of the world, and when it interacts with objects, it changes the state of the objects. This applies not only to the sense of vision but to all senses. When a robot interacts with an object, for instance, it receives visual feedback about its actions, but also tactile and auditory. For a robot to truly understand the world, it needs to be able to predict the sensory consequences of its actions, the so-called sensorimotor contingencies (O'Regan and Noë, 2001). This requires that the robot has a model of the world containing knowledge about the state and how this will change as a consequence of an action, including object affordances containing the knowledge of the actions that can be applied to objects and the resulting effects of these actions (Gibson, 2014; Krüger et al., 2011). This world model needs to be continuously updated to be kept in sync with the physical world based on new sensory observations and experiences of the robot.

The predominant sense–think–act cycle in robotics then needs to change to an act-predict-sense-compare cycle, where the predictions of the consequences of actions are compared to the new observations after applying the action. When the predictions are correct, the world model is correct and the next planned action can be executed. When the predictions do not match the observations, the world changes differently than expected, requiring attention to update the model. This is in line with current theories on human cognition (see, e.g. Clark, 2015).

The consequences of actions applied to an object can be learned by the robot autonomously by interacting with that object, thereby grounding the object in its sensory-motor experience (Harnad, 1990; Pfeifer and Bongard, 2006). This process of playful manipulation of the world allows the robot to learn causality (Pearl, 2019). Furthermore, during execution, the robot continuously observes the sensory consequences of actions, allowing self-supervised and

Published by Burleigh Dodds Science Publishing Limited, 2024.

life-long online learning of the prediction models. Importantly, the robot should not only be equipped with visual sensors but also with other sensory modalities.

Inspiration for robot learning can furthermore be taken from developmental psychology on how humans learn to perceive and interact with the world (Cangelosi and Schlesinger, 2015; Lungarella et al., 2003). An important aspect is that learning is incremental, starting from learning basic skills, like moving your hand to an object in space and grasping that object, which allows learning about objects in the world and on which more complex skills can be built, like harvesting all ripe apples in an orchard. Initially, learning is more constraint due to limitations in the sensorimotor system and through scaffolding by caretakers, simplifying the learning process, which gradually becomes less constraint to learn more complex relations. Another aspect is that of curiosity. When a certain skill is learned and predictions of the sensory consequence can be made successfully, it is time to explore the world for new challenges.

With this sensorimotor learning, robots can start to understand the world from their own perspective in an unsupervised way. This builds the foundation for further learning. To connect to the human world and perform a variety of agricultural tasks, future robots will no longer be programmed, but learn from human demonstrations (Ravichandar et al., 2020). Similar to how infants learn, imitation learning will provide a common taxonomy of the world and bootstrap learning of tasks. With this in place, robots can learn to fine-tune the execution of tasks with their own body morphology using reinforcement learning (Kober et al., 2013; Arulkumaran et al., 2017). These three levels of learning (unsupervised, supervised, and reinforced) correspond to the three ingredients of LeCun's cake analogy.

7 Conclusion

To conclude this survey of the advances of visual perception for agricultural robotics, it can be observed that current robots can operate in environments that are relatively structured and controlled. The challenges of variation and incomplete information, however, hamper the application of agricultural robots on a larger scale. If the amount of variation increases, for instance, by going to a variety of fields, the generalization of current perception methods is not yet good enough to be commercially applied. At the same time, when the amount of occlusion increases in cluttered environments, such as high-wire greenhouses and orchards, current robots are not able to locate all relevant objects. To improve future agricultural robots, a number of promising research directions to deal with the challenges of variation and incomplete information have been provided. This will allow the field to develop robots that can successfully perform specific tasks in agricultural environments in the coming

decade. To achieve more versatile robots that can perform any task on a farm, steps need to be taken to get closer to artificial general intelligence in the coming decades. As an important principle, robot perception should no longer be studied in isolation, but in tight integration with the robot's actions within the framework of sensorimotor systems. Furthermore, for effective learning, robots should be able to explore the world and learn from their own experiences.

8 References

Abbas, I., Liu, J., Faheem, M., Noor, R. S., Shaikh, S. A., Solangi, K. A. and Raza, S. M. Different sensor based intelligent spraying systems in agriculture, *Sensors and Actuators. Part A* 316, 112265 2020, issn: 0924-4247. doi: 10.1016/j.sna.2020.112265. [Online]. Available at: https://www.sciencedirect.com/science/article/ pii/S0924424720309274.

Abbas, A., Jain, S., Gour, M. and Vankudothu, S. Tomato plant disease detection using transfer learning with c-gan synthetic images, *Computers and Electronics in Agriculture* 187, 106279 2021.

Arad, B., Kurtser, P., Barnea, E., Harel, B., Edan, Y. and Ben-Shahar, O. Controlled lighting and illumination-independent target detection for real-time cost-efficient applications. The case study of sweet pepper robotic harvesting, *Sensors* 19(6) 2019, issn: 1424-8220. doi: 10.3390/s19061390. [Online]. Available at: https://www.mdpi.com/1424-8220/19/6/1390.

Arad, B., Balendonck, J., Barth, R., Ben-Shahar, O., Edan, Y., Hellström, T., Hemming, J., Kurtser, P., Ringdahl, O., Tielen, T. and van Tuijl, B. Development of a sweet pepper harvesting robot, *Journal of Field Robotics* 37(6), 1027–1039 2020. doi: 10.1002/rob.21937.

Arsenovic, M., Karanovic, M., Sladojevic, S., Anderla, A. and Stefanovic, D. Solving current limitations of deep learning based approaches for plant disease detection, *Symmetry* 11(7), 939 2019.

Arulkumaran, K., Deisenroth, M. P., Brundage, M. and Bharath, A. A. Deep reinforcement learning: A brief survey, *IEEE Signal Processing Magazine* 34(6), 26–38 2017.

Balestriero, R., Bottou, L. and LeCun, Y. *"The Effects of Regularization and Data Augmentation Are Class Dependent," arXiv Preprint ArXiv:2204.03632* 2022.

Bargoti, S. and Underwood, J. P. Image segmentation for fruit detection and yield estimation in apple orchards, *Journal of Field Robotics* 34(6), 1039–1060 2017. doi: 10.1002/rob.21699. eprint. [Online]. Available at: https://onlinelibrary.wiley.com/doi/abs/10.1002/rob.21699.

Barth, R., Hemming, J. and van Henten, E. J. Design of an eye-in-hand sensing and servo control framework for harvesting robotics in dense vegetation, *Biosystems Engineering* 146, 71–84 2016. doi: 10.1016/j.biosystemseng.2015.12.001.

Barth, R., IJsselmuiden, J., Hemming, J. and Van Henten, E. J. Data synthesis methods for semantic segmentation in agriculture: A capsicum annuum dataset, *Computers and Electronics in Agriculture* 144, 284–296 2018. doi: 10.1016/j.compag. 2017.12.001. [Online]. Available at: https://www.sciencedirect.com/science/article/pii/S0168169917305689.

Barth, R., IJsselmuiden, J., Hemming, J. and Van Henten, E. J. Synthetic bootstrapping of convolutional neural networks for semantic plant part segmentation, *Computers*

and Electronics in Agriculture. Hemming 161, 291–304 2019. doi: 10.1016/j.compag.2017.11.040. [Online]. Available at: http://www.sciencedirect.com/science/article/pii/S0168169917307664.

Barth, R., Hemming, J. and Van Henten, E. J. Optimising realism of synthetic images using cycle generative adversarial networks for improved part segmentation, *Computers and Electronics in Agriculture* 173, 105378 2020, issn: 0168-1699. doi: 10.1016/j.compag.2020.105378. [Online]. Available at: https://www.sciencedirect.com/science/article/pii/ S0168169919320794.

Bedi, P. and Gole, P. Plant disease detection using hybrid model based on convolutional autoencoder and convolutional neural network, *Artificial Intelligence in Agriculture* 5, 90–101 2021.

Billingsley, J., Visala, A. and Dunn, M. Robotics in agriculture and forestry. In: *Springer Handbook of Robotics* 2008. Springer, Berlin, Heidelberg Siciliano, B. and Khatib, O. (Eds). doi: 10.1007/978-3-540-30301-5_47.

Blok, P. M., van Evert, F. K., Tielen, A. P. M., van Henten, E. J. and Kootstra, G. The effect of data augmentation and network simplification on the image-based detection of broccoli heads with Mask R-CNN, *Journal of Field Robotics*, 1–20 2020. doi: 10.1002/rob.21975. eprint. [Online]. Available at: https://onlinelibrary.wiley.com/doi/abs/10.1002/rob.21975.

Blok, P. M., van Henten, E. J., van Evert, F. K. and Kootstra, G. Image-based size estimation of broccoli heads under varying degrees of occlusion, *Biosystems Engineering* 208, 213–233 2021.

Blok, P. M., Kootstra, G., Elghor, H. E., Diallo, B., van Evert, F. K. and van Henten, E. J. Active learning with maskal reduces annotation effort for training mask r-cnn on a broccoli dataset with visually similar classes, *Computers and Electronics in Agriculture* 197, 106917 2022, issn: 0168-1699. doi: 10.1016/j.compag.2022.106917. [Online]. Available at: https: //www.sciencedirect.com/science/article/pii/S0168169922002344.

Blok, P. M., van Boheemen, K. and van Evert, F. K., iJsselmuiden, J., and Kim, G.-H., Robot navigation in orchards with localization based on particle filter andabelln filter, *Computers and Electronics in Agriculture* 157, 261–269 n.d. doi: 10.1016/j.compag.2018.12.046. [Online]. Available at: https://www.sciencedirect.com/science/article/ pii/S0168169918315230.

Bonadies, S. and Gadsden, S. A. An overview of autonomous crop row navigation strategies for unmanned ground vehicles, *Engineering in Agriculture, Environment and Food* 12(1), 24–31 2019, issn: 1881-8366. doi: 10.1016/j.eaef.2018.09.001. [Online]. Available at: https://www.sciencedirect.com/science/article/pii/S188183661730188X.

Boogaard, F. P., Rongen, K. S. A. H. and Kootstra, G. W. Robust node detection and tracking in fruit-vegetable crops using deep learning and multi-view imaging, *Biosystems Engineering* 192, 117–132 2020, issn: 1537-5110. doi: 10.1016/j.biosystemseng.2020.01.023. [Online].

Boogaard, F. P., van Henten, E. J. and Kootstra, G. Boosting plant-part segmentation of cucumber plants by enriching incomplete 3D point clouds with spectral data, *Biosystems Engineering* 211, 167–182 2021. doi: 10.1016/j.biosystemseng.2021.09.004.

Boogaard, F. P., van Henten, E. J. and Kootstra, G. Improved point-cloud segmentation for plant phenotyping through class-dependent sampling of training data to battle class imbalance, *Frontiers in Plant Science* 13, 838190–838190 2022.

Bosch, A., Zisserman, A. and Munoz, X. Representing shape with a spatial pyramid kernel. In: Proceedings of the 6th ACM International Conference on Image and Video Retrieval 2007, pp. 401–408.

Bulanon, D. M., Burks, T. F. and Alchanatis, V. Fruit visibility analysis for robotic citrus harvesting, *Transactions of the ASABE* 52(1), 277–283 2009.

Bulanon, D. M., Hestand, T., Nogales, C., Allen, B. and Colwell, J. Machine vision system for orchard management. In: *Machine Vision and Navigation* Sergiyenko, O., Flores-Fuentes, W. and Mercorelli, P. (Eds). Springer, 2020.

Burusa, A., van Henten, E. and Kootstra, G. *"Attention-driven active vision for efficient reconstruction of plants and targeted plant parts," arXiv Preprint ArXiv:2206.10274* 2022. doi: 10.48550/arXiv.2206.10274.

Cangelosi, A. and Schlesinger, M. *Developmental Robotics: From Babies to Robots*. MIT Press 2015.

Caron, M., Misra, I., Mairal, J., Goyal, P., Bojanowski, P. and Joulin, A. Unsupervised learning of visual features by contrasting cluster assignments, *Advances in Neural Information Processing Systems* 33, 9912–9924 2020.

Chandra, A. L., Desai, S. V., Balasubramanian, V. N., Ninomiya, S. and Guo, W. Active learning with point supervision for cost-effective panicle detection in cereal crops, *Plant Methods* 16(1), 34 2020.

Chaudhury, A., Ward, C., Talasaz, A., Ivanov, A. G., Brophy, M., Grodzinski, B., Huner, N. P. A., Patel, R. V. and Barron, J. L. Machine vision system for 3d plant phenotyping, *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 16(6), 2009–2022 2019. doi: 10.1109/TCBB.2018.2824814.

Chen, J., Chen, J., Zhang, D., Sun, Y. and Nanehkaran, Y. A. Using deep transfer learning for imagebased plant disease identification, *Computers and Electronics in Agriculture* 173, 105393 2020.

Chlingaryan, A., Sukkarieh, S. and Whelan, B. Machine learning approaches for crop yield prediction and nitrogen status estimation in precision agriculture: a review, *Computers and Electronics in Agriculture* 151, 61–69 2018, issn: 0168-1699. doi: 10.1016/j.compag.2018.05.012. [Online]. Available at: https://www.sciencedirect.com/science/article/pii/ S0168169917314710.

Clark, A. *Surfing Uncertainty: Prediction, Action, and the Embodied Mind*. Oxford University Press 2015.

Corke, P. *Robotics, Vision and Control, Ser., Springer Tracts in Advanced Robotics*. Springer, Cham 2017.

Dabral, R., Mundhada, A., Kusupati, U., Afaque, S., Sharma, A. and Jain, A. Learning 3d human pose from structure and motion. In: Proceedings of the European Conference on Computer Vision (ECCV) 2018, pp. 679–696.

Doornweerd, J. E., Kootstra, G., Veerkamp, R. F., Ellen, E. D., van der Eijk, J. A. J., van de Straat, T. and Bouwman, A. C. Across-species pose estimation in poultry based on images using deep learning, *Frontiers in Animal Science* 2 2021. doi: 10.3389/fanim.2021.791290. [Online].

dos Santos Ferreira, A., Freitas, D. M., da Silva, G. G., Pistori, H. and Folhes, M. T. Unsupervised deep learning and semi-automatic dataabellingg in weed discrimination, *Computers and Electronics in Agriculture* 165, 104963 2019.

Dutagaci, H., Rasti, P., Galopin, G. and Rousseau, D. Rose-x: an annotated data set for evaluation of 3d plant organ segmentation methods, *Plant Methods* 16(1), 28 2020.

Fu, L., Sun, S., Li, R. and Wang, S. Classification of kiwifruit grades based on fruit shape using a single camera, *Sensors* 16(7) 2016, issn: 1424-8220. doi: 10.3390/s16071012. [Online]. Available at: https://www.mdpi.com/1424-8220/16/7/1012.

Fu, L., Gao, F., Wu, J., Li, R., Karkee, M. and Zhang, Q. Application of consumer rgb-d cameras for fruit detection and localization in field: A critical review, *Computers and Electronics in Agriculture* 177, 105687 2020, issn: 0168-1699. doi: 10.1016/j.compag.2020.105687. [Online]. Available at: https://www.sciencedirect.com/science/article/pii/ S0168169920319530.

Gal, Y. and Ghahramani, Z. Dropout as aabellinn approximation: representing model uncertainty in deep learning. In: International Conference *on* Machine Learning, PMLR, 2016 pp. 1050–1059.

Gan, H., Lee, W. S., Alchanatis, V., Ehsani, R. and Schueller, J. K. Immature green citrus fruit detection using color and thermal images, *Computers and Electronics in Agriculture* 152, 117–125 2018.

Gibson, J. J. *The Ecological Approach to Visual Perception: Classic Edition*. Psychology Press 2014.

Golbach, F., Kootstra, G., Damjanovic, S., Otten, G. and van de Zedde, R. Validation of plant part measurements using a 3d reconstruction method suitable for high-throughput seedling phenotyping, *Machine Vision and Applications* 27(5), 663–680 2016.

Goodfellow, I., Bengio, Y. and Courville, A. *Deep Learning*. MIT Press 2016. Available at: http://www.deeplearningbook.org.

Güldenring, R. and Nalpantidis, L. Self-supervised contrastive learning on agricultural images, *Computers and Electronics in Agriculture* 191, 106510 2021.

Guo, N., Zhang, B., Zhou, J., Zhan, K. and Lai, S. Pose estimation and adaptable grasp configuration with point cloud registration and geometry understanding for fruit grasp planning, *Computers and Electronics in Agriculture* 179, 105818 2020, issn: 0168-1699. doi: 10.1016/j.compag.2020.105818. [Online]. Available at: https://www.sciencedirect.com/science/ article/pii/S0168169920314046.

Halstead, M., McCool, C., Denman, S., Perez, T. and Fookes, C. Fruit quantity and ripeness estimation using a robotic vision system, *IEEE Robotics and Automation Letters* 3(4), 2995–3002 2018. doi: 10.1109/LRA.2018.2849514.

Halstead, M., Ahmadi, A., Smitt, C., Schmittmann, O. and McCool, C. Crop agnostic monitoring driven by deep learning, *Frontiers in Plant Science* 12, 786702 2021, issn: 1664-462X. doi: 10.3389/fpls.2021.786702. [Online]. Available at: https://www.frontiersin.org/articles/10.3389/fpls.

Han, Y., Zhan, I. H., Zhao, W. and Liu, Y.-J. A double branch next-best-view network and novel robot system for active object reconstruction. In: IEEE Publications International Conference on Robotics and Automation (ICRA) (vol. 2022), 2022, pp. 7306–7312.

Harnad, S. The symbol grounding problem, *Physica D: Nonlinear Phenomena* 42(1–3), 335–346 1990.

Hemming, J., Ruizendaal, J., Hofstee, J. W. and van Henten, E. J. Fruit detectability analysis for different camera positions in sweet-pepper, *Sensors* 14(4), 6032–6044 2014. doi: 10.3390/s140406032.

Hitelman, A., Edan, Y., Godo, A., Berenstein, R., Lepar, J. and Halachmi, I. Biometric identification of sheep via a machine-vision system, *Computers and Electronics in Agriculture* 194, 106713 2022.

Horaud, R., Hansard, M., Evangelidis, G. and Ménier, C. An overview of depth cameras and range scanners based on time-of-flight technologies, *Machine Vision and Applications* 27(7), 1005–1020 2016. doi: 10.1007/s00138-016-0784-4.

Jamil, N., Kootstra, G. and Kooistra, L. Evaluation of individual plant growth estimation in an intercropping field with UAV imagery, *Agriculture* 12(1), 10 2022, issn: 2077-0472. doi: 10.3390/agriculture12010102. [Online]. Available at: https://www.mdpi.com/2077-0472/12/1/102.

Kamilaris, A. and Prenafeta-Boldú, F. X. Deep learning in agriculture: A survey, *Computers and Electronics in Agriculture* 147, 70–90 2018, issn: 0168-1699. doi: 10.1016/j.compag.2018.02.016. [Online]. Available at: http://www.sciencedirect.com/science/ article/pii/S0168169917308803.

Kang, H. and Wang, X. Semantic segmentation of fruits on multi-sensor fused data in natural orchards, *Computers and Electronics in Agriculture* 204, 107569 2023.

Kang, H., Zhou, H., Wang, X. and Chen, C. Real-time fruit recognition and grasping estimation for robotic apple harvesting, *Sensors* 20(19) 2020, issn: 1424-8220. doi: 10.3390/s20195670. [Online]. Available at: https://www.mdpi.com/1424-8220/20/19/5670.

Kendall, A. and Gal, Y. What uncertainties do we need inabellinn deep learning for computer vision?, *Advances in Neural Information Processing Systems* 30 2017.

Kierdorf, J., Weber, I., Kicherer, A., Zabawa, L., Drees, L. and Roscher, R. Behind the leaves: estimation of occluded grapevine berries with conditional generative adversarial networks, *Frontiers in Artificial Intelligence* 5, 830026 2022.

Kober, J., Bagnell, J. A. and Peters, J. Reinforcement learning in robotics: A survey, *The International Journal of Robotics Research* 32(11), 1238–1274 2013.

Kootstra, G., Bender, A., Perez, T. and van Henten, E. J. 2020. Robotics in agriculture. In *Encyclopedia of Robotics*, pp. 1–19. Springer, isbn: 978-3-642-41610-1. doi: 10.1007/9783-642-41610-1_43-1. [Online]. doi: 10.1007/978-3-642-41610-1_43-1.

Kootstra, G., Wang, X., Blok, P. M., Hemming, J. and van Henten, E. Selective harvesting robotics: current research, trends, and future directions, *Current Robotics Reports* 2(1), 95–104 2021, issn: 2662-4087. doi: 10.1007/s43154-020-00034-1. [Online].

Krizhevsky, A., Sutskever, I. and Hinton, G. E. 2012. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems* Pereira, F., Burges, C. J. C., Bottou, L. and Weinberger, K. Q. (Eds) (vol. 25), pp. 1097–1105. Curran Associates, Inc. [Online]. Available at: https://proceedings.neurips.cc/paper/2012/file/c399862d3b9d6b76c8436e924a68c45bPaper.pdf.

Krüger, N., Geib, C., Piater, J., Petrick, R., Steedman, M., Wörgötter, F., Ude, A., Asfour, T., Kraft, D., Omrčen, D., Agostini, A. and Dillmann, R. Object–action complexes: grounded abstractions of sensory–motor processes, *Robotics and Autonomous Systems* 59(10), 740–757 2011.

Kurtser, P. and Edan, Y. Statistical models for fruit detectability: spatial and temporal analyses of sweet peppers, *Biosystems Engineering* 171, 272–289 2018.

Kurtser, P. and Edan, Y. Planning the sequence of tasks for harvesting robots, *Robotics and Autonomous Systems* 131, 103591 2020, issn: 0921-8890. doi: 10.1016/j.robot.2020.103591. [Online]. Available at: https://www.sciencedirect.com/science/article/ pii/S0921889020304310.

Lamping, C., Derks, M., Groot Koerkamp, P. and Kootstra, G. ChickenNet-- an end-to-end approach for plumage condition assessment of laying hens in commercial farms using computer vision, *Computers and Electronics in Agriculture* 194, 106695

2022, issn: 0168-1699. doi: 10.1016/j.compag.2022.106695. [Online]. Available at: https://www.sciencedirect. com/science/article/pii/S0168169922000126.

LeCun, Y. [Online], *"Predictive Learning, Keynote at NIPS2016."* 2016. Available at: https://youtu.be/Ount2Y4qxQo (visited on 24/07/2022).

LeCun, Y. and Misra, I. Self-supervised learning: the dark matter of intelligence, Facebook 2022, [Online] Available at: https://ai.facebook.com/blog/self-supervised-learning-the-dark-matter-ofintelligence (visited on 15/07/2022).

Lehnert, C., Tsai, D., Eriksson, A. and McCool, C. 3d move to see: multi-perspective visual servoing towards the next best view within unstructured and occluded environments 2019. In: IEEE Publications International, R. S. J. (Ed.) Conference on Intelligent Robots and Systems (IROS) (vol. 2019), pp. 3890–3897. doi: 10.1109/IROS40897.2019.8967918.

Li, T., Feng, Q., Qiu, Q., Xie, F. and Zhao, C. Occluded apple fruit detection and localization with a frustum-based point-cloud-processing approach for robotic harvesting, *Remote Sensing* 14(3) 2022, issn: 2072-4292. doi: 10.3390/rs14030482. [Online]. Available at: https://www.mdpi.com/20724292/14/3/482.

Liu, X., Yu, S. Y., Flierman, N. A., Loyola, S., Kamermans, M., Hoogland, T. M. and De Zeeuw, C. I. Optiflex: multi-frame animal pose estimation combining deep learning with optical flow, *Frontiers in Cellular Neuroscience* 15, 621252 2021.

Lowe, D. G. Distinctive image features from scale-invariant keypoints, *International Journal of Computer Vision* 60(2), 91–110 2004.

Lungarella, M., Metta, G., Pfeifer, R. and Sandini, G. Developmental robotics: A survey, *Connection Science* 15(4), 151–190 2003.

Lyubova, N., Ivaldi, S. and Filliat, D. From passive to interactive object learning and recognition through self-identification on a humanoid robot, *Autonomous Robots* 40(1), 33–57 2016.

Marks, E., Magistri, F. and Stachniss, C. Precise 3d reconstruction of plants from UAV imagery combining bundle adjustment and template matching. In: IEEE Publications International Conference on Robotics and Automation (ICRA) (vol. 2022) 2022, pp. 2259–2265.

Mathis, A., Mamidanna, P., Cury, K. M., Abe, T., Murthy, V. N., Mathis, M. W. and Bethge, M. Deeplabcut: markerless pose estimation of user-defined body parts with deep learning, *Nature Neuroscience* 21(9), 1281–1289 2018. doi: 10.1038/s41593-018-0209-y.

Milioto, A., Lottes, P. and Stachniss, C. Real-time semantic segmentation of crop and weed for precision agriculture robots leveraging background knowledge in cnns. In: IEEE Publications IEEE International Conference on Robotics and Automation (ICRA) (vol. 2018) 2018, pp. 2229–2235.

Mishra, P., Polder, G. and Vilfan, N. Close range spectral imaging for disease detection in plants using autonomous platforms: a review on recent studies, *Current Robotics Reports* 1(2), 43–48 2020. doi: 10.1007/s43154-020-00004-7.

Monarch, R. M. *Human-in-the-Loop Machine Learning: Active Learning and Annotation for Humancentered AI*. Simon and Schuster 2021.

Morrison, D., Milan, A. and Antonakos, E. Uncertainty-aware instance segmentation using dropout sampling. In: Proceedings of the Robotic Vision Probabilistic Object Detection Challenge (CVPR 2019 Workshop), Long Beach, CA 2019, pp. 16–20.

Nasiri, A., Yoder, J., Zhao, Y., Hawkins, S., Prado, M. and Gan, H. Pose estimation-based lameness recognition in broiler using cnn-LSTM network, *Computers and*

Electronics in Agriculture 197, 106931 2022, issn: 0168-1699. doi: 10.1016/j.compag.2022.106931. [Online]. Available at: https://www.sciencedirect.com/science/article/pii/S0168169922002484.

Neupane, K. and Baysal-Gurel, F. Automatic identification and monitoring of plant diseases using unmanned aerial vehicles: a review, *Remote Sensing* 13(19), 3841 2021.

Nir, O., Parmet, Y., Werner, D., Adin, G. and Halachmi, I. 3D Computer-vision system for automatically estimating heifer height and body mass, *Biosystems Engineering* 173, 4–10 2018.

O'Regan, J. K. and Noë, A. A sensorimotor account of vision and visual consciousness, *Behavioral and Brain Sciences* 24(5), 939–73; discussion 973 2001.

Oliveira, L. F. P., Moreira, A. P. and Silva, M. F. Advances in forest robotics: A state-of-the-art survey, *Robotics* 10(2), 53 2021. doi: 10.3390/robotics10020053.

Pantazi, X.-E., Moshou, D. and Bravo, C. Active learning system for weed species recognition based on hyperspectral sensing, *Biosystems Engineering* 146, 193–202 2016.

Pearl, J. The seven tools of causal inference, with reflections on machine learning, *Communications of the ACM* 62(3), 54–60 2019.

Pfeifer, R. and Bongard, J. *How the Body Shapes the Way We Think: a New View of Intelligence*. MIT Press 2006.

Polder, G. and Gowen, A. The hype in Spectral Imaging, *Spectroscopy Europe* 33(3), 12–14 2021.

Polder, G., Blok, P. M., de Villiers, H. A. C., van der Wolf, J. M. and Kamp, J. Potato virus Y detection in seed potatoes using deep learning on hyperspectral images, *Frontiers in Plant Science* 10, 209 2019. doi: 10.3389/fpls.2019.00209. [Online]. Available at: https://www.frontiersin.org/ article/10.3389/fpls.2019.00209.

Pretto, A., Aravecchia, S., Burgard, W., Chebrolu, N., Dornhege, C., Falck, T., Fleckenstein, F., Fontenla, A., Imperoli, M., Khanna, R., Liebisch, F., Lottes, P., Milioto, A., Nardi, D., Nardi, S., Pfeifer, J., Popovic, M., Potena, C., Pradalier, C., Rothacker-Feder, E., Sa, I., Schaefer, A., Siegwart, R., Stachniss, C., Walter, A., Winterhalter, W., Wu, X. and Nieto, J. Building an aerial & ground robotics system for precision farming: an adaptable solution, *IEEE Robotics and Automation Magazine* 28(3), 29–49 2021. doi: 10.1109/MRA.2020.3012492.

Qin, J., Chao, K., Kim, M. S., Lu, R. and Burks, T. F. Hyperspectral and multispectral imaging for evaluating food safety and quality, *Journal of Food Engineering* 118(2), 157–171 2013.

Ratner, A. J., Ehrenberg, H., Hussain, Z., Dunnmon, J. and Ré, C., Learning to compose domain-specific transformations for data augmentation, *Advances in Neural Information Processing Systems* 30 2017.

Ravichandar, H., Polydoros, A. S., Chernova, S. and Billard, A. Recent advances in robot learning from demonstration, *Annual Review of Control, Robotics, and Autonomous Systems* 3(1). ARTICLE, 297–330 2020.

Rawat, S., Chandra, A. L., Desai, S. V., Balasubramanian, V. N., Ninomiya, S. and Guo, W. How useful is image-based active learning for plant organ segmentation?, *Plant Phenomics* 2022, 9795275 2022.

Reed, S., et al. *"A generalist agent," arXiv Preprint ArXiv:2205.06175* 2022.

Rückin, J., Jin, L., Magistri, F., Stachniss, C. and Popović, M. Informative path planning for active learning in aerial semantic mapping. In: IEEE Publications International Conference on Robotics and Automation (ICRA) 2022.

Ruigrok, T., van Henten, E., Booij, J., van Boheemen, K. and Kootstra, G. Application-specific evaluation of a weed-detection algorithm for plant-specific spraying, *Sensors* 20(24), 7262 2020. doi: 10.3390/s20247262.

Ruigrok, T., van Henten, E., Dirks, J. P. and Kootstra, G. Generalization of deep neural networks for weed detection, *Computers and Electronics in Agriculture* 204, 107554 2023. doi: 10.1016/j.compag.2022.107554. Available: https://doi.org/10.1016/j.compag.2022.107554.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C. and Fei-Fei, L. Imagenet large scale visual recognition challenge, *International Journal of Computer Vision* 115(3), 211–252 2015.

Russello, H., van der Tol, R. and Kootstra, G. T-leap: occlusion-robust pose estimation of walking cows using temporal information, *Computers and Electronics in Agriculture* 192, no. 106559 2022.

Sa, I., Ge, Z., Dayoub, F., Upcroft, B., Perez, T. and McCool, C. Deepfruits: A fruit detection system using deep neural networks, *Sensors* 16(8), 1222 2016, issn: 1424-8220. doi: 10.3390/s16081222. [Online]. Available at: https://doi.org/10.3390/s16081222.

Sa, I., Lehnert, C., English, A., McCool, C., Dayoub, F., Upcroft, B. and Perez, T. Peduncle detection of sweet pepper for autonomous crop harvesting—combined color and 3-d information, *IEEE Robotics and Automation Letters* 2(2), 765–772 2017.

Saar, M., Edan, Y., Godo, A., Lepar, J., Parmet, Y. and Halachmi, I. A machine vision system to predict individual cow feed intake of different feeds in a cowshed, *Animal* 16(1), 100432 2022.

Settles, B. Active learning, *Synthesis Lectures on Artificial Intelligence and Machine Learning* 6(1), 1–114 2012.

Shi, W., van de Zedde, R., Jiang, H. and Kootstra, G. Plant-part segmentation using deep learning and multi-view vision, *Biosystems Engineering* 187, 81–95 2019.

Siegwart, R., Nourbakhsh, I. and Scaramuzza, D. *Introduction to Autonomous Mobile Robots*. The MIT Press 2011.

Silwal, A., Davidson, J. R., Karkee, M., Mo, C., Zhang, Q. and Lewis, K. Design, integration, and field evaluation of a robotic apple harvester, *Journal of Field Robotics* 34(6), 1140–1159 2017. doi: 10.1002/rob.21715. [Online]. Available at: https://onlinelibrary. Available at: wiley.com/doi/abs/10.1002/rob.21715.

Smitt, C., Halstead, M., Zaenker, T., Bennewitz, M. and McCool, C. Pathobot: A robot for glasshouse crop phenotyping and intervention. In: IEEE International Conference on Robotics and Automation (ICRA) (vol. 2021), 2021, pp. 2324–2330. doi: 10.1109/ICRA48506.2021.9562047.

Spoliansky, R., Edan, Y., Parmet, Y. and Halachmi, I. Development of automatic body condition scoring using a low-cost 3-dimensional Kinect camera, *Journal of Dairy Science* 99(9), 7714–7725 2016.

Su, D., Kong, H., Qiao, Y. and Sukkarieh, S. Data augmentation for deep learning based semantic segmentation and crop-weed classification in agricultural robotics, *Computers and Electronics in Agriculture* 190, 106418 2021, issn: 0168-1699. doi: 10.1016/j.compag.2021.106418. [Online]. Available at: https://www.sciencedirect.com/science/article/pii/ S016816992100435X.

Suh, H. K., Hofstee, J. W. and van Henten, E. J. Improved vegetation segmentation with ground shadow removal using an hdr camera, *Precision Agriculture* 19(2), 218–237 2018a. doi: 10.1007/s11119-017-9511-z.

Suh, H. K., Hofstee, J. W., IJsselmuiden, J. and van Henten, E. J. Sugar beet and volunteer potato classification using bag-of-visual-words model, scale-invariant feature transform, or speeded up robust feature descriptors and crop row information, *Biosystems Engineering* 166, 210–226 2018b, issn: 1537-5110. doi: 10.1016/j.biosystemseng.2017.11.015. [Online]. Available at: https://www.sciencedirect.com/science/article/pii/S1537511017301629.

Thenmozhi, K. and Reddy, U. S. Crop pest classification based on deep convolutional neural network and transfer learning, *Computers and Electronics in Agriculture* 164, 104906 2019.

Tsouros, D. C., Bibi, S. and Sarigiannidis, P. G. A review on UAV-based applications for precision agriculture, *Information* 10(11) 2019, issn: 2078-2489. doi: 10.3390/info10110349. [Online]. Available at: https://www.mdpi.com/2078-2489/10/11/349.

Turgut, K., Dutagaci, H., Galopin, G. and Rousseau, D. Segmentation of structural parts of rosebush plants with 3d point-based deep learning methods, *Plant Methods* 18(1), 20 2022.

van Essen, R., Nguyen, L., van Helmond, E., Batsleer, J., Poos, J.-J. and Kootstra, G. doi: 10.1093/icesjms/fsab233. [Online]. Automatic bycatch registration using deep learning and object tracking, *International Journal of Marine Science* 78(10), 3834–3846 2021. doi: 10.1093/icesjms/fsab233.

van Essen, R., Harel, B., Kootstra, G. and Edan, Y. Dynamic viewpoint selection for sweet pepper maturity classification using online economic decisions, *Applied Sciences* 12(9), 4414 2022.

van Marrewijk, B., Polder, G. and Kootstra, G. Investigation of the added value of cyclegan on the plant pathology dataset. In Proceedings of the 7th IFAC Conference on Sensing, Control and Automation Technologies for Agriculture. Munich, Germany 2022.

Virlet, N., Sabermanesh, K., Sadeghi-Tehran, P. and Hawkesford, M. J. Field scanalyzer: an automated robotic field phenotyping platform for detailed crop monitoring, *Functional Plant Biology* 44(1), 143–153 2016.

Wagner, N., Kirk, R., Hanheide, M. and Cielniak, G. Efficient and robust orientation estimation of strawberries for fruit picking applications. In: IEEE International Conference on Robotics and Automation (ICRA) (vol. 2021), 2021, pp. 13857–13863. doi: 10.1109/ICRA48506.2021.9561848.

Williams, H. A. M., Jones, M. H., Nejati, M., Seabright, M. J., Bell, J., Penhall, N. D., Barnett, J. J., Duke, M. D., Scarfe, A. J., Ahn, H. S., Lim, J. and MacDonald, B. A. Robotic kiwifruit harvesting using machine vision, convolutional neural networks, and robotic arms, *Biosystems Engineering* 181, 140–156 2019, issn: 1537-5110. doi: 10.1016/j.biosystemseng.2019.03.007. [Online]. Available at: https: //www.sciencedirect.com/science/article/pii/S153751101830638X.

Winterhalter, W., Fleckenstein, F., Dornhege, C. and Burgard, W. Localization for precision navigation in agricultural fields—beyond crop row following, *Journal of Field Robotics* 38(3), 429–451 2021. doi: 10.1002/rob.21995. eprint. [Online]. Available at: https://onlinelibrary.wiley/abs/10.1002/rob.21995.

Xie, C. and Yang, C. A review on plant high-throughput phenotyping traits using UAV-based sensors, *Computers and Electronics in Agriculture* 178, 105731 2020, issn: 0168-1699. doi: 10.1016/j.compag.2020.105731. [Online]. Available at: https://www.sciencedirect. com/science/article/pii/S0168169919320046.

York, L. M. Functional phenomics: an emerging field integrating high-throughput phenotyping, physiology, and bioinformatics, *Journal of Experimental Botany* 70(2), 379–386 2019, issn: 0022-0957. doi: 10.1093/jxb/ery379. eprint. Available at: https://academic.oup.com/jxb/article-pdf/70/2/379/27435247/ery379.pdf. [Online].

Zaenker, T., Smitt, C., McCool, C. and Bennewitz, M. Viewpoint planning for fruit size and position estimation. In: IEEE Publications International, R. S. J. (Ed.) Conference on Intelligent Robots and Systems (IROS) (vol. 2021), 2021 pp. 3271–3277.

Zaenker, T., Lehnert, C., McCool, C. and Bennewitz, M. Combining local and global viewpoint planning for fruit coverage. In: IEEE Publications European Conference on Mobile Robots (ECMR) (vol. 2021), 2021 pp. 1–7.

Zapotezny-Anderson, P. and Lehnert, C. Towards active robotic vision in agriculture: A deep learning approach to visual servoing in occluded and unstructured protected cropping environments, *IFAC-PapersOnLine* 52(30), 120–125 2019, 6th IFAC Conference on Sensing, Control and Automation Technologies for Agriculture AGRICONTROL, issn: 2405-8963. doi: 10.1016/j.ifacol.2019.12.508. [Online]. Available at: https://www.sciencedirect.com/science/article/pii/S2405896319324243.

Zeng, W., Karaoglu, S. and Gevers, T. 2018. Inferring point clouds from single monocular images by depth intermediation. doi: 10.48550/ARXIV.1812.01402. [Online] Available at: https://arxiv.org/ abs/1812.01402.

Zeng, X., Zaenker, T. and Bennewitz, M. Deep reinforcement learning for next-best-view planning in agricultural applications. In: IEEE Publications International Conference on Robotics and Automation (ICRA) (vol. 2022), 2022, pp. 2323–2329.

Zhang, Z., Wang, X., Lai, Q. and Zhang, Z. Review of variable-rate sprayer applications based on realtime sensor technologies, *Automation in Agriculture-Securing Food Supplies for Future Generations*, 53–79 2018.

Zhao, Y., Gong, L., Huang, Y. and Liu, C. A review of key techniques of vision-based control for harvesting robot, *Computers and Electronics in Agriculture* 127, 311–323 2016, issn: 01681699. doi: 10.1016/j.compag.2016.06.022. [Online]. Available at: https://www.sciencedirect.com/science/article/pii/S0168169916304227.

Zhou, C., Zhang, B., Lin, K., Xu, D., Chen, C., Yang, X. and Sun, C. Near-infrared imaging to quantify the feeding behavior of fish in aquaculture, *Computers and Electronics in Agriculture* 135, 233–241 2017.

Zhu, N., Liu, X., Liu, Z., Hu, K., Wang, Y., Tan, J., Huang, M., Zhu, Q., Ji, X., Jiang, Y. and Guo, Y. Deep learning for smart agriculture: concepts, tools, applications, and opportunities, *International Journal of Agricultural and Biological Engineering* 11(4), 21–28 2018.