

## Review

# RGB-D datasets for robotic perception in site-specific agricultural operations—A survey

Polina Kurtser<sup>a,b,\*</sup>, Stephanie Lowry<sup>a</sup>

<sup>a</sup> Centre for Applied Autonomous Sensor Systems, Örebro University, Sweden

<sup>b</sup> Department of Radiation Science, Radiation Physics, Umeå University, Sweden

## ARTICLE INFO

## Keywords:

3D perception  
Color point clouds  
Datasets  
Computer vision  
Agricultural robotics

## ABSTRACT

Fusing color (RGB) images and range or depth (D) data in the form of RGB-D or multi-sensory setups is a relatively new but rapidly growing modality for many agricultural tasks. RGB-D data have potential to provide valuable information for many agricultural tasks that rely on perception, but collection of appropriate data and suitable ground truth information can be challenging and labor-intensive, and high-quality publicly available datasets are rare. This paper presents a survey of the existing RGB-D datasets available for agricultural robotics, and summarizes key trends and challenges in this research field. It evaluates the relative advantages of the commonly used sensors, and how the hardware can affect the characteristics of the data collected. It also analyzes the role of RGB-D data in the most common vision-based machine learning tasks applied to agricultural robotic operations: visual recognition, object detection, and semantic segmentation, and compares and contrasts methods that utilize 2-D and 3-D perceptual data.

## 1. Introduction

The growing world population, climate change, and rising demand for food and crops have been driving forces of global efforts to increase crop production yields in the last few decades (Zhao et al., 2019a; Bac et al., 2014; Kamilaris and Prenafeta-Boldú, 2018). These increasing demands coupled with the declining number of skilled workforce and rising labor costs are currently major market forces behind the rapid development and implementation of both precision agriculture data-driven algorithms for crop monitoring and increase of yield and resilience, and automation efforts to replace manual labor with robots (Bechar and Vigneault, 2016; Bac et al., 2014). Automation efforts are introduced into all labor-intensive aspects of agricultural work, such as transplanting and seeding (Bechar and Vigneault, 2016; Haibo et al., 2015; Ruangurai et al., 2015), harvesting (Arad et al., 2020; Bac et al., 2014; Tang et al., 2020), and pruning (He and Schupp, 2018).

Visual sensing is a key part of any automated agricultural robotic system. Integrated autonomous in-field robots use color (RGB) or color fused with depth data (RGB-D) cameras for target localization (Fig. 1) in operations such as harvesting (Bac et al., 2014; Arad et al., 2020), weeding (Harders et al., 2021) pruning (Zahid et al., 2021) and crop monitoring (Kurtser et al., 2020b). Furthermore, range data provided by RGB-D cameras or Light Detection And Ranging sensors (LiDAR) are essential for many agricultural tasks: integrated robots that have

access to range data can perform harvesting and other manipulation tasks more efficiently (Barth et al., 2016; Ringdahl et al., 2019).

However, despite the growing interest in RGB-D data as a vital part of an agricultural robot's sensor suite (Fu et al., 2020), there are limited available open-access RGB-D datasets in the agricultural domain.

This paper reviews the available open-access RGB-D datasets that could potentially benefit the development of robots for site-specific agricultural operations. It summarizes the datasets available, the hardware and technology used to acquire such datasets, and the applications for which they can be used. It also discusses the gaps in available datasets and how it affects the development of the research field.

The main contribution of this paper is twofold. Firstly, it provides what is, to the authors' knowledge, the first comprehensive survey of RGB-D datasets that can be used for site-specific agricultural operations, to assist researchers to select a fitting dataset for their needs. Secondly, it evaluates the relationship between the available datasets and the algorithms that use those data.

The addressed topic is relevant and timely. There has been a rapid improvement in the field of computer vision over the last decade, due to an improvement in hardware and sensor technology, computational power, and the subsequent improvement in algorithms. At the same time, the related but separate field of RGB-D perception is also growing rapidly, with the development of capable, accurate, and

\* Correspondence to: Fakultetsgatan 1 Örebro, 70281, Sweden.

E-mail addresses: [polina.kurtser@umu.se](mailto:polina.kurtser@umu.se) (P. Kurtser), [stephanie.lowry@oru.se](mailto:stephanie.lowry@oru.se) (S. Lowry).



Fig. 1. Agricultural robots employing RGB-D sensors for site specific tasks. Left: SWEEPER- sweet pepper harvesting robot. Source — <http://www.sweeper-robot.eu/> (Arad et al., 2020); Right: Yield estimation robot in vineyards described by Kurtser et al. (2020b).

affordable RGB-D sensors, with greater functionality in the outdoor lighting conditions required by site-specific agricultural operations (Vit and Shani, 2018; Neupane et al., 2021; He et al., 2017b). A recent survey article (Lopes et al., 2022) presented 231 accessible RGB-D datasets. However, it contained none of the agricultural datasets included in this survey article. The supporting algorithms for RGB-D data are similarly rapidly developing, with the first widely used deep learning algorithms specifically for point cloud data becoming available in recent years, such as PointNet (Qi et al., 2017a). Since the release of PointNet in 2017, new algorithms are being released, such as LatticeNet in 2022 (Rosu et al., 2022).

The paper proceeds as follows. Section 2 provides background on the most relevant research strands. Section 3 introduces the datasets, and brief summaries of them can be found in the appendix. Section 4 discusses the possible applications that these datasets can be used for. The paper closes with discussion (Section 5) and conclusion (Section 6).

## 2. Background

### 2.1. Visual sensing and range data for agriculture

Visual sensing plays a key role in agricultural applications. Both precision agriculture algorithms and in-field automation efforts rely heavily on the processing of sensor data for navigation and carrying out site-specific tasks. Visual information is used for generating collections of phenotypic traits — observable expressions of underlying genes. In the past decade plant phenomics (the study of phenotypes) has evolved to a thriving research field supporting breeders in the search for crops with needed agronomic traits (Zhao et al., 2019a). Intelligent solutions for precision breeding and automated phenotyping platforms in controllable and field environments are emerging to support the collection of high-throughput phenotypic data (Zhao et al., 2019a; Yang et al., 2020; Araus and Cairns, 2014).

Visual sensing is also vital for agricultural automation tasks. While phenotypic traits data are often collected using a variety of sensors (i.e. hyper and multi spectral cameras, thermal cameras, and 3-D scanners), integrated autonomous in-field robots are typically reliant on color or RGB-D cameras for target localization in operations such as harvesting (Bac et al., 2014), weeding (Harders et al., 2021) and pruning (Zahid et al., 2021). Autonomous robots also require range data to perform visual servoing operations, as relying only on 2-D imagery rather than accessing range data directly makes a robot prone to long cycle times due to the correction of position (Barth et al., 2016; Ringdahl et al., 2019).

### 2.2. Dataset collection for agricultural robotics

There are many tasks within agricultural robotics that are based on machine learning algorithms. For example, the deep learning algorithm YOLO (Redmon et al., 2016a) has shown great success for detection and counting of crops such as tomatoes (Liu et al., 2020) and apples (Tian et al., 2019a), and plant diseases detection (Loey et al., 2020). However, machine learning algorithms rely on large variable datasets to advance the modeling and operative capabilities (Bac et al., 2014; Kamilaris and Prenafeta-Boldú, 2018; Kamilaris et al., 2017). As a result, for the rapid development of vision and perception systems for site-specific operating agricultural robots, the ongoing creation of relevant datasets is required.

Collection and annotation of large datasets have been at the base of the success of many other industries. Great advancements in autonomous urban driving have to do with the availability of public datasets such as KITTI (Geiger et al., 2013), CityScapes (Cordts et al., 2015) and Berkley's deep drive (Yu et al., 2020). Similarly, the rapid development of deep learning techniques for detection, recognition, and classification (LeCun et al., 2015) was driven by access to large labeled datasets such as ImageNet (Deng et al., 2009) and COCO (Lin et al., 2014).

Large annotated datasets are also increasingly collected in the agricultural application domain (Chi et al., 2016). Agriculture ecosystems are analyzed and understood through monitoring and measuring continuously various aspects of the physical environment. Such aspects include weather and environmental data such as ambient and soil temperature and humidity, light, wind, amount of perception, soil acidity, and many more (Kamilaris et al., 2017).

In-field imaging datasets for phenotyping, plant monitoring and site-specific operations are scarce (Kamilaris and Prenafeta-Boldú, 2018; Lu and Young, 2020) and focused almost entirely on RGB data. In fact, despite many robotic application relying heavily on RGB-D data (Zahid et al., 2021), open access RGB-D datasets are almost non-existent for in-field conditions. Until recently, this could have been partially attributed to the lack of commercial grade RGB-D sensors operating reliably in outdoor conditions (Vit and Shani, 2018). However, with recent developments in stereo-IR and LiDAR, RGB-D and range sensors are more widely available for data collection (Fu et al., 2020).

Some aspects of operation of in-field agricultural robotics can be reliant on datasets collected in non-agricultural conditions. For example, mobile autonomous navigation using algorithms and datasets from other applications domains, such as KITTI and CityScapes, has been shown to be successful in agricultural settings (Mousazadeh, 2013). Similarly, agricultural tasks such as fruit detection use transfer learning (Hameed et al., 2018). The detection system can be based on

existing deep learning algorithms such as Faster R-CNN (Ren et al., 2015) and Mask R-CNN (He et al., 2017a) using pre-trained models learned from large, generic image datasets. The models are then fine-tuned using the agriculture-specific data. However, even when transfer learning is applied, task-specific datasets are still required for fine-tuning the algorithms.

### 3. Datasets

#### 3.1. Criteria for inclusion

To survey the relevant publicly available agriculture datasets that include both visual and depth data, research databases including Google Scholar, Web of Science, and IEEE Xplore were searched with combinations of the following keywords: “point cloud”, “datasets”, “RGB-D”, “agriculture”. Database repositories Quantitative-plant (Lobet et al., 2013) and Zenodo (European Organization For Nuclear Research and OpenAIRE, 2013) were searched. Quantitative-plant was comprehensively surveyed while searching on Zenodo was performed with a combination of the following keywords: “plant name (e.g. broccoli, orange)”, “3-D”, “RGB-D”. Other review papers were checked for Lu and Young (2020), Patrício and Rieder (2018), Kamilaris and Prenafeta-Boldú (2018). However, it is noted that publication of datasets is not always systematic: some publications are dedicated to the dataset and contain that in the title (Schunck et al., 2021; Bender et al., 2020; Chebrolo et al., 2017; Akbar et al., 2016; Kitzler et al., 2023), but many papers do not (Kusumam et al., 2016, 2017; Khanna et al., 2019; Halstead et al., 2020; Arad et al., 2019; Kurtser et al., 2020a,b) and the publication of the data was often incidental to the algorithm or systems development aspect of the paper. Therefore, many of these datasets were only found by close reading of the papers to confirm if a publicly available dataset was included. The authors would therefore encourage researchers to consider publishing a specific dataset paper if they release their data, even in a specific data journal (such as Data in Brief) as was done by Gené-Mola et al. (2019a, 2020b,d, 2021b). The interested reader might find additional useful guidelines on dataset release published by Lu and Young (2020) relevant for heavy visual information files like images and point clouds, including suggested repositories. There are also recent initiatives such as the Eden Library (Mylonas et al., 2022) which aim to provide a platform for open access crop and plant databases covering proximal and aerial images. A similar platform that allowed for the inclusion RGB-D and depth data would be greatly beneficial.

Most datasets were collected in outdoor field conditions, but there are also three datasets from greenhouses (Khanna et al., 2019; Halstead et al., 2020; Arad et al., 2020) and one indoor phenotypic dataset (Schunck et al., 2021), where plants grown in pots were carefully scanned to generate dense high-quality point clouds with the aim of complete coverage, plant stillness, and lack of occlusion. However, the datasets were restricted to those that included growing plants. Therefore, datasets that exclusively contained scanned fruit after harvesting, such as Durand-Petiteville et al. (2018), were excluded.

Because of the sparsity of available datasets in the agricultural domain, this paper not only includes RGB-D datasets but LiDAR datasets where intensity values are also included. This paper also includes two datasets that do not include point cloud data (Bender et al., 2020; Milella et al., 2019), but that do include stereo imagery that could potentially be used to generate depth information.

It is to be noted that visible light (RGB and some LiDARs), NIR and IR (in some stereo RGBD and LiDARs) sensing are not the only visual sensing modalities frequently used in agricultural applications. Other sensors such as X-ray tomography (Dutagaci et al., 2020), hyperspectral cameras (Bender et al., 2020; Khanna et al., 2019), and magnetic resonance imaging (MRI) (Pflugfelder et al., 2017) are also sometimes used for plant analysis and phenotyping. However, since this paper

focuses on visual sensor modalities that are commonly used for robotic operations, we include only papers that produce 3-D data.

Additionally, this paper focuses on site-specific acquisitions — measurements made from ground level, enabling a proximal viewpoint and a direct manipulation of the crop by the robot. As a result, papers describing data acquisitions from remote platforms (Vélez et al., 2022) such as UAVs at high flight distance, were not included. Nevertheless, papers describing data collections for site specific plant monitoring applications without direct plant manipulations were included.

Finally, this paper reviews exclusively dataset papers that provide sensory data along ground truth information. Papers that did not contain either manually acquired image labels or phenotypic metadata collected in the field were excluded from the review (Barbole and Jadhav, 2023).

#### 3.2. Summary of datasets

A total of 24 papers describing 16 public datasets were identified according to the criteria described in Section 3.1. All the papers reviewed are summarized in Tables 1–2. The available datasets represent a range of agricultural plants, including maize (Schunck et al., 2021), tomato (Schunck et al., 2021), broccoli (Kusumam et al., 2016, 2017; Blok et al., 2021a,b; Bender et al., 2020), apples (both growing (Gené-Mola et al., 2019a,b, 2020a,b,c,d) and dormant (Akbar et al., 2016) trees), grapes (Milella et al., 2019; Marani et al., 2021; Kurtser et al., 2020a,b), cauliflower (Bender et al., 2020), sugar beets (Chebrolo et al., 2017; Khanna et al., 2019; Kitzler et al., 2023), sweet peppers (Halstead et al., 2020; Arad et al., 2019). However, these plants represent only a small fraction of the plants cultivated in global agriculture and horticulture (see, for example, Hameed et al., 2018).

Furthermore, it is evident that RGB-D datasets for agriculture are a relatively recent phenomenon. All the datasets found during the literature search were from 2016 or later (see Fig. 2). However, publicly available datasets for agriculture based on other visual modalities may also be similarly recent: a survey of 34 computer vision datasets for agriculture (Lu and Young, 2020) also only contained datasets from 2015. A reader interested in a survey of visual datasets in agriculture, is kindly referred to Lu and Young (2020).

#### 3.3. Hardware and technology

A common denominator of the reviewed papers hardware is the significant preference to employ commercial grade sensors. Microsoft Kinect V2 (Microsoft, Redmond, Washington, United States) which relies on time-of-flight technology (ToF) has been employed in Kusumam et al. (2016, 2017), Gené-Mola et al. (2019a,b), Chebrolo et al. (2017), Akbar et al. (2016) to collect fully registered RGB and range (D) data, while another less common ToF-based RGB-D camera (produced by Fotonic) was employed by Arad et al. (2019). Similarly, registered RGB-D images were collected by a number of different stereo-IR sensors produced by Intel (Intel, Santa Clara, California, United States) in their RealSense series by Milella et al. (2019), Marani et al. (2021), Halstead et al. (2020), Khanna et al. (2019), Kurtser et al. (2020a,b), Blok et al. (2021a,b). Stereo vision was collected by Bender et al. (2020), Blok et al. (2021b,b) in the form of overlapping RGB or monochrome images that can be transferred into registered color point clouds. Overlapping RGB images were also used by Gené-Mola et al. (2020c,d) for 3-D point cloud reconstruction using structure from motion.

Exceptions to this rule are Gené-Mola et al. (2020a,b), Chebrolo et al. (2017) employing a 3-D Velodyne LiDAR (Velodyne LiDAR, San Jose, California, United States) for their point cloud collections. While Gené-Mola et al. (2020a,b) employed the LiDAR to extract reflectance-based features for site specification operations, Chebrolo et al. (2017) employed the LiDAR for mainly autonomous mobile robot navigation purposes and equipped their platform with a Kinect V2 as well. Finally Schunck et al. (2021) employed a laser triangulation scanner for collection of detailed point clouds.

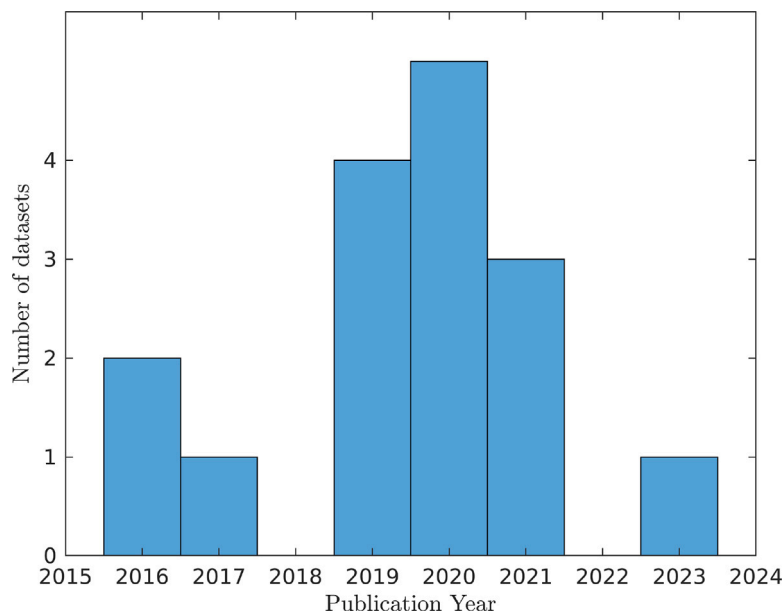


Fig. 2. Publication years of the dataset papers in this survey.

### 3.3.1. Effect of chosen hardware on dataset measures

Sensors relying on similar technology generally generate comparative datasets in terms of specifications, and therefore are more useful for training and fine tuning of models reliant on the collected datasets. Depth measurements collected with very different technology can be assumed to generate point clouds very different in nature and therefore may limit the transferability of the results.

As mentioned above, the data collection most commonly relied on commercial grade sensors, such as Microsoft Kinect and Intel RealSense sensors. The different technologies to produce point clouds that were used in the reviewed papers can be divided into three types: Time-of-Flight (ToF) cameras, LiDAR, and Stereo RGB/IR cameras. These technologies are often fused with an additional (in the case of ToF and LiDAR) or already integrated (in the case of stereo) RGB camera for the production of colored point clouds.

**Time-of-flight (ToF) cameras.** Time-of-flight (ToF) sensors estimate the distance based on the time difference between an emitted light signal and its return time to the sensor. Typically ToF cameras would project a light or IR pattern on the captured scene and measure the time it takes for the reflected pattern to be measured by the sensor. In the case the projected signal is periodic, the phase shift between the transmitted and the received signal can be used as an indicator of the round-trip time (Giancola et al., 2018). As a result, the technology is highly hindered by light obstructions, which makes its use in outdoor agriculture environments challenging. The use of IR or Near Infra-Red (NIR) light signals helps mitigating the affect of natural light interference to some extent, but coping with ambient light remains a source of large errors in depth estimation in natural outdoor scenes.

ToF cameras were used to create datasets in a few of the reviewed papers. In these cases the limitations of the ToF camera were mitigated by collecting data only during night conditions Gené-Mola et al. (2019a,b) or with the use of specially constructed enclosures to block the incoming ambient light Kusumam et al. (2016, 2017). To collect complementary color information artificial lighting was used.

**LiDAR.** LiDAR sensors operate in a similar way to the ToF sensors, where distance to objects is calculated based on the time between an emitted and received signal. LiDAR sensors often use emitted light at a higher frequency than the ToF cameras. Commercial grade LiDAR have shown to be significantly more robust to light abstractions compared to the cheaper and weaker ToF cameras (He et al., 2017b).

However, non solid-state LiDARs, used by the authors of the datasets collected in this paper, rely on a set of laser emitters, rotating in one

or several 2-D planes, and therefore their resolution is limited by the number of rotating lasers in 3-D. For example, the Puck VLP-16 LiDAR sensor – one of the most expensive sensors used for dataset collection in the reviewed papers – exhibits impressive abilities of measurements range of up to 100 m, using 16 channels. At that distance it provides accuracy of  $\pm 3$  cm.

Due to its long range and highly accurate distance data, the LiDAR is able to perform mapping and localization required for autonomous navigation, and is a common and popular choice for robot localization systems. However, the employed non solid state LiDAR performance is less suitable for robotic operations that require short-range, high-resolution performance, such as any tasks that requires manipulation of the plant. For example, the typical distance between growing rows in orchard, vineyard and greenhouse conditions is 1.5 m-2 m, and the industrial robotic manipulator has a reachability range of around 1 m. Thus if a LiDAR such as the VLP-16 is mounted on a mobile platform or on the robotic manipulator navigating between the growing rows, as often being done as part of the acquisition protocols, a reasonable acquisition distance is around 1 m-2 m.

At that distance range (1 m-2 m) the Puck VLP-16 LiDAR exhibits a significant limitation given the 2° vertical resolution and the 30° field of view resulting with a maximum of 3.5 cm vertical resolution for objects placed at 1 m straight in-front of the camera (Fig. 3). In other words vertical localization accuracy is subject to a several centimeters error from the sensor data itself, an unacceptable value for most robotic manipulation tasks. The vertical field of view is also rather limited for an optimized operation as it only covers about 0.5 m window at a 1 m distance, although this might be sufficient for some manipulation tasks as it is in line with the reachability area of typical industrial robotic manipulator.

The resolution limitations of the LiDAR will potentially be mitigated with the solid state 3-D LiDAR technology which recently drawn interests of the academic research community (Li et al., 2022b). Some of these LiDARs rely on a use of a photodetector (PD) array, to capture the entire target scene within a single shot, thus eliminating the need to mechanically rotate laser emitters. With incorporation of microelectromechanical mirrors the depth resolution and accuracy at close range sensing are expected to be sufficient for close range detection robotic operations. For example the Livox Mid-360 (Livox Technology Company Limited, Shenzhen, Guangdong, China) specifications<sup>1</sup> declare a

<sup>1</sup> <https://www.livoxtech.com/mid-360/specs>

**Table 1**  
Summary of reviewed papers-Part I.

Dataset	Crop	RGB-D technology	Hardware	Dataset size	Acquisition protocol	Ground truth/ Manual labels
Pheno4D <sup>a</sup> (Schunck et al., 2021)	Maize Tomato	Laser triangulation scanner	Perceptron Scan Works V5	DB#1 — 84 pcls DB#2 — 140 pcls	DB#1 — 12 days, 7 plants DB#2 — 20 days, 7 plants	DB#1 — 49 labeled pcls DB#2 — 77 labeledpcls
Broccoli3D <sup>b</sup> (Kusumam et al., 2016, 2017)	Broccoli	Time-of-flight (ToF)	Kinect V2	4 datasets: #1: 4227 pcls #2: 9239 pcls #3: 14336 pcls #4: 14420 pcls	Scans in different countries and different weather conditions	labeled point clouds and manual size measurements
KFuji RGB-DS <sup>c</sup> (Gené-Mola et al., 2019a,b)	Fuji Apples	Time-of-flight	Kinect V2	967 images	Single collection	12839 manual apple annotations
LFuji-air <sup>d</sup> (Gené-Mola et al., 2020a,b)	Fuji Apples	LiDAR	Puck VLP-16 Velodyne	88 pcls	11 Fuji apple trees 8 scans per tree multiple viewpoints	1353 manually annotated apples
S3CavVineyardDataset <sup>e</sup> (Milella et al., 2019; Marani et al., 2021)	Grapes	Stereo IR	RealSense R200	500 RGB images (D missing)	2 days of filming at 5 fps camera on a mobile platform single viewpoint	85 manual annotations
Ladybird Cobbitty Brassica <sup>f</sup> (Bender et al., 2020)	Cauliflower Broccoli	Stereo RGB	Grasshopper	3-4 images\plant	Weekly scans for 10 week Four growing beds, 144 per bed	Height, width, fresh and dry weight
Eschikon plant stress phenotyping <sup>g</sup> (Khanna et al., 2019)	Sugarbeet	Stereo IR	-RealSense ZR300	496 stereo IR pairs	biweekly (for 2 month)	Reference plant trait measurements
Sweet Pepper Detection <sup>h</sup> (Halstead et al., 2020)	Sweet-pepper	Stereo IR	- RealSense 200 - RealSense 435i	DB#1 — 1583 images DB#2 — 687 images DB#3 — 286 images	Different light conditions (sunlight, polytunnel, glasshouse)	DB#1 — 5774 annotations DB#2 — 3741 annotations DB#3 — 3724 annotations
SWEeper <sup>i</sup> (Arad et al., 2019)	Sweet-pepper	Time-of-flight	Fotonic F80	156 scenes, 468 images	2 illumination conditions	344 annotated peppers

<sup>a</sup><https://www.ipb.uni-bonn.de/data/4d-plant-registration/>.

<sup>b</sup>[https://icas.lincoln.ac.uk/nextcloud/shared/agritech-datasets/broccoli/broccoli\\_datasets.html](https://icas.lincoln.ac.uk/nextcloud/shared/agritech-datasets/broccoli/broccoli_datasets.html).

<sup>c</sup><https://www.grap.udl.cat/en/publications/kfuji-rgb-ds-database/>.

<sup>d</sup><https://www.grap.udl.cat/en/publications/lfuji-air-dataset/>.

<sup>e</sup><https://github.com/ispstiima/S3CavVineyardDataset>.

<sup>f</sup><https://ses.library.usyd.edu.au/handle/2123/20187>.

<sup>g</sup><https://projects.asl.ethz.ch/datasets/doku.php?id=2018plantstressphenotyping>.

<sup>h</sup><https://data.researchdatafinder.qut.edu.au/dataset/qut-hia-daf-capsicum-datasets>.

<sup>i</sup>[http://icvl.cs.bgu.ac.il/lab\\_projects/agrovision/DB/Sweeper04/#/scene](http://icvl.cs.bgu.ac.il/lab_projects/agrovision/DB/Sweeper04/#/scene).

0.15° angular resolution. None of the reviewed papers used such a sensor in their data collections and the sensor has yet to be widely tested in outdoor agricultural conditions. However, the continuous and ongoing improvement in sensor technologies will play an important role in future developments of perception-based agricultural systems.

**Stereo cameras (RGB or IR based).** Acquisition of depth information from stereo relies on registration of images from a number of viewpoints, from one viewpoint coordinate system to the other. This can be done either through acquisition from multiple mono lens cameras or from stereo cameras (RGB or IR based) that are equipped with 2 or more lenses with a separate image sensor, for each lens. The quality of the point cloud will rely highly on both the quality of the images as in their ability to provide detailed scene information, as well as the quality of the calibration procedure placing the relative localization of one viewpoint to another. Most of the stereo-IR cameras reviewed in this paper are multi-lens systems enclosed in a rigid factory made casing (i.e. the RealSense cameras) and therefore the authors opted to rely on the factory calibration procedures. Nevertheless, some authors opted to acquire stereo information from a number of mono lens cameras (Bender et al., 2020; Kitzler et al., 2023) or perform their own calibration (Milella et al., 2019) of a multi-lens camera.

Rather than simply using stereo RGB, a popular choice is to use stereo-IR imaging for stereo registration, due to the better outdoor performance of stereo-IR imaging over stereo-RGB. The most common stereo IR technology used to create the surveyed datasets was Intel's RealSense cameras.

The specified resolution of the Intel RealSense is up to 60° field of view at 1280 × 720 pixels maximum resolution. This implies that under close-range imaging conditions of 1 m distance (as required for manipulation tasks) the imaging windows would account for 1.15 m vertical windows of acquisition and 0.15 cm vertical resolution. This is a much higher resolution result than from the mentioned above and widely used LiDAR, which could only achieve 3.5 cm of vertical resolution. In other words, RGB-D data acquired by stereo IR sensors are expected to be significantly denser compared to LiDAR.

Furthermore, neither the LiDAR or the ToF cameras exhibit an intrinsic ability to provide colored point clouds (that is, point clouds with associated RGB data). An RGB camera must be coupled with the range sensor and registered further to provide color data, which is an additional source of potential error, while stereo cameras can provide colored point clouds without requiring additional sensors and calibration.

**Table 2**  
Summary of reviewed papers — Part II.

Dataset	Crop	RGB-D technology	Hardware	Dataset size	Acquisition protocol	Ground truth/ Manual labels
Grapes3D <sup>a</sup> (Kurtser et al., 2020a,b)	Grapes	Stereo IR	RealSense D435	DB#1 — c.a.1000 pcl (6 videos) DB#2 — c.a.7000 pcls (4 videos)	2-3 viewpoint condition, DB#1 — single row, DB#2 — 2 rows	DB#1 — 17 clusters, 10 plants, DB#2 — 8-19 clusters, 5 plants no annotations provided
Sugarbeets2016 <sup>b</sup> (Chebroul et al., 2017)	Sugarbeet	-Time-of-flight -LiDAR	-Kintect V2 -Puck VLP-16 Velodyne	5 TB of recordings rosbag format	30 days of recording 3 days a week	300 labeled images (non RGB-D)
Apple Trees <sup>c</sup> (Akbar et al., 2016)	Apple Trees	Time-of-flight	Kintect V2	c.a. 3200 pcls	Indoor and outdoor	9 trees
Broccoli heads <sup>d</sup> (Blok et al., 2021a,b)	Broccoli	- Stereo Mono - Stereo IR	- IDS Ensensio N35 - RealSense D435	DB#1 — 947 pcls DB#2 — 1613 pcls	DB#1 — Enclosed box DB#2 — natural light with umbrella	DB#1 — size (122 heads) DB#2 — size (250 heads)
Fuji-sfm <sup>e</sup> (Gené-Mola et al., 2020c,d)	Fuji Apple	Structure from motion	EOS 60D DSLR Cannon RGB camera	582 raw images single reconstruction	Free hand acquisition 5-6 img/position two row sides	1455 3-D apple box annotations
PFuji-size <sup>f</sup> (Gené-Mola et al., 2021a,b)	Fuji Apple	-Structure from motion -Multiview Stereo	EOS 60D DSLR Cannon RGB camera	-DB#1:856 raw images single reconstruction -DB#2 — 4500 images 25 apple point clouds	-DB#1: Free hand acquisition 5-6 img/position -DB#2:multiview turn table	-DB#1: 615 3-D apple segmented point clouds and size annotations -DB#2: 25 size annotations
WE3DS <sup>g</sup> (Kitzler et al., 2023)	Field seedlings 17 species	Stereo Mono	XIMEA MC023CG-SY	6224 image pairs	Top view continues (1 fps) different fields 25 measurement dates	2568 annotated label maps of 17 plant species classes

<sup>a</sup><https://sites.google.com/view/grapes3d>.

<sup>b</sup><http://www.ipb.uni-bonn.de/data/sugarbeets2016/>.

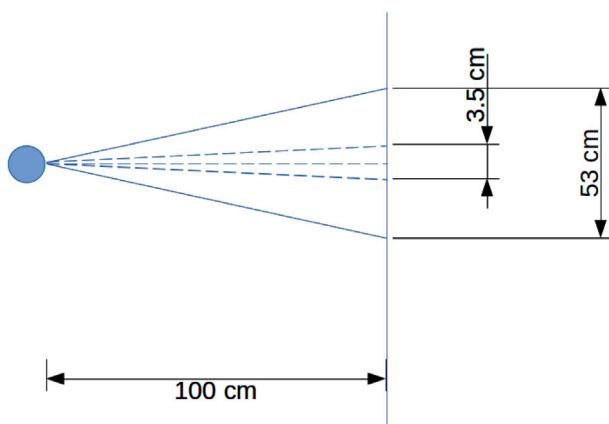
<sup>c</sup>[https://engineering.purdue.edu/RVL/CVPRW\\_Dataset/](https://engineering.purdue.edu/RVL/CVPRW_Dataset/).

<sup>d</sup><https://research.wur.nl/en/datasets/data-underlying-the-publication-image-based-size-estimation-of-br>.

<sup>e</sup><https://zenodo.org/record/3712808#.XnD82iNCe01>.

<sup>f</sup><https://doi.org/10.34810/data141>.

<sup>g</sup><https://zenodo.org/record/7457983#.ZFpl3NJBxhE>.



**Fig. 3.** Vertical resolution and field of view of a Puck VLP-16 LiDAR in realistic acquisition scenario. Any objects smaller than 3.5 cm in vertical dimension will be missed or subject to quantization error at acquisition distance of 1 m.

Overall, commercial grade stereo IR cameras (specifically Intel RealSense D435), and commercial grade ToF cameras, currently appear to be very popular sensors employed in agricultural robotics applications. This is not necessarily the most suitable sensor for all applications (Neupane et al., 2021), but as we show in this review, a very wide spread. A stereo IR camera provides high-resolution, high field-of-view images,

with factory calibrated colored point clouds, and exhibits often good performance in outdoor environments in the tested lighting conditions. But in some cases, as presented by Neupane et al. (2021), commercial grade ToF cameras such as the Microsoft Azure Kinect, exhibit more accuracy in depth measurements. This specific sensor was not used by any of the dataset papers reviewed. It is important to note though that stereo sensors usually require less power than ToF, which allows smaller and lighter devices, very suitable for integration in robotic arms. A disadvantage of stereo IR sensing compared to LiDAR is that LiDAR based point clouds are not so dependent on the individual calibration of the specifically used sensor, and therefore LiDAR point clouds will exhibit less variation in range measurement accuracy. But, for the acquisition of colored point cloud registration between a LiDAR point cloud and a color image is needed, exhibiting the same registration issues mentioned above.

### 3.3.2. Acquisition protocol

In order for an acquired dataset to be usable in the agricultural robotics setting, it most often needs to be acquired in similar conditions in which the robot is to operate. Representative scenarios include both environmental factors (i.e., datasets acquired in field conditions during different times of day and season) as well as acquisition protocol (Kurtser et al., 2016). For example, manually acquired imaging data focusing on in-field phenotyping procedures will place the target fruit in the center of the image. This scenario can be reasonable when in-field semi-manual phenotyping procedures are planned, but cannot be assumed in conditions of continuous acquisition along crop rows for

the detection of crops. In continuous robotic acquisitions, the target can be assumed in a wider variety of locations, illumination conditions and presence of occlusion. Therefore the most realistic scenario in acquisition of in-field dataset is equipping a robot with RGB-D sensors to perform continuous acquisition.

In the reviewed papers, many of the in-field datasets were indeed acquired using a sensor suite mounted on either an autonomous robot (Bender et al., 2020) or a manually driven tractor or mobile platform (Kusumam et al., 2016, 2017; Gené-Mola et al., 2020a,b; Milella et al., 2019; Marani et al., 2021; Chebroly et al., 2017; Blok et al., 2021a,b; Gené-Mola et al., 2019a,b; Kurtser et al., 2020a,b; Kitzler et al., 2023). Some of the datasets also manually selected a subset of the images, post-capture (Milella et al., 2019; Marani et al., 2021).

The other datasets used manual positioning of cameras: either for the full dataset (Gené-Mola et al., 2020c,d, 2021b,a; Akbar et al., 2016), or for part of the dataset (Blok et al., 2021a,b). For the sweet pepper datasets, acquisition was manual in one case (Halstead et al., 2020), and in another the process included a combination of manual placing of the robot with automatic viewpoint selection (Arad et al., 2019). A special case was the phenotypic dataset (Schunck et al., 2021) which used a precise measuring arm to obtain the position and orientation of the scanner relative to the plants, and took several minutes to scan each plant.

In addition to the physical positioning of the sensor suite, the variability of environmental conditions within the dataset is of importance to both test the developed algorithm on the acquired data as well as to produce more general and robust learning algorithms trained on the data. Details on the environmental conditions in which the surveyed datasets were acquired can be found in Tables 1–2 and in the appendixes.

#### 4. Dataset applications

There are many applications that RGB-D data can be used for in agricultural operations, including harvesting fruits and vegetables (Bac et al., 2014), weeding (Harders et al., 2021) and pruning (Zahid et al., 2021). However, in each of these cases, the first goal of the RGB-D data is perception. For example, before weeding can occur, the weed must be both detected and reliably identified as a weed rather than a crop. These perception tasks are the focus of the applications evaluated on the RGB-D datasets in this review paper.

In many cases, RGB-D data are used for the same applications that RGB data can be used for, such as detection and identifying elements in the image or point clouds. However, since RGB data does not contain an absolute scale, it cannot naturally be used for measuring sizes, distances, or volumes. In contrast, RGB-D data provide the opportunity to extract geometric information, which allows estimation of fruit diameter or volume, leaf area, and other relevant agricultural data.

Since publicly available datasets enable training and testing of algorithms for agricultural operations, an important aspect of the dataset is the ground truth. For example, if a system is designed to measure fruit diameters using RGB-D data, it is important to have reliable width measurements for each fruit so that the results from the algorithm can be evaluated. Therefore, this section also discusses the important role of ground truth data (Section 4.4).

This section summarizes the background of visual detection and recognition in the agricultural domain using 2-D data (Section 4.1). It then presents the current state-of-the-art for visual detection and recognition in the agricultural domain using 3-D data (Section 4.2). Section 4.3 presents efficiency considerations, while Section 4.4 discusses the role of the ground truth data in a RGB-D agricultural dataset. Section 4.5 presents the use of 3-D data for geometric and morphological characterizations, and finally Section 4.6 discusses the use of the datasets for other inference and learning tasks, beyond detection and geometric characterizations.

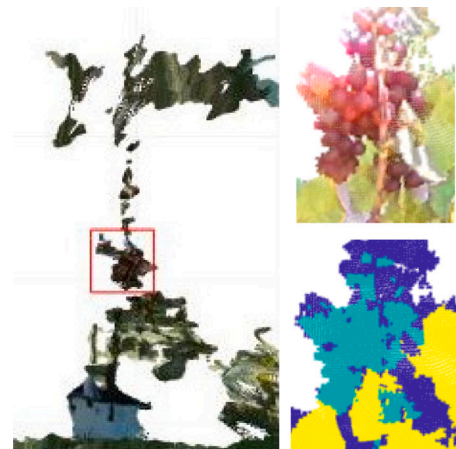


Fig. 4. Example of an annotated point cloud. Point cloud source: Kurtser et al. (2020b). Left: Example of a bounding box label. Right: Close up on a grape cluster (top) and example of a segmentation mask (bottom). Note: the low resolution of the closeup images is due to the sparsity of the point cloud.

#### 4.1. Visual detection and recognition in agricultural datasets

Object detection is an essential technique in agriculture, and is used for many essential applications including fruit and weed detection. It is the most common use of RGB data: according to a recent survey of public datasets for computer vision in agriculture (Lu and Young, 2020), all but one of the 34 datasets identified were used to perform detection tasks.<sup>2</sup> In Lu and Young (2020), two key applications of detection were identified: weed control (with 15 datasets) and fruit detection (with 10 datasets). The other 9 datasets had a variety of other detection tasks including flower detection, disease detection, and detection of harvest-induced damage.

For the 16 datasets included in this survey, all but one explicitly identified recognition and detection tasks as a key application of the dataset.<sup>3</sup> Visual recognition – the ability of an algorithm to recognize objects in an image – is one of the computer vision tasks that has developed and advanced most rapidly in the last decade. This development is due in part to the availability of large labeled datasets, which (in parallel with increases in computational power) spurred radical advances in deep learning techniques which increased the performance of recognition systems. Performance on the ImageNet image classification benchmark task is a clear example. From 2012 to 2014, image classification errors reduced by 2.4× (from 16.4% to 6.7%) (Russakovsky et al., 2015), and since then the classification errors have decreased yet further to better-than-human levels of accuracy.

Object detection is a refinement of image classification. In image classification, the classification system provides an output that identifies what object appears in an image. Object detection systems can not only identify that an object appears in an image, but locates where it appears, using a bounding box (see Fig. 4). While image classification is evaluated using a binary measure (that is, either the system correctly

<sup>2</sup> The Lu and Young (2020) computer vision dataset survey and this RGB-D dataset survey contain some overlaps. Seven of the datasets included in Lu and Young (2020) also contained 3-D data and thus are included in the survey. The overlapping datasets are: Sugarbeets (Chebroly et al., 2017), Ladybird (Bender et al., 2020), KFuji (Gené-Mola et al., 2019a,b), Fuji-SfM (Gené-Mola et al., 2020c,d), LFuji-air (Gené-Mola et al., 2020a,b), Apple Trees (Akbar et al., 2016), and 3-D Broccoli Kusumam (Kusumam et al., 2016, 2017). The non-detection dataset was the same as in this survey, specifically the Apple Trees dataset (Akbar et al., 2016).

<sup>3</sup> The exception was the Apple Trees dataset (Akbar et al., 2016) which focused on 3-D reconstruction.

labels the object within the image, or it does not), object detection also evaluates how accurately the bounding box matches the object. Object detection has shown similarly dramatic performance increases over recent years as image classification, with results on the benchmark COCO object detection dataset (Lin et al., 2014) becoming three times better between 2015 and 2018 (Szeliski, 2022).

In the past decade, the state-of-the-art in object detection has been based upon deep learning techniques (Zhao et al., 2019b; Jiao et al., 2019) and in particular Convolutional Neural Networks (CNNs) (LeCun et al., 2015). An early successful deep-learning object detection system was R-CNN (Girshick et al., 2014), with updated versions Fast R-CNN (Girshick, 2015) and Faster R-CNN (Ren et al., 2015) quickly following.

R-CNN and its variants are known as two-stage detectors. Two-stage object detection algorithms first determine regions in the image that might contain objects, and then take these object region proposals and classify the type of object in the region, based on the features extracted from the region. More recently, single-stage detectors have been proposed, which perform bounding box extraction and object classification in a single step. These more efficient single-stage detectors include the well-known framework YOLO (Redmon et al., 2016b) and its variants (Redmon and Farhadi, 2017, 2018), and SSD (Liu et al., 2016).

Semantic segmentation (Luo et al., 2023) is conceptually similar to object detection, but in semantic segmentation, the image is labeled at a pixel level instead of using bounding boxes. That is, every pixel in the image is assigned to a class (see Fig. 4 to compare bounding box and segmentation mask labels). As a result, it is more common that semantically segmented images contain multiple class labels. For example, soil, leaf, fruit, and stem might all be labeled separately in a semantically segmented image.

Semantic segmentation can be performed using very similar techniques to object detection. In fact, one of the most commonly used semantic segmentation systems, Mask R-CNN (He et al., 2017a), is simply an extension of the object detection framework Faster R-CNN, with an extra branch that predicts a pixel-level object segmentation mask, as well as the standard bounding box provided by an object detection network.

Historically, visual recognition and object detection has been performed using 2-D images (Szeliski, 2022). Even as RGB-D data becomes more accessible through the availability of more accurate and affordable sensors, there is still reliance in the agricultural domain on existing and widely available object detection algorithms based on 2-D data. Many of the surveyed papers that performed detection and segmentation used standard algorithms designed for 2-D data. For example, sweet pepper detection (Halstead et al., 2020) was demonstrated using both Faster R-CNN (using bounding boxes as ground truth) and Mask R-CNN (using pixel segmentation masks as ground truth). The Fuji-SfM dataset (Gené-Mola et al., 2020c,d), also demonstrated Mask R-CNN to perform instance segmentation of apples. On the S3CavVineyard dataset (Milella et al., 2019; Marani et al., 2021), the authors used image classification networks on image patches, rather than using an end-to-end object detection system, and compared a variety of networks including AlexNet (Krizhevsky et al., 2012), GoogLeNet (Szegedy et al., 2015), and two architectures based on VGG (Simonyan and Zisserman, 2015). In Blok et al. (2021b), broccoli head detection was demonstrated using ORCNN (Follmann et al., 2019), a semantic segmentation framework specifically designed for handling occlusion in an image.

However, the 3-D data can be integrated into traditionally 2-D detection algorithms. In Gené-Mola et al. (2019b) the Faster R-CNN input layer was extended from 3 color channels to 5 channels, adding range-corrected infra-red (IR) data and the re-projected depth data from the Kinect v2 sensor employed. Furthermore, the paper demonstrated that the 5-channel data out-performed the 3-channel RGB data in the evaluated detection task. Moreover, there are other methods to employ the 3-D data for detection and recognition, including an increasing

number of object detection and segmentation algorithms that take point clouds as input rather than 2-D images, and are specifically designed for RGB-D data. The next section discusses the role of 3-D data in visual detection and recognition algorithms in more detail, and in particular its application in the agricultural domain.

#### 4.2. Visual detection and recognition techniques for RGB-D data

Depth information can be used to improve 2-D detection and segmentation by manual pre-filtering of the point cloud based on domain knowledge to assist the segmentation from the 2-D space. For example, the point cloud can be pre-filtered to include only points a certain distance from the camera. Pre-filtering the data by distance information helps detection systems by generating a region of interest to perform detection and segmentation within it. These kinds of tailor-made pre-filtering procedures are effective, but also limit the ability of the algorithm to generalize, as they require fine tuning between datasets.

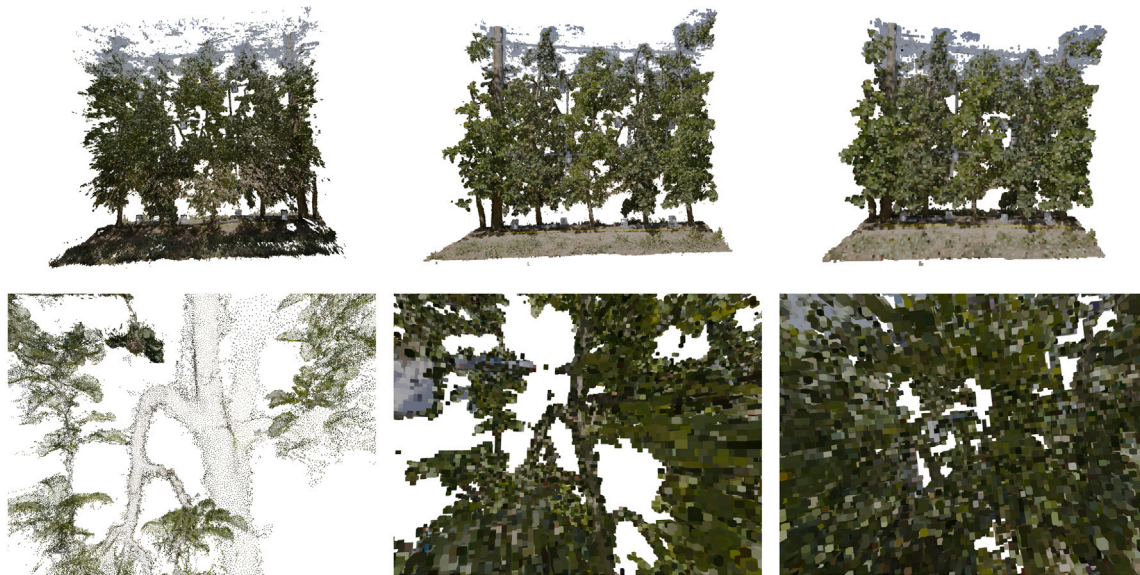
The development of algorithms that use the entire domain of visual and range data to perform both detection and physical size extraction via an end-to-end extraction of measures from acquired RGB-D data is crucial for the advancement of perception algorithms in agricultural robotics. One of the challenges in achieving these end-to-end detection pipelines for 3-D data is that typical deep learning convolutional architectures are designed to take simple and regular grid input formats, such as 2-D images or 3-D voxel grids (Charles et al., 2017). In contrast, point clouds are derived from range sensors and are therefore highly sparse compared to other visual data formats.

Point clouds can be discretized to 3-D space to serve as input to volumetric CNNs such as 3dshapenets (Wu et al., 2015). The discretized point clouds can be then used as part of a detection and recognition deep learning pipeline. However, volumetric representation of the 3-D space can be computationally expensive and introduces discretization artefacts. The effects of discretization artefacts can be seen in Figs. 5 and 6, which clearly demonstrate the impact of voxel size on the quality of the 3-D representation.

Discretization error is also intensified by the challenges of imaging in outdoor field conditions, compared to dense scanning performed in controlled environments. The increase in discretization artefacts outdoors can be seen by comparing Figs. 5 and 6. Fig. 5 shows example point clouds from the PFuji-size dataset (Gené-Mola et al., 2021b). These point clouds were generated using Structure-from-Motion algorithm applied to RGB images acquired in field conditions. Fig. 6 presents an example of a point cloud from the Pheno4D dataset (Schunck et al., 2021) acquired indoors using a laser triangulation scanner. The images show that the scans collected in controlled indoor conditions are less affected by artefacts and quantization error compared to outdoor data collected using a less accurate 3-D sensing technology.

An increasingly popular approach is to address a point cloud as a set of vectors, invariant to permutations of their members. Several deep learning architectures, such as the pioneering PointNet (Qi et al., 2017a) and its variants (such as PointNet++ (Qi et al., 2017b)), follow this approach. These networks can typically be computationally heavy, and demanding a lot of memory, compared to the state of the art 2-D segmentation networks, but recent developments in deep learning architectures capable of processing raw point cloud data, such as LatticeNet (Rosu et al., 2022), promise significant increase in performance.

Despite the growing literature of point cloud based methods, this survey identified only two of the datasets that demonstrated detection or segmentation via point cloud-based methods. Kurtser et al. (2020a) employed PointNet (Qi et al., 2017a) to segment grape clusters from raw point clouds, while Schunck et al. (2021) compared the performance of three deep learning architectures (PointNet, PointNet++ and LatticeNet). In these experiments, LatticeNet showed overall better semantic segmentation performance, especially for segmentation classes containing a smaller number of points.



**Fig. 5.** Example of a point cloud generated using Structure-from-Motion (RGB cameras) in field conditions, including discretization errors generated as a result of translation from point clouds to voxels. Images in left column — original point clouds as appearing in the dataset. Middle column — discretization into voxel size = 0.01 m. Right — discretization into voxel size = 0.03 m. Second row of images provides a close up view from the top row for easier view. The images show significant structural artifacts generated by quantization of erroneous points acquired. Increase in voxel size shows significant structural disruptions. Point cloud source: [Gené-Mola et al. \(2021b\)](#).



**Fig. 6.** Example of a point cloud acquired using a Perceptron Scan Works V5 in laboratory conditions, including discretization errors generated as a result of translation from point clouds to voxels. The point clouds are colored according to an annotation mask provided in the dataset. Left — original point clouds. Middle — discretization into voxel size = 0.01 m. Right — discretization into voxel size = 0.03 m. The images show less significant structural artifacts generated by quantization of both error points acquired compared to point clouds generated from images collected in field conditions (see [Fig. 5](#)). Increase in voxel size shows significant structural disruptions. Point cloud source: [Schunck et al. \(2021\)](#).

For PointNet or other point cloud based methods to be used within agricultural applications, it is essential to have the appropriate training data. In particular, this means that the ground truth classification needs to label the points in the 3-D point clouds, rather than simply providing bounding boxes or pixel labels in 2-D image space. As well as the two datasets noted in this survey as evaluating point cloud-based methods ([Kurtser et al., 2020a](#); [Schunck et al., 2021](#)), two other datasets also provided labeled 3-D point clouds: one for broccoli heads ([Kusumam et al., 2017, 2016](#)), and one for apples ([Gené-Mola et al., 2020c,d](#)).

#### 4.3. Performance and efficiency of detection and recognition in agricultural datasets

There has been a rapid increase in efficiency of visual detection and recognition algorithms in recent years. In the 2-D image domain, modern object detection algorithms are consistently decreasing inference time while maintaining high quality detection results ([Wang et al., 2022](#)). Similarly, the development of point cloud-based systems such as PointNet ([Qi et al., 2017a](#)) represented a significant efficiency increase over previous deep learning architectures for 3-D data classification ([Qi et al., 2016](#); [Su et al., 2015](#)), due to PointNet's linear complexity in time and space. While 3-D detection and recognition approaches are undoubtedly more computationally intensive than 2-D frameworks ([Wang et al., 2022](#); [Rosu et al., 2022](#)), there has also been

notable improvements in performance of the 3-D systems over recent years ([Rosu et al., 2022](#)).

In domains such as autonomous driving, detection algorithms must, for safety reasons, be both highly efficient and highly accurate ([Szeliski, 2022](#)). In contrast, object detection and recognition in agricultural applications is focused on high quality detection and recognition performance rather than highly efficient approaches. While cycle times are of high importance when determining the commercial feasibility of an agricultural robot, they are most often hindered by localization accuracy, false positive detections, high uncertainty and mechanical limitation of manipulation in dense environments rather than the processing time and power required for image analysis ([Bac et al., 2014](#); [Bechar and Vigneault, 2017](#)). In the surveyed papers, the focus was on achieving reliable detection rather than highly efficient detection. The slower two-stage object detectors such as Faster R-CNN ([Ren et al., 2015](#)) were used by some agricultural applications ([Halstead et al., 2020](#); [Gené-Mola et al., 2020c,d](#)), while the more efficient one-stage object detectors such as YOLO ([Redmon et al., 2016a](#)) were used by others ([Tian et al., 2019b](#)).

In the agricultural domain, where the amount of available training data is still relatively limited, it is to be expected that currently, the focus will be on acquiring larger datasets for training. Nevertheless, some agricultural robotics applications, specifically those relying on operations while in motion, are somewhat hindered already now by the inefficiency of the detection algorithms. Examples of such robots

include weeding robots performing mechanical weeding or precision spraying while the robot traverses the field (Li et al., 2022a). The authors expect more and more applications to overcome the challenges of accurate detection in complex environments, and that over time, as the research field matures and the performance of the underlying algorithms becomes more reliable, the efficiency of the resulting algorithms will become of greater interest.

#### 4.4. Ground truth labeling for detection tasks in RGB-D agricultural datasets

The ground truth for detection algorithms is either a bounding box or a segmentation mask which provides semantic labels for the image at the pixel level (see Fig. 4). This ground truth information is typically manually created and its creation can be highly labor-intensive.

Lack of or false annotations are major sources of error and a limiting factor in producing large quantities of relevant data. Acquiring accurate annotations is challenging for several reasons. First, some annotations requires expert knowledge. For example, in the Pheno4D dataset (Schunck et al., 2021) the authors claim the segmentation of tomato leaves point clouds is a simple task due to clear separation between the leaves and the stem, while for maize leaves the separation is not as apparent. As a result, the authors had to employ the *LeafColor* method for staging corn plants where a special point in the leaf is identified as the barrier between the leaf and the stem. Second, some annotations are simply harder and more prone to human errors. For example, annotation of weed in mature fields such as done in 2-D images in Sørensen et al. (2017) is a highly complex task due to the high overlap between the weed and the plants, and the low color contrast between them. Finally, annotations of other color spectra (i.e. hyperspectral images), pose a complex task for the human annotator. Bender et al. (2020) coped with this challenge by generating 2-D reconstruction images from the acquired 3-D hypercubes to create images of good contrast used for annotation.

As a result, the fact that the authors of the cited papers provided labels to the released data is highly valuable. Manually labeled bounding boxes for crop detection evaluation are provided for apples (Gené-Mola et al., 2019a,b, 2020a,b), sweet peppers (Halstead et al., 2020; Arad et al., 2019), and broccoli and cauliflower (Bender et al., 2020). In Blok et al. (2021a,b), broccoli heads were labeled, not with bounding boxes, but with a polygonal mask for the visible part of each head, and a circle mask that estimated the size of the whole head, including the areas that were occluded by leaves.

While bounding boxes typically offer detection for a single class of objects (e.g. apples), segmentation masks can provide multiple class labels. Segmentation masks by Chebroly et al. (2017), include labels for sugar beets and nine different types of weed, while in Milella et al. (2019), Marani et al. (2021), segmentation ground truth is provided for 5 different classes (grape bunch, pole, wood, leaves, background). Bender et al. (2020) also evaluated semantic segmentation for both broccoli, soil, and weed classes. However, data from a hyperspectral camera was used and therefore it is not clear whether this could be directly applied to the stereo imagery.

Attempts to automatically generate a segmentation mask from an annotated bounding box have been reported in Kurtser et al. (2020a,b) where the authors perform unsupervised K-means and Euclidean segmentation of grape bunches from regions of interest annotated manually. This method reduces significantly the manual labor involved in generating a segmentation mask, but reduces the accuracy of the annotation compared to a human annotator.

These bounding box and semantic segmentation labels are generated on the 2-D color images, but it is also possible to label the points in the 3-D point clouds. This is done for broccoli heads (Kusumam et al., 2017, 2016), grape clusters (Kurtser et al., 2020a,b), and apples (Gené-Mola et al., 2020c,d). In the Pheno4D (Schunck et al., 2021) dataset, the point clouds from maize and tomato plants are manually labeled for soil, stem, and leaf classes. Furthermore, in the Pheno4D dataset the

points on each leaf were uniquely labeled over the measurement period, so instance segmentation (i.e. identifying leaves with different labels) could also be performed. These labeled point clouds can be used as input to neural network architectures such as PointNet (Charles et al., 2017) that are specifically designed for point cloud input (Schunck et al., 2021; Kurtser et al., 2020a).

#### 4.5. Geometric characterizations

One important area in which RGB-D data has a significant advantage over RGB data is measuring and predicting geometric properties of plants. Standard 2-D image data have no inherent notion of scale, while depth data provide information about size, distances, areas, and volumes. This is vital in agricultural operations where evaluation of size and volume of crops is valuable information often used by farmers and growers in adaptation of the growing protocol to environmental stressors.

The 3-D data play a vital role in tasks that require measuring geometric properties of the plant. However, for the 3-D data to be evaluated and tested, the datasets should contain ground truth information about the true geometric character of the plants, in the form of manual measurements. The manual ground truth measurements that are available in these datasets include estimation of canopy width, canopy cross-section area, and leaf area (Gené-Mola et al., 2020a,b), fruit diameters (Gené-Mola et al., 2021a,b; Kurtser et al., 2020b), volume estimation of grape clusters (Milella et al., 2019; Kurtser et al., 2020b), as well as leaf area (Schunck et al., 2021). In Kusumam et al. (2017, 2016) additional measurements of head diameter (using calipers) and weight measurements were performed on the broccoli heads, and Bender et al. (2020) also included criteria such as brassica height, width, and weight.<sup>4</sup>

The geometric characterization is fundamentally linked to the detection tasks performed using the RGB-D data: if the system is intended to measure the size of broccoli heads, for example, it also needs to detect and recognize the broccoli head in the image in order to estimate its dimensions. In Blok et al. (2021b) a segmentation mask was generated using standard 3-channel RGB images to detect broccoli heads in an image. The depth mask was then used to estimate the size of the broccoli head in a secondary step.

In Akbar et al. (2016), the topic of interest was the pruning of dormant apple trees. To facilitate this, the dataset included so-called ground truth images of the tree, which were captured using a color (RGB) camera rather than a depth camera, with the branches individually labeled in each image. The ground truth data also included the diameters of the branches and the distances between branches.

#### 4.6. Other inference tasks

The majority of surveyed datasets focused on object detection tasks, with geometric characterizations also being frequently included. However, there are other tasks that can be inferred from RGB and RGB-D data, and a number of datasets included a broader range of ground truth data, from which learning systems could be trained and evaluated.

For the sweet pepper datasets, where the fruit can have a variety of colors (such as green, yellow, red, mixed, or black), additional color information about each fruit was included (Arad et al., 2019; Halstead et al., 2020). In Halstead et al. (2020), different sweet pepper cultivars were grown, and this information was also included.

In a number of datasets, information about canopy cover was included (Gené-Mola et al., 2020a,b; Milella et al., 2019; Marani et al., 2021). Since the canopy characterizations are geometric and include

<sup>4</sup> While weight is not directly measurable using 3-D imagery, the ability to estimate diameter, area, and volume of various crops can then be used in some cases to infer approximate information about weight.

measures of height, width, and cross-sectional areas, the 3-D data are particularly necessary to perform these calculations.

Other ground truth measurements included measuring greenness (Bender et al., 2020), and plant stress caused by weed pressure, nitrogen surplus or deficit, or lack of water (Khanna et al., 2019). While the 3-D data are not absolutely necessary to calculate these values, an interesting result was that the 3-D data provided improved performance over just RGB images for the plant stress inference tasks (Khanna et al., 2019). In these experiments, the best results were achieved with a combination of RGB, 3-D data, hyperspectral imagery, and a temporal component.

## 5. Discussion

As can be seen from the reviewed publications, while the number of datasets has rapidly increased over the last few years, publicly available datasets for agricultural applications that provide RGB-D or even depth data are still relatively rare. As a result, the ability to develop and improve relevant algorithms is limited to those groups that have the resources to collect datasets themselves. Furthermore, there is limited ability to perform benchmarking of algorithms when the results are only presented on a closed dataset, making a fair comparison of algorithms impossible.

The effect of required resources on the acquisitions can also be seen in the choice to use commercial grade cameras in data acquisition. The choice can be explained by the fact that a high accuracy RGB-D sensor can be almost 10 folds more expensive than the commercial grade one (Fu et al., 2020). With price being an important consideration in the choice of hardware when developing site-specific agricultural robots (Bac et al., 2014), the high-end sensors are usually not considered feasible. It is also reasonable to assume that some of the high-end sensors are not accessible to some research groups. With factory calibrated commercial RGB-D sensors providing increasingly better accuracy and performance (Vit and Shani, 2018) and with sensor costs decreasing, one can expect this obstacle to become less critical with time (Fu et al., 2020).

However, in some cases adaptation to the acquisition protocol needed to be made in order to provide the needed accuracy for the specific application. For example, Kusumam et al. (2016, 2017) and Blok et al. (2021a,b) constructed special enclosures to block direct sunlight, and Gené-Mola et al. (2019a,b) collected the point clouds only at night with artificial lighting.

There are still only limited applications explored in these datasets: the majority of the datasets do detection or segmentation as a main goal. Since detection is generally considered a computer vision problem, this particular application does not fully exploit the value of the depth data within these datasets. However, Gené-Mola et al. (2019b) demonstrates that extending the learning model from a 3-channel RGB image to a 5-channel image including range-corrected intensity data and depth data can improve detection capabilities. The authors see this inclusion of depth data into 2-D detection algorithms as an interesting future direction of research.

At the same time, computationally feasible detection and segmentation frameworks that operate directly on the 3-D point cloud are now becoming a reality (Qi et al., 2017a; Rosu et al., 2022). While detection methods that depend on 2-D data are more mature and more efficient computationally, the authors foresee a role for both approaches in agricultural applications, as 2-D and 3-D methods can complement each other depending on the needs of the particular application.

Furthermore, there are other applications emerging that do explicitly depend on depth data. In particular, many of the datasets include size measurements, such as size of the fruit or vegetable, which is vital information for the agricultural domain. Having accurate 3-D information is a key input for providing scale to these measurements that 2-D images often cannot provide. It is possible that the range data can also provide indirect measurements about values such as weight,

which normally can only be accurately evaluated manually, and after harvesting. Providing manual ground truth for these values can be a labor-intensive process that often requires expert knowledge, and only some research groups have the capability to achieve it. Therefore, having publicly available datasets that provide this information can be an essential ingredient of a vibrant research landscape.

The types of plants for which these datasets are available is still limited, demonstrating how small and challenging the research landscape is, and the difficulty of creating high-quality agricultural datasets. However, having even these datasets available allows broader opportunities for researchers to contribute and will hopefully spur on innovation and encourage broadening the range of plants that may be available. Furthermore, the experience gained from these datasets can be indirectly used to guide future research into a wider variety of crops and plants.

A more daring question is whether any algorithms trained or evaluated on these datasets can be used directly for transfer learning between different crops. A restricted version of this problem was explored by Halstead et al. (2020) where generalizability of detection algorithms was tested across different sweet pepper cultivars, and the potential for transfer learning between crop types has been demonstrated for sugar beets, carrots, and onions (Bosilj et al., 2020). These results are particularly notable as other research (Autz et al., 2022) has suggested that “naïve” transfer learning from a non-agricultural dataset such as COCO (Lin et al., 2014) to an agricultural dataset (Chebroul et al., 2017) is not always beneficial.

The RGB-D datasets for each specific plant type have been published by only one or two research groups. However, there are signs that RGB-D dataset collection may be maturing, albeit slowly. The most recently published RGB-D broccoli dataset (Blok et al., 2021a,b) explicitly extended the existing broccoli datasets (Kusumam et al., 2016, 2017; Bender et al., 2020) by systematically making the broccoli heads in the images more occluded. Similarly a recent Fuji apple dataset (Gené-Mola et al., 2021a,b) extends on the group’s previous datasets (Gené-Mola et al., 2020c,d) by introducing occlusions and including both apple diameter data and different apples at different maturity levels. It can be expected that over time similar developments in creating progressively more challenging datasets will be seen for other crops, despite the small number of research groups currently capable of generating the datasets required.

## 6. Conclusions

There is great potential for robots and autonomous systems supplied with visual and range data from on-board RGB-D sensors to assist in a wide range of agricultural operations. However, to develop methods that depend on this visual and range data requires access to datasets containing quality data that adequately represents the field conditions in which agricultural robots must operate. Therefore, publicly available datasets are essential to the development of this research field. Since high quality data is scarce and often hard to obtain, there has been a major bottleneck in the development of machine learning algorithms that rely heavily on high quality data.

As a result, each acquired dataset has potential to provide research groups with the ability to build upon previous work, compare performance and advance the state of the art. As presented in this dataset review paper, there are still only a limited number of available datasets. However 16 public datasets covering a wide range of crops were identified, and a certain critical mass has accumulated, with new datasets extending older datasets via a wider range of scenarios (Blok et al., 2021a,b; Kusumam et al., 2016, 2017; Bender et al., 2020; Gené-Mola et al., 2020c,d, 2021a,b).

The increase of RGB-D datasets in agricultural operations has been spurred by the accessibility of increasingly capable and affordable sensors. At the same time, there is also rapid development in the underlying available algorithms, particularly in the field of visual detection

and segmentation. The object detection and segmentation techniques based on 2-D data such as Faster R-CNN, Mask R-CNN, and YOLO are increasingly reliable and efficient, while the equivalent 3-D methods such as PointNet and LatticeNet are comparatively recent developments but showing a great deal of promise.

The release of RGB-D data in the agricultural domain collected in outdoor conditions has only recently begun, but the existence of the publicly available datasets surveyed in this paper can already provide valuable information for use in precision agriculture and automation of agricultural procedures. These datasets can provide research groups with the ability to develop state-of-the-art agricultural algorithms, as well as compare the performance of different systems, without the effort for data collection and preparation.

### Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

### Data availability

No data was used for the research described in the article.

## Appendix A

This section provides a summary of each of the datasets included in this survey.

### A.1. Pheno4D (Schunck et al., 2021)

This dataset consists of scans of maize and tomato plants from sprouting (early growth). It tracks 7 maize plants for 12 days, generating 84 point clouds of which 49 have ground truth labels, and 7 tomato plants for 20 days: generating 140 point clouds of which 77 have ground truth labels. The point clouds are manually labeled with ‘soil’, ‘stem’, and ‘leaf’ where each leaf is uniquely labeled over the measurement period. The dataset consists of approximately 440 million 3-D points, of which 260 million are labeled.

The data were acquired using a Perceptron Scan Works V5 laser triangulation scanner (with 660 nm wavelength), using a high precision measuring arm (ROMER Infinite 2.0) with an accuracy of measuring the tip of the arm of 45  $\mu\text{m}$ . From this the 2-D scan profiles could be registered into a 3-D point cloud (with no color). The scanning was done to achieve maximal coverage and plant stillness, although some occlusion remained.

With the maize plants, it is difficult to differentiate between leaves and stems. Therefore two methods of manual segmentation were performed — the Leaf Collar Method (a common maize staging method) and the Leaf Tip Method, and both ground truth results are included.

The paper demonstrates numerous applications on the dataset, including semantic segmentation, instance segmentation (i.e. identifying leaves with different labels), spatio-temporal registration, surface reconstruction, and phenotyping (including calculating leaf area and leaf length).

### A.2. Broccoli 3D (Kusumam et al., 2016, 2017)

This dataset consists of scans of broccoli plants, with the application goal of an autonomous selective broccoli harvester. The broccoli was imaged during four data collection sessions: three in Lincolnshire, UK during the early harvesting season and one in Murcia, Spain, during the late harvesting season. The data consisted of a training set of 32 broccoli point clouds and 324 “other” point clouds (both of which are segmented out). The test set consists of point cloud frames and 1619 annotated instances of broccoli heads. Furthermore, the Spain dataset

had 100 broccoli heads measured manually for size (using calipers) and weight.

The data was acquired with a Kinect V2 sensor. The sensor was fixed inside a specially constructed enclosure, which was mounted on the front of a tractor. The enclosure blocked direct sunlight and includes an “umbrella” during rainy conditions, and artificial lighting source, comprising strip LED lighting, for increase light homogeneity and allow night time acquisition. The sensor is mounted at range of heights from ground 125 cm–140 cm.

The papers demonstrated a number of applications, including size estimation (for maturity evaluation), localization (for manipulation of the robotic arm) and creation of a 3-D map with broccoli locations.

### A.3. KFuji RGB-DS (Gené-Mola et al., 2019a,b)

This paper consists of 967 multi-modal images of Fuji apples on trees. Contains color images, depth and range-corrected IR intensity. The apples were scanned 3 weeks prior to harvesting. The data was acquired with two Kinect V2s. The scanning was carried out at night with artificial lighting.

The data was processed with 3-D point clouds projected onto images and manual annotated fruit locations were provided, with 12839 apples annotated in total. The papers presented experiments using fruit detection (based on Faster R-CNN (Ren et al., 2015)).

### A.4. LFuji-air (Gené-Mola et al., 2020a,b)

This paper consist of point clouds of 11 Fuji apple trees using LiDAR imagery. The dataset consists of 8 scans per tree, from 2 different heights, 2 different sides of the tree and also 2 different air flow conditions — fan off and fan on, to see if this helped with occlusions. The dataset included 1353 manually annotated apples (1444 were counted in the orchard but the remaining apples could not be identified in the point clouds).

The data was captured with a Puck VLP-16 Velodyne and a RTK-GNSS system, which were pulled by a tractor parallel to the trees, and used the fact that apples have higher IR reflectance than the background to identify them within the scans.

The papers demonstrated applications including fruit detection, yield prediction, and geometric characterization (canopy height, width estimation, cross-section area, leaf area).

### A.5. S3CavVineyardDataset (Milella et al., 2019; Marani et al., 2021)

This dataset consists of images of grape cluster on grapevines. The images were captured using an Intel RealSense R200 which was mounted on a mobile Niko caterpillar platform. However, only the RGB images are currently publicly available<sup>5</sup>

The data consists of 500 RGB images of grapevines and was collected over 2 days of filming with camera facing the rows of grapes at 0.75–1 m distance at 5 fps and mobile platform moving on average at 1.5 m/s. The data was processed using the authors’ own registration algorithms from the IR inputs of the camera for registration in a better range of distances.

For ground truth, 84 images were picked for manual annotation in 5 classes (Bunch, pole, wood, leaves, background). Applications include canopy volume estimation, and detection/segmentation of grape bunches in in-field images.

<sup>5</sup> <https://github.com/ispstiima/S3CavVineyardDataset> — last accessed August 2023.

#### A.6. Ladybird Cobbitty 2017 brassica dataset (Bender et al., 2020)

This dataset consists of weekly scans of cauliflower and broccoli, covering a 10-week period from transplant to harvest. The data scanned 4 beds with approximately 144 Brassica per bed (50 cm apart). These beds were from different treatment zones (with low irrigation and normal irrigation, and different fertilizer treatments).

This dataset consists of color, thermal, and hyperspectral data. It does not contain any point clouds, but stereo images were taken of the crops which could be used to extract 3-D information. The dataset also contains ground truth (manual) measures as well as in-situ weather station and soil sensor data. Manual measurements were taken the day after the autonomous scan. The ground truth measurements included greenness (using a SPAD chlorophyll meter), height and width. From week 4, destructive measures – fresh weight, dry weight, and relative water content – were taken on 4 plants from each bed and each species. The weather station measured temperature, wind speed, humidity, pressure, and rainfall. The soil sensors measured temperature, electrical conductivity, dielectric and volumetric water content at 16 locations, measuring 10 cm and 30 cm deep at each location.

The data was captured with a Ladybird robot platform that autonomously followed GPS waypoints in a farm map and localized using RTK-GPS. It had an imaging canopy to block direct sunlight. The imagery data consists of stereo images from Grasshopper cameras, approximate 3–4 images per plant with image overlap of 70%–85%, a thermal camera, and a hyperspectral camera. The paper demonstrated examples of applications that include semantic segmentation (from the hyperspectral images), and object detection.

#### A.7. Eschikon plant stress phenotyping (Khanna et al., 2019)

This dataset consists of sugar beets grown in a greenhouse chamber. The images were taken top-down onto a growing box. The data was captured with an Intel RealSense ZR300 (RGB-D stereo IR), Ximea MQ022HG-IM-SM5X5-NIR hyperspectral camera.

The plants were imaged bi-weekly for 2 months producing color, IR stereo, and hyperspectral data. A total of 1984 images were captured (31 boxes captured on 16 dates and consisting of 4 images (RGB, 2IR, 1HS)). The paper demonstrated the application of identifying water, nitrogen and weed stress, using as ground truth treatment and environment parameters that were included with the image data.

#### A.8. Sweet pepper detection dataset (Halstead et al., 2020)

This dataset consists of scans of sweet peppers. The dataset is in three parts, two from Australia (QHDF, QHDP) and one from Germany (BUP). One part each from a field in sunlight (QHDF), polytunnel with some sun protection (QHDP), and glass house (BUP). Different cameras with different resolutions were used: Intel RealSense 200 @ 640 × 480 (QHDF/QHDP) and Intel RealSense 435i @ 1280 × 720 px (BUP).

The data was manually annotated with locations of the sweet peppers. The data consists of 1583 images (QHDF) with 5774 ground truth labels, 687 images (QHDP) with 3741 ground truth labels, and 286 images (BUP) with 3724 ground truth labels. Additional ground truth information included the color of the fruit (green, red, mixed, and black), and the cultivar of the fruit (four different cultivars were imaged).

The paper demonstrates detection using Faster R-CNN (Ren et al., 2015) and Mask R-CNN (He et al., 2017a) and focuses on the potential for generalizability of detection of sweet peppers across different locations and cultivars.

#### A.9. SWEEPER — Sweet pepper dataset (Arad et al., 2019)

This dataset consists of images of sweet peppers under complex illumination conditions. The data was captured in a greenhouse and contained 156 scenes with 468 images containing a total of 344 manually labeled yellow sweet peppers in 2 illumination conditions — with flash and without. A Fotonic F80 camera (a hybrid RGB-ToF depth camera) was used, with an external illumination trigger and illumination rig (Effilux brand LED strips). The dataset is captured with the goal of evaluation of a Flash-no-Flash acquisition technique for stabilization of ambient light.

#### A.10. Grapes 3D (Kurtser et al., 2020a,b)

This dataset consists of grape clusters captured in a vineyard. The dataset contained two parts — the first part was captured outdoors in controlled conditions, and the second part was captured outdoors in a vineyard in different conditions (including different viewpoint angles, and with/without a contrasting background). The outdoor part consisted of 17 grape clusters on 10 grapevines (1–3 clusters on each vine), and the vineyard part consists of 5 grapevines with 8–19 grape clusters on each.

The data was captured with an Intel RealSense D435 RGB-D camera mounted on a mobile robot platform. The outdoor part contained 6 videos of 30–90 s each at 30 fps and vineyard part contained 4 videos of 45–80 s each. The ground truth of physical size of grapes was provided for the outdoor dataset. The paper demonstrated detection for harvesting and grape cluster sizing for yield estimation.

#### A.11. Sugarbeets2016 (Chebrolu et al., 2017)

This dataset consists of images of sugar beets in fields, captured using the BoniRob field robot. Its sensor suite consists of an Kinect V2 (RGB-D), JAI AD-130GE camera (RGB), a Velodyne VLP16 Puck (3-D LiDAR), a NipponSignal FX8 (laser range sensor), GPS RTK and Global GPS. The data was collected 2 or 3 days each week, with a total of 30 days of recordings, capturing different growth stages of the crop, as well as different weather and soil conditions. Each recording was of between 4 and 8 crop rows measuring 400 m in length.

The dataset contains labeled ground truth segmentation masks for that encode sugar beets, nine different types of weed, and non-vegetative parts.

#### A.12. Apple trees (Akbar et al., 2016)

This dataset consists of images of 9 apple trees in the dormant phase. 3 trees were imaged indoors and 6 outdoors in the USA (Indiana and Pennsylvania) using a Kinect V2 sensor. The intended application of this dataset is development of 3-D reconstruction algorithms to assist the pruning of apple trees.

The data consists of depth images and color images from the Kinect, and ground truth images. The ground truth images are color images with number labels on them for each branch, where images are captured from different angles using a regular (non-depth) camera. Diameters of branches and distances between branches were also provided. The paper includes a detailed description of the denoising and preprocessing techniques performed on the images.

#### A.13. Broccoli heads (Blok et al., 2021a,b)

This data consists of images of broccoli heads under varying degrees of occlusion. Two experiments were conducted. The first experiment contains 947 RGB-D images (approximately 4–10 images per head for 122 broccoli heads). Occluding leaves were removed and the broccoli heads were re-imaged. Then the heads were measured with a ruler. RGB color camera (IDS UI-5280FA-C-HQ) and monochrome stereo-vision

camera (IDS Ensenso N35) were used, and an enclosed box was used for uniform illumination.

The second experiment used an Intel RealSense D435 in natural light but with an umbrella to diffuse light. 250 occluded broccoli heads were imaged. First, 5–10 frames per broccoli head with natural occlusion were captured and then different occlusions were created by cutting leaves from neighboring plants and placing over to create an artificial occlusion, and also with no occlusion (1613 RGB-D images total). The size of the broccoli heads were also measured with a ruler.

The images were manually annotated with a polygon for the visible head and a circle for the estimated broccoli head (not just the visible part). The paper demonstrates segmentation and size estimation from the images.

#### A.14. *Fuji-SfM (Gené-Mola et al., 2020c,d)*

This data consists of images of 11 scanned apple trees, with point clouds generated from structure from motion (SfM). The images were captured in a single experiment in morning and afternoon conditions.

The data was captured with an EOS 60D DSLR Cannon RGB camera. Images were taken freehand following a scanning pattern of 5–6 images from imaging locations at 0.2 m apart. A multi-view structure-from-motion photogrammetry based on bundle adjustment was applied to generate the 3-D point cloud.

The dataset contains 582 raw images, and 1455 annotations (with a 3-D box annotation in the reconstructed dataset). The paper demonstrates apple detection, segmentation and localization algorithms.

#### A.15. *PFuji-size dataset (Gené-Mola et al., 2021a,b)*

This dataset consists of images of apples based on structure from motion (SfM). Three experiments were conducted. The first two experiments were identical with 6 scanned apple trees in 2 different maturity stages (3 trees in each maturity stage). The third experiment was an indoor lab experiment on a turn table.

The data was captured with an EOS 60D DSLR Cannon RGB camera. The images were taken freehand following a scanning pattern of 5–6 images from imaging locations at 0.2 m apart. A total of 168 and 180 images were acquired respectively from the eastern and western sides of the tree row captured in experiment 1, and a total of 228 and 280 images respectively from the eastern and western sides of the tree row captured in experiment 2. 4500 images were captured for the indoor case (25 apples, 180 images per apple). Apple segmented point clouds are also provided.

The paper demonstrates apple detection and size estimation. The data was manually annotated with 615 apple diameter annotations for experiments 1 and 2 and 25 apple diameter annotations for experiment 3.

#### A.16. *WE 3DS (Kitzler et al., 2023)*

This dataset consists of image pairs of 17 young field crops for stereo vision. Experiments were conducted at 25 measurement dates, and includes manually collected metadata (i.e., camera mounting height, light conditions and wind), as well as geo-locations information and camera parameters.

The acquisition was performed using a pair of industrial RGB cameras (XIMEA MC023CG-SY) mounted on a manually driven cart at fixed position in a top-down viewpoint facing the ground. The cart is also equipped with a GPS sensor.

Collected images are manually annotated (2568 labeled images) using a segmentation mask classifying into 17 classes. Distance maps are generated using the stereo information. The segmentation masks are used for evaluation of accuracy using ESANet as a benchmark.

## References

- Akbar, S.A., Chattopadhyay, S., Elfiky, N.M., Kak, A., 2016. A novel benchmark RGBD dataset for dormant apple trees and its application to automatic pruning. *CVPRW*, pp. 347–354.
- Arad, B., Balendonck, J., Barth, R., Ben-Shahar, O., Edan, Y., Hellström, T., Hemming, J., Kurtser, P., Ringdahl, O., Tiele, T., et al., 2020. Development of a sweet pepper harvesting robot. *J. Field Robotics* 37 (6), 1027–1039.
- Arad, B., Kurtser, P., Barnea, E., Harel, B., Edan, Y., Ben-Shahar, O., 2019. Controlled lighting and illumination-independent target detection for real-time cost-efficient applications. the case study of sweet pepper robotic harvesting. *Sensors* 19 (6), 1390.
- Araus, J.L., Cairns, J.E., 2014. Field high-throughput phenotyping: the new crop breeding frontier. *Trends Plant Sci.* 19 (1), 52–61.
- Autz, J., Mishra, S.K., Herrmann, L., Hertzberg, J., 2022. The pitfalls of transfer learning in computer vision for agriculture. In: Gandorfer, M., Hoffmann, C., El Benni, N., Cockburn, M., Anken, T., Floto, H. (Eds.), 42. GIL-Jahrestagung, Künstliche Intelligenz in der Agrar- und Ernährungswirtschaft. Gesellschaft für Informatik e.V., Bonn, pp. 51–56.
- Bac, C.W., van Henten, E.J., Hemming, J., Edan, Y., 2014. Harvesting robots for high-value crops: State-of-the-art review and challenges ahead. *J. Field Robotics* 31 (6), 888–911.
- Barbole, D.K., Jadhav, P.M., 2023. GrapesNet: Indian RGB & RGB-D vineyard image datasets for deep learning applications. *Data Brief* 48, 109100.
- Barth, R., Hemming, J., van Henten, E.J., 2016. Design of an eye-in-hand sensing and servo control framework for harvesting robotics in dense vegetation. *Biosyst. Eng.* 146, 71–84.
- Bechar, A., Vigneault, C., 2016. Agricultural robots for field operations: Concepts and components. *Biosyst. Eng.* 149, 94–111.
- Bechar, A., Vigneault, C., 2017. Agricultural robots for field operations. Part 2: Operations and systems. *Biosyst. Eng.* 153, 110–128.
- Bender, A., Whelan, B., Sukkarieh, S., 2020. A high-resolution, multimodal data set for agricultural robotics: A ladybird's-eye view of Brassica. *J. Field Robotics* 37 (1), 73–96.
- Blok, P.M., van Henten, E.J., van Evert, F.K., Kootstra, G., 2021a. Data Underlying the Publication: Image-Based Size Estimation of Broccoli Heads Under Varying Degrees of Occlusion. 4TU.ResearchData.
- Blok, P.M., van Henten, E.J., van Evert, F.K., Kootstra, G., 2021b. Image-based size estimation of broccoli heads under varying degrees of occlusion. *Biosyst. Eng.* 208, 213–233.
- Bosilj, P., Aptoula, E., Duckett, T., Cielniak, G., 2020. Transfer learning between crop types for semantic segmentation of crops versus weeds in precision agriculture. *J. Field Robot.* 37 (1).
- Charles, R.Q., Su, H., Kaichun, M., Guibas, L.J., 2017. PointNet: Deep learning on point sets for 3D classification and segmentation. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition. *CVPR*, pp. 77–85.
- Chebrou, N., Lottes, P., Schaefer, A., Winterhalter, W., Burgard, W., Stachniss, C., 2017. Agricultural robot dataset for plant classification, localization and mapping on sugar beet fields. *Int. J. Robot. Res.* 36 (10), 1045–1052.
- Chi, M., Plaza, A., Benediktsson, J.A., Sun, Z., Shen, J., Zhu, Y., 2016. Big data for remote sensing: Challenges and opportunities. *Proc. IEEE* 104 (11), 2207–2219.
- Cordts, M., Omran, M., Ramos, S., Scharwächter, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B., 2015. The cityscapes dataset. In: *CVPR Workshop on the Future of Datasets in Vision*, Vol. 2.
- Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., Fei-Fei, L., 2009. Imagenet: a large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition. *IEEE*, pp. 248–255.
- Durand-Petitville, A., Sadowski, D., Vougioukas, S., 2018. A strawberry database: Geometric properties, images and 3D scans.
- Dutagaci, H., Rasti, P., Galopin, G., Rousseau, D., 2020. ROSE-X: an annotated data set for evaluation of 3D plant organ segmentation methods. *Plant Methods* 16 (1), 1–14.
- European Organization For Nuclear Research, OpenAIRE, 2013. Zenodo. CERN.
- Follmann, P., König, R., Härtinger, P., Klostermann, M., Böttger, T., 2019. Learning to see the invisible: end-to-end trainable amodal instance segmentation. In: *IEEE Winter Conference on Applications of Computer Vision. WACV2019, Waikoloa Village, HI, USA, January 7–11, 2019*, IEEE, pp. 1328–1336.
- Fu, L., Gao, F., Wu, J., Li, R., Karkee, M., Zhang, Q., 2020. Application of consumer RGB-D cameras for fruit detection and localization in field: A critical review. *Comput. Electron. Agric.* 177, 105687.
- Geiger, A., Lenz, P., Stiller, C., Urtasun, R., 2013. Vision meets robotics: The KITTI dataset. *Int. J. Robot. Res.* 32 (11), 1231–1237.
- Gené-Mola, J., Gregorio, E., Auat Cheein, F., Guevara, J., Llorens, J., Sanz-Cortiella, R., Escolà, A., Rosell-Polo, J.R., 2020a. Fruit detection, yield prediction and canopy geometric characterization using LiDAR with forced air flow. *Comput. Electron. Agric.* 168, 105121.
- Gené-Mola, J., Gregorio, E., Auat Cheein, F., Guevara, J., Llorens, J., Sanz-Cortiella, R., Escolà, A., Rosell-Polo, J.R., 2020b. lFuji-air dataset: annotated 3D LiDAR point clouds of Fuji apple trees for fruit detection scanned under different forced air flow conditions. *Data Brief* 29, 105248.

- Gené-Mola, J., Sanz-Cortiella, R., Rosell-Polo, J.R., Escolà, A., Gregorio, E., 2021a. In-Field Apple Size Estimation Using Photogrammetry-Derived 3D Point Clouds: Comparison of 4 Different Methods Considering Fruit Occlusions, Vol. 188. Elsevier, 106343.
- Gené-Mola, J., Sanz-Cortiella, R., Rosell-Polo, J.R., Escolà, A., Gregorio, E., 2021b. PFuji-size dataset: A collection of images and photogrammetry-derived 3D point clouds with ground truth annotations for Fuji apple detection and size estimation in field conditions. *Data Brief* 39, 107629.
- Gené-Mola, J., Sanz-Cortiella, R., Rosell-Polo, J.R., Morros, J.-R., Ruiz-Hidalgo, J., Vilaplana, V., Gregorio, E., 2020c. Fruit detection and 3D location using instance segmentation neural networks and structure-from-motion photogrammetry. *Comput. Electron. Agric.* 169, 105165.
- Gené-Mola, J., Sanz-Cortiella, R., Rosell-Polo, J.R., Morros, J.-R., Ruiz-Hidalgo, J., Vilaplana, V., Gregorio, E., 2020d. Fuji-SFM dataset: A collection of annotated images and point clouds for Fuji apple detection and location using structure-from-motion photogrammetry. *Data Brief* 30, 105591.
- Gené-Mola, J., Vilaplana, V., Rosell-Polo, J.R., Morros, J.-R., Ruiz-Hidalgo, J., Gregorio, E., 2019a. KFuji RGB-DS database: Fuji apple multi-modal images for fruit detection with color, depth and range-corrected IR data. *Data Brief* 25, 104289.
- Gené-Mola, J., Vilaplana, V., Rosell-Polo, J.R., Morros, J.-R., Ruiz-Hidalgo, J., Gregorio, E., 2019b. Multi-modal deep learning for Fuji apple detection using RGB-D cameras and their radiometric capabilities. *Comput. Electron. Agric.* 162, 689–698.
- Giancola, S., Valenti, M., Sala, R., 2018. A Survey on 3D Cameras: Metrological Comparison of Time-Of-Flight, Structured-Light and Active Stereoscopy Technologies. Springer.
- Girshick, R., 2015. Fast R-CNN. In: *Proceedings of the IEEE International Conference on Computer Vision. ICCV*, pp. 1440–1448.
- Girshick, R., Donahue, J., Darrell, T., Malik, J., 2014. Rich feature hierarchies for accurate object detection and semantic segmentation. In: *2014 IEEE Conference on Computer Vision and Pattern Recognition*. pp. 580–587.
- Haibo, L., Shuliang, D., Zunmin, L., Chuijie, Y., 2015. Study and experiment on a wheat precision seeding robot. *J. Robot.* 12–12.
- Halstead, M., Denman, S., Fookes, C., McCool, C., 2020. Fruit detection in the wild: The impact of varying conditions and cultivar. In: *2020 Digital Image Computing: Techniques and Applications. DICTA*, pp. 1–8.
- Hameed, K., Chai, D., Rassau, A., 2018. A comprehensive review of fruit and vegetable classification techniques. *Image Vis. Comput.* 80, 24–44.
- Harders, L.O., Czymmek, V., Knoll, F.J., Hussmann, S., 2021. Area yield performance evaluation of a nonchemical weeding robot in organic farming. In: *2021 IEEE International Instrumentation and Measurement Technology Conference. I2MTC, IEEE*, pp. 1–6.
- He, K., Gkioxari, G., Dollár, P., Girshick, R., 2017a. Mask R-CNN. In: *2017 IEEE International Conference on Computer Vision. ICCV*, pp. 2980–2988.
- He, Y., Liang, B., Zou, Y., He, J., Yang, J., 2017b. Depth errors analysis and correction for time-of-flight (ToF) cameras. *Sensors* 17 (1), 92.
- He, L., Schupp, J., 2018. Sensing and automation in pruning of apple trees: A review. *Agronomy* 8 (10), 211.
- Jiao, L., Zhang, F., Liu, F., Yang, S., Li, L., Feng, Z., Qu, R., 2019. A survey of deep learning-based object detection. *IEEE Access* 7, 128837–128868.
- Kamilaris, A., Kartakoullis, A., Prenafeta-Boldú, F.X., 2017. A review on the practice of big data analysis in agriculture. *Comput. Electron. Agric.* 143, 23–37.
- Kamilaris, A., Prenafeta-Boldú, F.X., 2018. Deep learning in agriculture: A survey. *Comput. Electron. Agric.* 147, 70–90.
- Khanna, R., Schmid, L., Walter, A., Nieto, J., Siegart, R., Liebis, F., 2019. A spatio-temporal spectral framework for plant stress phenotyping. *Plant Methods* 15 (1), 1–18.
- Kitzler, F., Barta, N., Neugschwandtner, R.W., Gronauer, A., Motsch, V., 2023. WE3DS: An RGB-D image dataset for semantic segmentation in agriculture. *Sensors* 23 (5), 2713.
- Krizhevsky, A., Sutskever, I., Hinton, G.E., 2012. ImageNet classification with deep convolutional neural networks. In: *Pereira, F., Burges, C., Bottou, L., Weinberger, K. (Eds.), Advances in Neural Information Processing Systems*, Vol. 25. Curran Associates, Inc.
- Kurtser, P., Arad, B., Ben-Shahar, O., van Bree, M., Moonen, J., van Tuijl, B., Edan, Y., 2016. Robotic data acquisition of sweet pepper images for research and development. In: *The 5th Israeli Conference on Robotics 2016. Air Force Conference Center Hertzilya, Israel*, 13–14 April, 2016.
- Kurtser, P., Ringdahl, O., Rotstein, N., Andreasson, H., 2020a. PointNet and geometric reasoning for detection of grape vines from single frame RGB-D data in outdoor conditions. In: *3rd Northern Lights Deep Learning Workshop*, Vol. 1. Tromsø, Norway 20–21 January, 2019, NLDL, pp. 1–6.
- Kurtser, P., Ringdahl, O., Rotstein, N., Berenstein, R., Edan, Y., 2020b. In-field grape cluster size assessment for vine yield estimation using a mobile robot and a consumer level RGB-D camera. *IEEE Robot. Autom. Lett.* 5 (2), 2031–2038.
- Kusumam, K., Krajník, T., Pearson, S., Cielniak, G., Duckett, T., 2016. Can you pick a broccoli? 3D-vision based detection and localisation of broccoli heads in the field. In: *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems. IROS, IEEE*, pp. 646–651.
- Kusumam, K., Krajník, T., Pearson, S., Duckett, T., Cielniak, G., 2017. 3D-vision based detection, localization, and sizing of broccoli heads in the field. *J. Field Robotics* 34 (8), 1505–1518.
- LeCun, Y., Bengio, Y., Hinton, G., 2015. Deep learning. *Nature* 521 (7553), 436.
- Li, Y., Guo, Z., Shuang, F., Zhang, M., Li, X., 2022a. Key technologies of machine vision for weeding robots: A review and benchmark. *Comput. Electron. Agric.* 196, 106880.
- Li, N., Ho, C.P., Xue, J., Lim, L.W., Chen, G., Fu, Y.H., Lee, L.Y.T., 2022b. A progress review on solid-state LiDAR and nanophotonics-based LiDAR sensors. *Laser Photonics Rev.* 16 (11), 2100511.
- Lin, T.-Y., Maire, M., Belongie, S., Bourdev, L., Girshick, R., Hays, J., Perona, P., Ramanan, D., Zitnick, C.L., Dollár, P., 2014. Microsoft COCO: Common objects in context. *arXiv:1405.0312*.
- Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.-Y., Berg, A.C., 2016. SSD: Single shot multibox detector. In: *Leibe, B., Matas, J., Sebe, N., Welling, M. (Eds.), Computer Vision. ECCV 2016*, Springer International Publishing, Cham, pp. 21–37.
- Liu, G., Nouaze, J.C., Touko Mbouembe, P.L., Kim, J.H., 2020. YOLO-tomato: A robust algorithm for tomato detection based on YOLOv3. *Sensors* 20 (7), 2145.
- Lobet, G., Draye, X., Périlleux, C., 2013. An online database for plant image analysis software tools. *Plant Methods* 9 (1), 1–8.
- Loey, M., ElSawy, A., Afify, M., 2020. Deep learning in plant diseases detection for agricultural crops: a survey. *Int. J. Serv. Sci. Manag. Eng. Technol. (IJSSMET)* 11 (2), 41–58.
- Lopes, A., Souza, R., Pedrini, H., 2022. A survey on RGB-D datasets. *Comput. Vis. Image Underst.* 222, 103489.
- Lu, Y., Young, S., 2020. A survey of public datasets for computer vision tasks in precision agriculture. *Comput. Electron. Agric.* 178, 105760.
- Luo, Z., Yang, W., Yuan, Y., Gou, R., Li, X., 2023. Semantic segmentation of agricultural images: A survey. *Inf. Process. Agric.*
- Marani, R., Milella, A., Petitti, A., Reina, G., 2021. Deep neural networks for grape bunch segmentation in natural images from a consumer-grade camera. *Precis. Agric.* 22 (2), 387–413.
- Milella, A., Marani, R., Petitti, A., Reina, G., 2019. In-field high throughput grapevine phenotyping with a consumer-grade depth camera. *Comput. Electron. Agric.* 156, 293–306.
- Mousazadeh, H., 2013. A technical review on navigation systems of agricultural autonomous off-road vehicles. *J. Terramech.* 50 (3), 211–232.
- Mylonas, N., Malounas, I., Mouseti, S., Vali, E., Espejo-García, B., Fountas, S., 2022. Eden library: A long-term database for storing agricultural multi-sensor datasets from UAV and proximal platforms. *Smart Agric. Technol.* 2, 100028.
- Neupane, C., Koirala, A., Wang, Z., Walsh, K.B., 2021. Evaluation of depth cameras for use in fruit localization and sizing: Finding a successor to kinect v2. *Agronomy* 11 (9), 1780.
- Patrício, D.I., Rieder, R., 2018. Computer vision and artificial intelligence in precision agriculture for grain crops: A systematic review. *Comput. Electron. Agric.* 153, 69–81.
- Pflugfelder, D., Metzner, R., van Dusschoten, D., Reichel, R., Jahnke, S., Koller, R., 2017. Non-invasive imaging of plant roots in different soils using magnetic resonance imaging (MRI). *Plant Methods* 13 (1), 102.
- Qi, C.R., Su, H., Mo, K., Guibas, L.J., 2017a. Pointnet: Deep learning on point sets for 3D classification and segmentation. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. pp. 652–660.
- Qi, C.R., Su, H., Niessner, M., Dai, A., Yan, M., Guibas, L.J., 2016. Volumetric and multi-view CNNs for object classification on 3D data. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. CVPR*.
- Qi, C.R., Yi, L., Su, H., Guibas, L.J., 2017b. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *Adv. Neural Inf. Process. Syst.* 30.
- Redmon, J., Divvala, S., Girshick, R., Farhadi, A., 2016a. You only look once: Unified, real-time object detection. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. pp. 779–788.
- Redmon, J., Divvala, S., Girshick, R., Farhadi, A., 2016b. You only look once: Unified, real-time object detection. In: *2016 IEEE Conference on Computer Vision and Pattern Recognition. CVPR*, pp. 779–788.
- Redmon, J., Farhadi, A., 2017. YOLO9000: Better, faster, stronger. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. CVPR*.
- Redmon, J., Farhadi, A., 2018. YOLOv3: An incremental improvement. *arXiv:1804.02767Comment: Tech Report*.
- Ren, S., He, K., Girshick, R., Sun, J., 2015. Faster R-CNN: Towards real-time object detection with region proposal networks. In: *Cortes, C., Lawrence, N., Lee, D., Sugiyama, M., Garnett, R. (Eds.), Advances in Neural Information Processing Systems*, Vol. 28. Curran Associates, Inc.
- Ringdahl, O., Kurtser, P., Edan, Y., 2019. Evaluation of approach strategies for harvesting robots: Case study of sweet pepper harvesting. *J. Intell. Robot. Syst.* 95 (1), 149–164.
- Rosu, R.A., Schütt, P., Quenzel, J., Behnke, S., 2022. LatticeNet: fast spatio-temporal point cloud segmentation using permutohedral lattices. *Auton. Robots* 46 (1), 45–60.
- Ruangurai, P., Ekpanyapong, M., Pruetong, C., Watwai, T., 2015. Automated three-wheel rice seeding robot operating in dry paddy fields. *Maejo Int. J. Sci. Technol.* 9 (3), 403.
- Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A.C., Fei-Fei, L., 2015. ImageNet large scale visual recognition challenge. *Int. J. Comput. Vis.* 115 (3), 211–252.

- Schunck, D., Magistri, F., Rosu, R.A., Cornelißen, A., Chebrolov, N., Paulus, S., Léon, J., Behnke, S., Stachniss, C., Kuhlmann, H., et al., 2021. Pheno4D: A spatio-temporal dataset of maize and tomato plant point clouds for phenotyping and advanced plant analysis. *PLoS One* 16 (8), e0256340.
- Simonyan, K., Zisserman, A., 2015. Very deep convolutional networks for large-scale image recognition. In: *International Conference on Learning Representations*.
- Sørensen, R.A., Rasmussen, J., Nielsen, J., Jørgensen, R.N., 2017. Thistle detection using convolutional neural networks. In: *EFITA WCCA 2017 Conference*. Montpellier Supagro, Montpellier, France, pp. 2–6.
- Su, H., Maji, S., Kalogerakis, E., Learned-Miller, E., 2015. Multi-view convolutional neural networks for 3D shape recognition. In: *2015 IEEE International Conference on Computer Vision*. ICCV, pp. 945–953.
- Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A., 2015. Going deeper with convolutions. In: *2015 IEEE Conference on Computer Vision and Pattern Recognition*. CVPR, pp. 1–9.
- Szeliski, R., 2022. *Computer Vision - Algorithms and Applications*, second ed. In: *Texts in Computer Science*, Springer.
- Tang, Y.-C., Wang, C., Luo, L., Zou, X., et al., 2020. Recognition and localization methods for vision-based fruit picking robots: a review. *Front. Plant Sci.* 11, 510.
- Tian, Y., Yang, G., Wang, Z., Wang, H., Li, E., Liang, Z., 2019a. Apple detection during different growth stages in orchards using the improved YOLO-V3 model. *Comput. Electron. Agric.* 157, 417–426.
- Tian, Y., Yang, G., Wang, Z., Wang, H., Li, E., Liang, Z., 2019b. Apple detection during different growth stages in orchards using the improved YOLO-V3 model. *Comput. Electron. Agric.* 157, 417–426.
- Vélez, S., Vacas, R., Martín, H., Ruano-Rosa, D., Álvarez, S., 2022. High-resolution UAV RGB imagery dataset for precis. agric. and 3D photogrammetric reconstruction captured over a pistachio orchard (*Pistacia vera* L.) in Spain. *Data* 7 (11).
- Vit, A., Shani, G., 2018. Comparing RGB-D sensors for close range outdoor agricultural phenotyping. *Sensors* 18 (12), 4413.
- Wang, C.-Y., Bochkovskiy, A., Liao, H.-Y.M., 2022. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors.
- Wu, Z., Song, S., Khosla, A., Yu, F., Zhang, L., Tang, X., Xiao, J., 2015. 3D shapenets: A deep representation for volumetric shapes. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. pp. 1912–1920.
- Yang, W., Feng, H., Zhang, X., Zhang, J., Doonan, J.H., Batchelor, W.D., Xiong, L., Yan, J., 2020. Crop phenomics and high-throughput phenotyping: Past decades, current challenges and future perspectives. *Mol. Plant*.
- Yu, F., Chen, H., Wang, X., Xian, W., Chen, Y., Liu, F., Madhavan, V., Darrell, T., 2020. Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In: *Proceedings Of The IEEE/CVF Conference On Computer Vision And Pattern Recognition*. pp. 2636–2645.
- Zahid, A., Mahmud, M.S., He, L., Heinemann, P., Choi, D., Schupp, J., 2021. Technological advancements towards developing a robotic pruner for apple trees: A review. *Comput. Electron. Agric.* 189, 106383.
- Zhao, C., Zhang, Y., Du, J., Guo, X., Wen, W., Gu, S., Wang, J., Fan, J., 2019a. Crop phenomics: current status and perspectives. *Front. Plant Sci.* 10, 714.
- Zhao, Z.-Q., Zheng, P., Xu, S.-T., Wu, X., 2019b. Object detection with deep learning: A review. *IEEE Trans. Neural Netw. Learn. Syst.* 30 (11), 3212–3232.