Improving Robotic Fruit Harvesting Within Cluttered Environments Through 3D Shape Completion

Federico Magistri Yue Pan Jake Bartels Jens Behley Cyrill Stachniss Chris Lehnert

***Abstract*—The world population is increasing and will, by 2050, nearly double its demand for food, feed, fuel, and fiber. Besides environmental challenges, labor shortage also poses crucial challenges to the agricultural production system. Automation of manual tasks in crop production can potentially increase efficiency but also lead to a change in agricultural practices for more effective usage of available land. In this paper, we address the problem of robotic fruit harvesting in challenging real-world scenarios such as vertical farms, where robotic sensing and acting need to cope with a cluttered environment. Robotic fruit harvesting is typically done by directly detecting a grasp point in the sensor reading, which can lie on the fruit itself or on its peduncle depending on crop harvesting requirements. However, grasp point detection is not always possible as the ideal grasp point may be hidden behind leaves or other fruits. Our approach exploits shape completion techniques allowing us to estimate the complete 3D shape of a target fruit together with its pose even under strong occlusions. In this way, we can estimate a grasp point even when the fruit is only partially visible. We evaluate our approach on a real robotic manipulator operating in a vertical farm growing different fruit species and employing different harvesting tools. Our experiments show that, on average, our proposed pipeline increases the success rate by 18.5 percentage points, in terms of end-effector positioning, compared to the most competitive baseline among the ones reported in this work, that does not rely on shape completion.**

***Index Terms*—Robotics and Automation in Agriculture and Forestry; Agricultural Automation; Perception for Grasping and Manipulation**

I. Introduction

WHILE the world population is rapidly growing, increasing the demand for food, feed, fuel, and fiber [14], our agricultural production systems are put under severe stress by labor shortage [9], [39] and loss of arable lands [15]. On the one hand, autonomous robotic systems have the potential to reduce the need for human labor [22] by taking over tasks currently performed by humans, e.g., weeding [2], [52], transportation [11], [17], or phenotyping [33], [42]. On the other hand, vertical farms offer a solution to increase yield while reducing land use [21], which are potentially easier to realize using autonomous robotic systems than human labor.

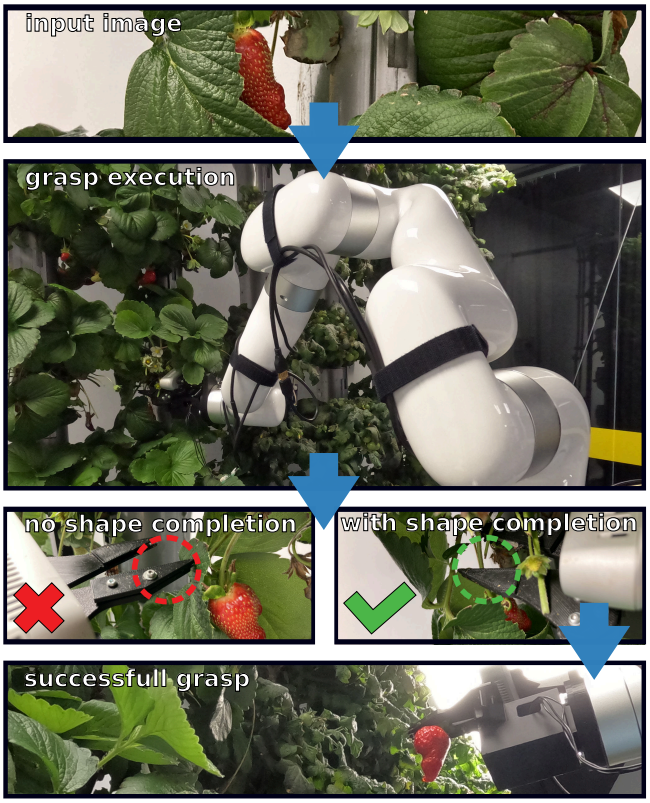


Fig. 1: Top: zoomed-in view of a partially visible fruit, which is common in real-world scenarios, making the harvesting task challenging. Second row: our manipulator performing an autonomous harvest in a vertical farm. Third row: (right) example of good end-effector positioning when estimating the fruit's shape and orientation, (left), the end-effector does not reach a good position for grasping when the fruit shape is not estimated. Bottom: successful harvest using our proposed pipeline exploiting shape completion.

In this paper, we consider the problem of autonomous robotic fruit harvesting, with experiments on real strawberries and tomatoes plants. To successfully harvest a fruit, a robot needs to precisely locate where to place its end-effector. Such a point can be on the fruit itself or on the fruit peduncle depending on the design of the end-effector tool and the fruit species [48]. However, the estimation of a grasp point is not trivial due to the cluttered nature of agricultural environments. As a practical example, grasp points may be hidden by leaves or other fruits as shown exemplarily in Fig. 1. To overcome such challenges, a robotic system should have a high-level, semantic, and geometric understanding of its surroundings and especially the task-relevant objects.

Manuscript received: December 17, 2023; Revised: March 22, 2024; Accepted: May 28, 2024.

This paper was recommended for publication by Editor Hyungpil Moon upon evaluation of the Associate Editor and Reviewers' comments.

F. Magistri, Y. Pan, J. Behley, and C. Stachniss are with the Center for Robotics, University of Bonn, Germany. F. Magistri, J. Bartels and C. Lehnert are with the Queensland University of Technology, Australia. C. Stachniss is additionally with the Department of Engineering Science at the University of Oxford, UK and with the Lamarr Institute for Machine Learning and Artificial Intelligence, Germany.

This work has partially been funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy, EXC-2070 – 390732324 – PhenoRob and under STA 1051/5-1 within the FOR 5351 (AID4Crops).

Digital Object Identifier (DOI): see top of this page.

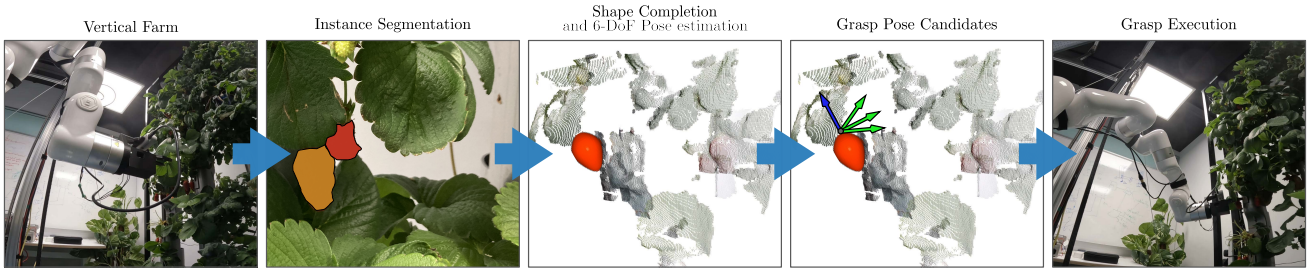


Fig. 2: An overview of our harvesting pipeline. From left to right: our robot arm in its initial configuration. We segment individual fruits using the color information of an RGB-D camera. For each segmented fruit, we estimate its complete 3D shape and its 6-DoF pose. Based on the fruit's pose and shape we evaluate different grasp pose candidates exploiting the fruit's symmetry before attempting the grasp.

In recent years there has been an increasing interest in robotic fruit harvesting leading to three main paradigms: (i) detecting and cutting the fruit peduncle [3] with applications to capsicums and strawberries, (ii) attaching to the fruit while using suction devices [38] with applications to tomatoes and apples, (iii) directly gripping the fruit and applying a twisting motion [45] with applications to tomatoes and berries. Entailing that often a specifically tailored vision system has to be implemented to comply with the tool design.

The main contribution of this paper is a novel approach integrating 3D shape completion of fruits plus a 6 Degrees of Freedom (DoF) pose estimation module into a robotic harvesting pipeline of real fruits and the experimental validation of the working pipeline on two fruit species. This is a novel setup and we demonstrate in this paper its benefits on real tomato and strawberry plants. This gain is achieved as the robot has a better understanding of the scene and thus can perform better robotic harvesting. Notably, our approach can be integrated seamlessly with different end-effector tools designed for cutting the peduncle or gripping a fruit.

In sum, we make three key claims: our approach is able to (i) yield a higher success rate by integrating shape completion into a harvesting pipeline; (ii) find good grasp candidates while using different end-effectors designed for different fruit species; (iii) increase the robustness of the harvesting pipeline by evaluating different grasp pose candidates.

II. Related Work

Due to the heterogeneous nature of agricultural environments, a diverse number of approaches have been proposed for robotic fruit harvesting such as cutting the peduncle, gripping and pulling the fruit, and vacuum sucking depending on the different tool design [20]. In the first case, Sa et al. [40] propose to detect the fruit's peduncle to directly estimate a cutting point. Estimating the fruit orientation is an orthogonal approach for fruit picking when cutting the peduncle giving the robot more spatial information. Such an estimation can be done by directly regressing the orientation by means of neural networks [49] or by estimating keypoints of the fruit to recover its main axis [19], [24], [44], [55]. In the second case, the end-effector tool is typically designed with 3 or more fingers to firmly grip the fruit [7], [43]. In this scenario, the robot needs to estimate the complete shape of the fruit to secure the grasp [10]. Lastly, for vacuum-sucking tools, a robot needs to estimate a planar patch on the fruit's surface to attach for the end-effector [38]. Note that each tool design needs an ad-hoc vision system to correctly estimate where to place the end-effector. In contrast, our proposed approach can be easily integrated into harvesting pipelines when using different tools.

Additionally, other factors hamper research in robotic harvesting such as simplistic testing environments [16] and robotic platforms tailored to a specific growing system [51]. In contrast, our proposed solution can be deployed in real agricultural environments.

Lehnert et al. [25] propose to recover the fruit's shape and pose by fitting an ellipsoid on the segmented fruits. Such an approach can fail in case of a limited number of sample points, for example, due to occlusions. Our approach, presented here, exploits a robust shape completion and pose estimation pipeline that learns a prior over fruit shapes exploiting high-resolution point clouds. Additionally, we exploit the almost symmetric fruit shape to evaluate different grasp directions to increase robustness to occlusions. Ren et al. [37] deploy a soft gripper to harvest strawberries in a vertical farm. They propose a scene categorization to discard occluded fruits from the harvesting process. In contrast, our approach can harvest fruit also in cluttered environments.

Shape completion has been used for mapping [30] and next best view planning [32], [53] by fitting an ellipsoid in 3D point clouds of fruits. Instead, we use a shape completion module to drive a harvesting pipeline. In prior work, we estimate complete 3D shapes of fruit by exploiting a high-precision LiDAR system to build prior knowledge over a general fruit shape [29]. Pan et al. [34] extend it by creating multi-resolution panoptic maps where the authors estimate the complete 3D shape and 6-DoF pose of fruits. Our work integrates our pose estimation module [34] in the harvesting pipeline, exploiting the almost symmetric shape of fruits to better reason the arm's end-effector placement.

Beyond fruit harvesting, shape completion approaches have been proposed in the agricultural context for yield estimation [6], [18] and plant phenotyping [28], [31]. These works focus mainly on monitoring the plant's status. In contrast, we use the estimated 3D shapes to interact with the plants for harvesting fruits.

III. Our Approach to Robotic Fruit Harvesting

Given an RGB-D image, we segment fruit instances using an image-based instance segmentation approach. Afterward, for each segmented fruit, we estimate its complete 3D shape together with its 6-DoF pose. At this point, exploiting the

almost rotational symmetric shape of the fruits, we evaluate different grasp pose candidates and select the candidate with the highest manipulability. We show an overview in Fig. 2

A. Vision Modules

The goal of our vision modules is to estimate a 3D mesh $\mathcal{S}$ and a corresponding pose $\mathsf{T}_{\mathcal{S}} \in \mathrm{SE}(3)$ for each fruit in an input RGB-D image of height H and width W, where we denote with $\boldsymbol{I} \in \mathbb{R}^{3\times H\times W}$ the RGB image and $\boldsymbol{D} \in \mathbb{R}^{H\times W}$ the depth channel. With $\boldsymbol{I}[u,v]$ and $\boldsymbol{D}[u,v]$ we refer to the RGB or depth value at pixel location (u,v), respectively.

Instance Segmentation: We use Mask R-CNN [12] for instance segmentation of fruits using the RGB image $\boldsymbol{I}$, where we use a ResNet18 [13] as backbone. Thus, we obtain F binary masks, $\boldsymbol{M}_1,\dots,\boldsymbol{M}_F$, $\boldsymbol{M}_i \in \{0,1\}^{H\times W}$, of the segmented fruits in image $\boldsymbol{I}$ with $\mathcal{M}_i = \{(u,v) \mid \boldsymbol{M}_i[u,v] = 1\}$ being the pixel locations of the foreground, i.e., $\boldsymbol{M}_i[u,v] = 1$, of each mask $\boldsymbol{M}_i$. The corresponding set of point clouds, $\mathcal{P}_1,\dots,\mathcal{P}_F$, where $\mathcal{P}_i = \{\boldsymbol{p}_1,\dots,\boldsymbol{p}_N\}$, $\boldsymbol{p}_j \in \mathbb{R}^3$, can be obtained by having access to the camera intrinsics and $\boldsymbol{D}$ resulting in the backprojected point clouds $\mathcal{P}_i$ of every fruit, where we use only depth values of $\boldsymbol{D}$ given by $\mathcal{M}_i$.

Shape Completion and Pose Estimation: In line with our previous approaches [29], [34], we use DeepSDF [35] to learn a shape prior over different fruit species by exploiting high-resolution point clouds of complete fruits. DeepSDF [35] takes as input a query position $\boldsymbol{x} \in \mathbb{R}^3$ and a latent shape code $\boldsymbol{z} \in \mathbb{R}^C$, and predicts the SDF value $s \in \mathbb{R}$ at $\boldsymbol{x}$ through a decoder network D_θ:

$$s = D_\theta(\boldsymbol{x}, \boldsymbol{z}), \tag{1}$$

where θ are the model weights of neural network D_θ.

At training time, we learn model weights θ such that the predicted SDF value is close to the ground truth value determined from the complete high-resolution point clouds. At inference time, we optimize a shape code $\boldsymbol{z}_i$ via backpropagation using as input a partial point cloud $\mathcal{P}_i$ using the trained D_θ with fixed model weights θ. We refer to our previous paper [34] for more details. At this point, we can compute a dense SDF volume by querying D_θ at a regular 3D grid of coordinates that we can convert to a complete mesh $\mathcal{S}_i$ via marching cubes [26]. This means that the shape code $\boldsymbol{z}_i$ univocally defines a 3D mesh. With this setup, fruit shapes $\mathcal{S}_i$ are represented in a so-called canonical pose corresponding to the peduncle pointing upwards, which can lead to a failed grasp as fruits are not always aligned with such a direction.

To jointly estimate a 3D shape, namely estimating its shape code $\boldsymbol{z}_i$, and its corresponding pose $\mathsf{T}_{\mathcal{S}_i}$, we define three loss functions: $\mathcal{L}_\mathrm{s}$, $\mathcal{L}_\mathrm{d}$, and $\mathcal{L}_\mathrm{m}$. The surface reconstruction loss $\mathcal{L}_\mathrm{s}$ is responsible for keeping the points from the target point cloud $\mathcal{P}_i$ close to the iso-surface, i.e., coordinates of the SDF volume with $s = 0$, of the SDF predicted by D_θ. In this way, the predicted 3D mesh will closely align with the input point cloud and exploit its encoded prior to estimate how the fruit may look in regions where the robot has no observations via the point cloud $\mathcal{P}_i$. Formally, our surface reconstruction loss $\mathcal{L}_\mathrm{s}$ is given by:

$$\mathcal{L}_\mathrm{s} = \frac{1}{|\mathcal{P}_i|} \sum_{\mathbf{p}\in\mathcal{P}_i} D_\theta\left(\mathsf{T}_{\mathcal{S}_i}\,\mathbf{p}, \boldsymbol{z}_i\right), \tag{2}$$

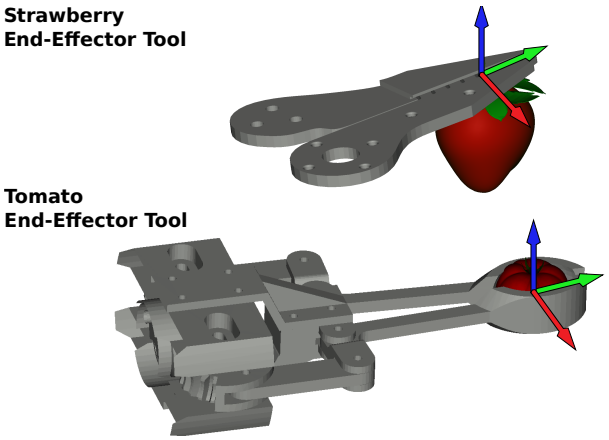


Fig. 3: Designs of our end-effectors and their envisioned grasp pose. Top: a strawberry grasping tool designed to cut the peduncle. Bottom: tomato grasping tool designed to grip and pull the fruit. Our proposed grasping pipeline can be easily adapted to both designs.

where $\mathbf{p}$ is the corresponding homogeneous coordinate of $\boldsymbol{p}$. Thus, $\mathcal{L}_s$ is minimized if the iso-surface is close to the observed point cloud, which means that $\mathcal{L}_s$ approaches zero when evaluated at the measured points $\boldsymbol{p} \in \mathcal{P}_i$.

Using differentiable rendering [50], [54], we render a depth image $\hat{\boldsymbol{D}}_i$ and a binary mask $\hat{\boldsymbol{M}}_i$ from the estimated shape $\mathcal{S}_i$ to obtain consistency between the robot's observation and the predicted shape using the depth rendering and mask rendering losses $\mathcal{L}_\mathrm{d}$ and $\mathcal{L}_\mathrm{m}$:

$$\mathcal{L}_\mathrm{d} = \frac{1}{|\mathcal{M}_i|} \sum_{(u,v)\in\mathcal{M}_i} \left|\left|\hat{\boldsymbol{D}}_i[u,v] - \boldsymbol{D}[u,v]\right|\right|^2, \tag{3}$$

$$\mathcal{L}_\mathrm{m} = \frac{1}{|\mathcal{M}_i|} \sum_{(u,v)\in\mathcal{M}_i} \left|\left|\hat{\boldsymbol{M}}_i[u,v] - \boldsymbol{M}_i[u,v]\right|\right|^2, \tag{4}$$

which ensures that the predicted shape $\mathcal{S}_i$ is also consistent with the observed RGB image and depth map.

In sum, we optimize the following loss for each fruit:

$$\mathcal{L} = w_\mathrm{s}\,\mathcal{L}_\mathrm{s} + w_\mathrm{d}\,\mathcal{L}_\mathrm{d} + w_\mathrm{m}\,\mathcal{L}_\mathrm{m} + w_\mathrm{r}\,\mathcal{L}_\mathrm{r}, \tag{5}$$

where $\mathcal{L}_\mathrm{r} = \|\boldsymbol{z}_i\|^2$ is a regularization term and w_s, w_d, w_m, w_r are the weights for each loss term. By minimizing the loss $\mathcal{L}$, we can estimate the shape code $\boldsymbol{z}_i$ and pose $\mathsf{T}_{\mathcal{S}_i}$ of the i^{th} fruit. We refer to Pan et al. [34] for more details on the implementation of the optimization routine to efficiently estimate the fruit's shape and pose.

B. Tool-specific Grasp Pose Initialization

In our approach, we engineered two distinct end-effectors tailored for strawberries and tomatoes, respecting each crop's unique harvesting requirements. The strawberry end-effector features a dual-function scissor mechanism, designed to sever and secure the peduncle above the strawberry, ensuring a gentle and efficient harvest. Enhanced with a safety guard, this tool prioritizes human and environmental safety. Conversely, the tomato end-effector incorporates an adaptive 4-bar linkage gripper with parabolic fingertips. This design adapts to various tomato sizes, enabling a smooth detachment from the vine without necessitating a cutting action.

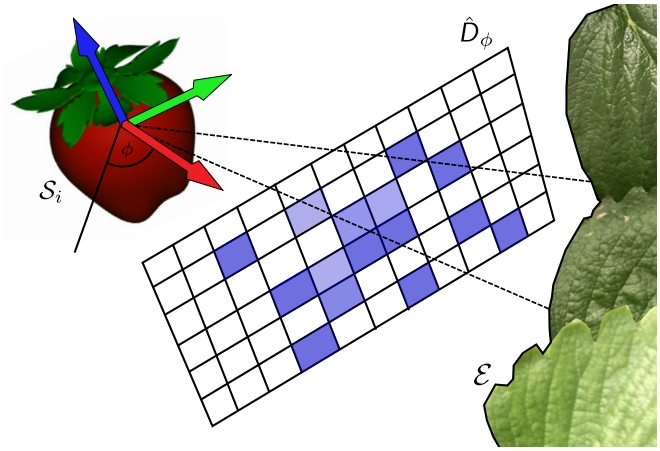


Fig. 4: Determining the free space of the grasp candidate is achieved via rendering of a depth image $\hat{D}_\phi$ using the candidate pose T_ϕ at the grasp point location determined from the completed mesh $\mathcal{S}_i$. The depth image $\hat{D}_\phi$ stores the result of the ray casting operation with the environment mesh $\mathcal{E}$.

With the estimated pose $\mathsf{T}_{\mathcal{S}_i}$ and knowing the end-effector design, we compute the grasp pose $\mathsf{T}'_{\mathcal{S}_i}$ using a tool-specific transformation T_g, which is relatively applied to the determined estimated fruit pose $\mathsf{T}_{\mathcal{S}_i}$ as follows:

$$\mathsf{T}'_{\mathcal{S}_i} = \mathsf{T}_{\mathcal{S}}\,\mathsf{T}_g, \tag{6}$$

where the rotation part of T_g is the identity matrix and the translation part is the vector $\boldsymbol{g} \in \mathbb{R}^3$ whose values depend on the used end-effector tool as illustrated in Fig. 3.

Specifically, when gripping the fruit itself, $\boldsymbol{g} = [0, 0, 0]^\top$ meaning that we target the center of the estimated fruit shape. While, for cutting the peduncle, $\boldsymbol{g} = [0, 0, z^*]^\top$ with the scalar z^* computed from the estimated complete mesh $\mathcal{S}_i$ as

$$z^* = \max_{(x,y,z)\in\mathcal{V}_{\mathcal{S}_i}} z, \tag{7}$$

where we denote by $\mathcal{V}_{\mathcal{S}_i}$ the set of vertices of mesh $\mathcal{S}_i$.

To summarize, the only change needed to adapt our grasp pose initialization to different tool designs used in this paper is a translation along the fruit estimated z-axis. Furthermore, this can be easily adapted to other fruit requirements.

C. Grasp Pose Estimation

Having estimated the fruit's shape $\mathcal{S}_i$ and its grasp pose $\mathsf{T}'_{\mathcal{S}_i}$, which now includes the target grasp point, we can now reason about how to approach the fruit to execute the harvesting. We note that fruits are almost symmetrical around their main axis, i.e., the z-axis estimated by our pose estimation module. Meaning that all poses obtained by rotating around this axis are potential grasp candidates. We define such a set with $\mathcal{T} = \{\mathsf{T}_{\phi_1}, \ldots, \mathsf{T}_{\phi_n}\}$ where ϕ_i corresponds to the rotation angle around the z-axis. For simplicity, we consider only angle increments of 15°, i.e., $\phi_1 = 0°$, $\phi_2 = 15°$, ...

At this point, we need to evaluate which poses are reachable by our manipulator. We start by integrating the current depth frame D using a standard volumetric mapping pipeline [46]. Again, using marching cubes, we can extract a mesh $\mathcal{F}$ that represents the current robot's observation. We can, now, define

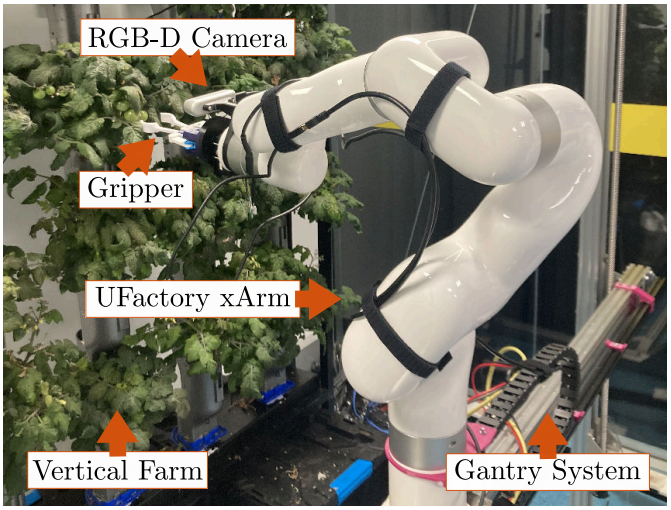


Fig. 5: A visual impression of our vertical growing system where we highlighted each component.

the mesh $\mathcal{E}$ representing the robot's surrounding environment, containing the mesh $\mathcal{F}$ and the known vertical farm 3D model, i.e., a CAD model of the pipes and walls blocking the path of the manipulator.

Using the mesh $\mathcal{E}$, we can now evaluate potential manipulator paths and determine how to approach the fruit such that we have the most possible free space. To this end, we render the mesh $\mathcal{E}$ from each candidate pose T_ϕ obtaining a depth image $\hat{D}_\phi$ via ray casting, where we initialize the rendered image with zero anywhere. Therefore, if the raycasting starting at the current candidate pose T_{ϕ_i} hits a surface it will receive a non-zero value at that image location, as shown in Fig. 4. We now define the set of grasp pose candidates as the set $\mathcal{T}_c \subseteq \mathcal{T}$ as follows:

$$\mathcal{T}_c = \{\mathsf{T}_\phi \in \mathcal{T} \mid [\![\hat{D}_\phi]\!] < t\}, \tag{8}$$

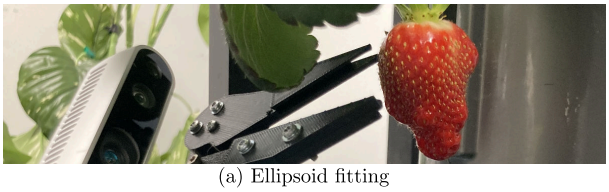
where the operator $[\![\cdot]\!]$ counts how many pixels of the rendered image $\hat{D}_\phi$ have a non-zero value, i.e., where raycasting hit a surface, and t is a user-defined threshold. Note that we only evaluate Eq. (8) in a patch around the image centered roughly corresponding to the robot footprint. For each candidate pose $\mathsf{T} \in \mathcal{T}_c$, we compute the inverse kinematic solution $\boldsymbol{q} = \mathrm{IK}(\mathsf{T})$, where $\boldsymbol{q} \in \mathbb{R}^J$ corresponds to a specific joint angle configuration of the J joints. For a configuration $\boldsymbol{q}$, we evaluate the arm manipulability μ using the Jacobian $J(\boldsymbol{q})$ evaluated at the specific joint configuration $\boldsymbol{q}$:

$$\mu(\boldsymbol{q}) = \boldsymbol{J}(\boldsymbol{q})\,\boldsymbol{J}^\top(\boldsymbol{q}). \tag{9}$$

We can then select the pose $\mathsf{T}^* \in \mathcal{T}_c$ that maximizes the arm's manipulability:

$$\mathsf{T}^* = \max_{\mathsf{T}\in\mathcal{T}_c} \mu(\mathrm{IK}(\mathsf{T})). \tag{10}$$

With the selected T^* and the corresponding joint angle configuration, we use RRT [23] within MoveIt [8] for motion planning with the TRAC-IK [5] solver for improved solution rate and time compared to standard solvers.

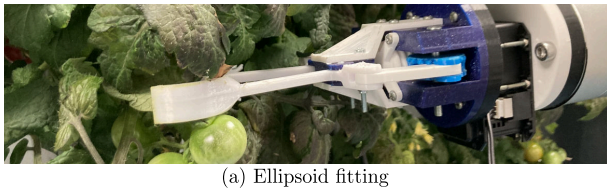


(a) Ellipsoid fitting

(b) Estimating orientation

(c) Our shape completion approach

Fig. 6: We show different end-effector positioning. In (a), the fruit pose and shape estimated by ellipsoid fitting [25] do not align well with the real fruit shape and pose leading to wrong positioning. In (b), the fruit pose is correctly estimated regressing the fruit orientation with a deep neural network [49]. However, the missing shape estimation leads to wrong positioning. In (c), we show the positioning obtained with our approach. The end-effector is correctly positioned behind a leaf approaching the fruit's peduncle. Note that detecting the grasp point [40] failed as it was covered by a leaf.



(a) Ellipsoid fitting

(b) Grasp point detection

(c) Our shape completion approach

Fig. 7: We show different end-effector positioning. In (a), the shape estimated by ellipsoid fitting [25] does not align well with the real fruit shape and pose leading to wrong positioning. In (b), we directly detect the grasp point [40]. However, the missing shape estimation leads to wrong positioning due to the partially occluded fruit. In (c), we show the positioning obtained with our proposed approach. The end-effector tool is correctly positioned to approach the fruit.

IV. Experimental Evaluation

The main focus of this work is the integration of a shape completion and pose estimation module in a robotic harvesting pipeline to robustly harvest fruits in a real-world scenario. Our experiments show that our approach can (i) yield a higher success rate by integrating shape completion systems in a harvesting pipeline; (ii) find good grasp candidates while using different end-effectors designed for different fruit species; (iii) increase the robustness of the harvesting pipeline by evaluating different grasp pose candidates.

A. Experimental Setup and Hyperparameters

Vision Modules: For segmenting individual fruits, we train Mask R-CNN [12] for 200 epochs using AdamW [27] as optimizer. We set the initial learning rate to $1 \cdot 10^{-4}$ and employ an exponentially decaying learning rate schedule. As training set, we use two public datasets LaboroTomato [1] and StrawDI [36] for tomatoes and strawberries, respectively. We collected two datasets to train our shape completion model, one for each species, using a high-precision LiDAR system as described in Schunck et al. [41]. Following the settings used in our previous work [29], [34], we train our model for 3000 epochs using AdamW [27]. We set the initial learning rate to $5 \cdot 10^{-4}$ and use a stepping strategy where we halve the learning rate every 300 epochs.

Robotic System: Our robotic platform consists of a custom-built gantry system and a UFactory xArm manipulator with 6-DoF. The gantry system provides two additional prismatic joints for horizontal and vertical movement of the arm at a fixed distance from the vertical growing system. For sensing, we use an Intel RealSense d435i RGB-D camera mounted on the manipulator's wrist together with our grippers, see Fig. 5. We additionally refer to Barthelme et al. [4] for more details on the vertical farm.

Metrics: We report three metrics to evaluate the end-effector positioning. Attempts ratio, ρ_a, i.e., the number of attempts over the number of detections. Success ratio, ρ_s, i.e., the number of successes over the number of detections. We finally report the success rate over the attempts rate ρ_s/ρ_a. To evaluate the harvesting performances, we report the ratio between the number of successful harvests over the number of attempts ρ_h/ρ_a. We evaluate such metrics over 32 and 27 trials for strawberries and tomatoes.

Baselines: We compare our approach with three baselines: (i) directly detecting the grasp point [40]; (ii) shape completion and pose estimation by ellipsoid fitting [25]; (iii) estimating the fruit pose by regressing its orientation with a deep neural network [49]. This last baseline cannot be adapted to tomatoes given their almost spherical shape. We train the baselines on the same publicly available datasets, StrawDI [36] and LaboroTomato [1]. Given that such datasets provide instance segmentation labels, we manually label bounding boxes for the peduncles and orientation vectors for estimating the pose of the fruits. This was done specifically to give the same data as input during training to ensure a fair evaluation.

B. End-Effector Positioning

The first experiment shows that our approach (i) yields a higher success rate by integrating a shape completion approach into a harvesting pipeline, and (ii) finds good grasp candidates while using different end-effectors designed for different fruit species, and thus supports our first two claims.

We report a quantitative evaluation of end-effector positioning for both, strawberries and tomatoes, in Tab. I. To

TABLE I: Quantitative analysis of end-effector positioning. Our proposed approach outperforms all the baselines.

Approach	Strawberry			Tomato		
	ρ_a ↑ [%]	ρ_s ↑ [%]	ρ_s/ρ_a ↑ [%]	ρ_a ↑ [%]	ρ_s ↑ [%]	ρ_s/ρ_a ↑ [%]
Grasp Point Detection [40]	31.25	21.88	70.00	**100.00**	66.67	66.67
Rotation Estimation [49]	78.13	40.63	52.00	-	-	-
Ellipsoid Fitting [25]	68.75	37.50	54.55	88.89	55.56	62.50
Ours	**84.38**	**62.50**	**74.07**	96.30	**81.48**	**84.62**

TABLE II: Inference time needed for estimating the grasp pose.

	Grasp Point Detection [40]	Rotation Estimation [49]	Ellipsoid Fitting [25]	Ours
Inference Time [s]	**0.03**	0.06	0.05	0.13

fairly evaluate the different approaches, we applied our pose selection strategy (Sec. III-C) to the baselines as well. Note that we vary the starting pose of our manipulator to test the different approaches with different camera viewpoints.

Strawberries experiment: We notice that our approach outperforms the baselines in all reported metrics. Specifically, we obtain a success ratio ρ_s of 62.5%, which is over 20 percentage points better than the closest baseline [49]. We additionally outperform the baselines in terms of attempt rate ρ_a even though by a smaller margin, 84.4% against 78.1% of the closest baseline [49]. The combination of such metrics indicates that our approach is able to find good grasp poses consistently. This is also reflected by the success over attempt ratio ρ_s/ρ_a where our approach reaches 74.1%. With the closest baseline slightly above 50%. Most of the failed grasps when using our proposed pipeline are caused by errors in the fruit pose estimation. While errors due to a wrong shape estimation are considerably less frequent. Additionally, we want to highlight that positioning the end-effector tool by directly detecting the grasp point, the peduncle, in this case, turns out to be extremely challenging due to two main reasons. First, the strawberry peduncle is rather small and thin, thus harder to detect for a deep neural network than the strawberries themselves. Second, the peduncle is often hidden behind leaves or other fruits demonstrating why a shape completion approach is needed to be able to grasp those fruits. We, additionally, show the different end-effector positioning when using different approaches for estimating the grasp pose in Fig. 6. It can be seen that, while our approach is able to correctly place the end-effector behind a leaf Fig. 6(c), the baselines fail because of wrong pose estimation Fig. 6(a) or wrong shape estimation Fig. 6(b).

Tomatoes experiment: We obtain the highest success rate ρ_s reaching 81.5% over 66.7% of the closest baseline [40]. Similarly, we obtain a success over attempt ratio ρ_s/ρ_a of 84.6% which is 18 percentage points higher than the closest baseline [40]. Detecting the grasp point directly for tomatoes means targeting the center of the segmented fruit explaining the 100% attempt rate ρ_a. Our approach yields a lower attempt rate, reaching 96.3% due to a wrong shape estimation. We want to highlight that directly detecting the grasping point is more effective for gripping the fruit instead of cutting the peduncle. In contrast, fitting an ellipsoid to tomatoes is less effective mostly due to an overestimation of the fruit shape. Such an outcome underlines the challenges in developing a general pipeline for robotic fruit harvesting. Additionally, we note a general increase in the attempt rate with the lowest score of 88.9% which is higher than the best approach in the strawberries experiment. This outcome can be explained by the different plant species. In the tomatoes experiment, fruits are more hidden in the dense canopy compared to the strawberries experiment. Meaning that our instance segmentation network is not able to detect unreachable tomatoes. In contrast, our network segments more strawberries thanks to the plant canopy being less dense. Afterward, our pose selection strategy evaluates some strawberries to be unreachable.

Fig. 8: Examples of good end-effector positioning in strawberry (top row) and tomato plants (bottom row) using our proposed harvesting pipeline. Where the end-effector often needs to be positioned behind a leaf or between multiple fruits.

TABLE III: We compare the harvest success rate with and without our pose selection strategy evaluating it on both fruit species. The results show the benefits of using our proposed approach.

Grasp Pose Selection	Strawberry ρ_h/ρ_a ↑ [%]	Tomato ρ_h/ρ_a ↑ [%]
✗	56.25	62.96
✓	**68.75**	**70.37**

Additionally, we show the different end-effector positioning when using different approaches for estimating the grasp pose in Fig. 7. It can be seen that correctly estimating the complete shape of the fruit is crucial for good end-effector positioning. Our approach robustly estimates the shape of a partially occluded tomato leading to a good end-effector positioning, Fig. 7(c). In Fig. 7(b) a complete shape is not estimated resulting in an end effector positioning slightly off, in contrast fitting an ellipsoid yields an inaccurate shape estimation, thus, wrong positioning Fig. 7(a). To better appreciate the challenging environment and the precise positioning needed to grasp fruits with our robot, we show qualitative end-effector positioning using our shape completion approach in Fig. 8. Finally, we report the inference time needed by each

approach to estimate the grasp pose in Tab. II. As expected, directly estimating the grasp point has the fastest inference time of $0.03\,\mathrm{s}$ given that it only requires one forward pass. Our proposed approach is the slowest given that we need multiple forward passes. However, we only need $0.13\,\mathrm{s}$ to estimate the grasp point which is sufficient as currently planning the path and performing the manipulation are the main bottleneck for grasping. We want to highlight that we did not optimize any of the approaches, meaning that the reported runtimes can be further improved. In our experiments, we used an NVIDIA Quadro RTX A5000 given the application in a vertical farm.

C. Autonomous Harvesting

Finally, we analyze the impact of our grasp pose selection strategy on robotic harvesting, showing that it increases the robustness of the harvesting pipeline by evaluating different grasp pose candidates. In Tab. III, we report harvesting success metrics with and without our pose selection strategy. When our pose selection strategy is not used, we select an approach direction closest to the perpendicular to the plane given by the vertical growing system. While harvesting strawberries, the results suggest that our pose selection strategy increases the harvesting success rate by 12 percentage points over the naive approach. As a concrete example of such a difference, consider a peduncle hidden by a leaf. Our pose selection module would find an approach direction to move around the leaf, while the naive approach would push the leaf toward the growing system making the peduncle unreachable. Similarly, for harvesting tomatoes, our proposed pose selection strategy increases the harvesting success rate from 63% to 70.4%. As for the strawberries experiment, when naively approaching a fruit, without considering its surroundings, we notice that most errors come from the robot pushing the target fruit away. Such a situation is particularly challenging for dense canopies such as tomatoes. Additionally, tomato branches are stiffer than strawberry peduncles, meaning that an interaction between robot and plant causes a much bigger plant movement. Interestingly, in the tomatoes experiment, there is a clear regression in the harvesting success rate compared to the positioning success, mostly coming from the tomato slipping outside the gripper when the arm retracts, suggesting that a more precise positioning is needed and that an ad-hoc path needs to be planned in such a scenario.

In summary, our evaluation suggests that our method provides better end-effector positioning than the baselines on different fruit species requiring different tool designs. This indicates that our proposed approach can be easily integrated into the harvesting pipeline regardless of the end-effector design. We additionally show that our pose selection strategy increases the harvesting success rate by exploiting the almost symmetrical fruit shape. Thus, we supported all our claims with this experimental evaluation.

V. Conclusion

In this paper, we presented a novel approach for robotic fruit harvesting exploiting shape completion techniques. Our approach takes as input a single RGB-D image to obtain a point cloud of individual fruits. Our method exploits recent developments in differentiable rendering to jointly estimate a fruit shape and its 6-DoF in challenging, cluttered agricultural environments. This allows us to successfully estimate a grasp pose even when the target fruit is occluded by leaves or other fruits. Despite these encouraging results, there is further space for improvement. The most common cause of failed grasps is a wrong pose estimation for the strawberries and a wrong shape estimation for the tomatoes. In both cases, considering more than just one frame could benefit the estimations by having a more informative input, especially when paired with active vision pipelines. Finally, our pose selection strategy can be further improved by additionally evaluating orientation changes along an axis other than the fruit z-axis. While being tested only on a vertical farm, our approach for robotic fruit harvesting is rather general. Transferring our approach to a mobile robot is straightforward and only requires an odometry estimation [47] and a mapping system [46]. We implemented and evaluated our approach on a robotic arm operating in a real vertical farm system with different fruit species and provided comparisons to other existing techniques, and supported all claims made in this paper. The experiments suggest that robustly estimating the shape and the 6-DoF pose of each fruit is a key element in successfully harvesting fruit. We show that our proposed approach can be easily integrated into harvesting pipelines irrespective of the envisioned end-effector. Additionally, we show that we can improve the harvesting performances by evaluating different grasp pose candidates exploiting the almost symmetrical shape of fruits. Thus, demonstrating how robotic fruit harvesting benefits from the integration of novel computer vision techniques.

References

[1] Laborotomato: Instance segmentation dataset. Available Online: https://github.com/laboroai/LaboroTomato, 2020.

[2] A. Ahmadi, M. Halstead, and C. McCool. BonnBot-I: A Precise Weed Management and Crop Monitoring Platform. In *Proc. of the IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems (IROS)*, 2022.

[3] B. Arad, J. Balendonck, R. Barth, O. Ben-Shahar, Y. Edan, T. Hellström, J. Hemming, P. Kurtser, O. Ringdahl, T. Tielen, et al. Development of a sweet pepper harvesting robot. *Journal of Field Robotics (JFR)*, 37(6):1027–1039, 2020.

[4] Q. Barthelme and C. Lehnert. Implementation of a vertical hydroponic farming end effector for harvesting. In *Proc. of the Australasian Conf. on Robotics and Automation (ACRA)*, 2021.

[5] P. Beeson and B. Ames. Trac-ik: An open-source library for improved solving of generic inverse kinematics. In *Proc. of the IEEE Intl. Conf. on Humanoid Robots*, 2015.

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

[7] J. Brown and S. Sukkarieh. Design and evaluation of a modular robotic plum harvesting system utilizing soft components. *Journal of Field Robotics (JFR)*, 38(2):289–306, 2021.

[8] D.T. Coleman, I.A. Sucan, S. Chitta, and N. Correll. Reducing the barrier to entry of complex robotic software: a moveit! case study. *Journal of Software Engineering in Robotics*, 5(1):3–16, 2014.

[9] A.F. Constant and B.N. Tien. Germany's immigration policy and labor shortages. Technical Report 41, IZA Institute of Labor Economics, 2011.

[10] L. Gong, W. Wang, T. Wang, and C. Liu. Robotic harvesting of the occluded fruits with a precise shape and position reconstruction approach. *Journal of Field Robotics (JFR)*, 39(1):69–84, 2022.

[11] L. Guevara, M. Hanheide, and S. Parsons. Implementation of a human-aware robot navigation module for cooperative soft-fruit harvesting operations. *Journal of Field Robotics (JFR)*, 2023.

[12] K. He, G. Gkioxari, P. Dollár, and R. Girshick. Mask R-CNN. In *Proc. of the IEEE Intl. Conf. on Computer Vision (ICCV)*, 2017.

[13] K. He, X. Zhang, S. Ren, and J. Sun. Deep Residual Learning for Image Recognition. In *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[14] L. Horrigan, R.S. Lawrence, and P. Walker. How sustainable agriculture can address the environmental and human health harms of industrial agriculture. *Environmental Health Perspectives*, 110(5):445–456, 2002.

[15] P. Horton, S.A. Banwart, D. Brockington, G.W. Brown, R. Bruce, D. Cameron, M. Holdsworth, S. Lenny Koh, J. Ton, and P. Jackson. An agenda for integrated system-wide interdisciplinary agri-food research. *Food Security*, 9(2):195–210, 2017.

[16] H. Kang, H. Zhou, X. Wang, and C. Chen. Real-time fruit recognition and grasping estimation for robotic apple harvesting. *Sensors*, 20(19):5670, 2020.

[17] M.W. Khan, G.P. Das, M. Hanheide, and G. Cielniak. Incorporating spatial constraints into a bayesian tracking framework for improved localisation in agricultural environments. In *Proc. of the IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems (IROS)*, 2020.

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

[19] T. Kim, D.H. Lee, K.C. Kim, and Y.J. Kim. 2d pose estimation of multiple tomato fruit-bearing systems for robotic harvesting. *Computers and Electronics in Agriculture*, 211:108004, 2023.

[20] G. Kootstra, X. Wang, P.M. Blok, J. Hemming, and E. Van Henten. Selective harvesting robotics: current research, trends, and future directions. *Current Robotics Reports*, 2:95–104, 2021.

[21] T. Kozai, G. Niu, and M. Takagaki. *Plant factory: an indoor vertical farming system for efficient quality food production*. Academic press, 2019.

[22] L. Kugler. Addressing labor shortages with automation. *Communications of the ACM*, 65(6):21–23, 2022.

[23] S. LaValle. Rapidly-exploring random trees: A new tool for path planning. Technical report, Iowa State University, 1998.

[24] J. Le Louëdec and G. Cielniak. Key point-based orientation estimation of strawberries for robotic fruit picking. In *Proc. of the Intl. Conf. on Computer Vision Systems (ICVS)*, 2023.

[25] C. Lehnert, I. Sa, C. McCool, B. Upcroft, and T. Perez. Sweet Pepper Pose Detection and Grasping for Automated Crop Harvesting. In *Proc. of the IEEE Intl. Conf. on Robotics & Automation (ICRA)*, 2016.

[26] W. Lorensen and H. Cline. Marching Cubes: a High Resolution 3D Surface Construction Algorithm. In *Proc. of the Intl. Conf. on Computer Graphics and Interactive Techniques (SIGGRAPH)*, 1987.

[27] I. Loshchilov and F. Hutter. Decoupled weight decay regularization. In *Proc. of the Intl. Conf. on Learning Representations (ICLR)*, 2019.

[28] F. Magistri, N. Chebrolu, J. Behley, and C. Stachniss. Towards In-Field Phenotyping Exploiting Differentiable Rendering with Self-Consistency Loss. In *Proc. of the IEEE Intl. Conf. on Robotics & Automation (ICRA)*, 2021.

[29] F. Magistri, E. Marks, S. Nagulavancha, I. Vizzo, T. Läbe, J. Behley, M. Halstead, C. McCool, and C. Stachniss. Contrastive 3D Shape Completion and Reconstruction for Agricultural Robots using RGB-D Frames. *IEEE Robotics and Automation Letters (RA-L)*, 7(4):10120–10127, 2022.

[30] S. Marangoz, T. Zaenker, R. Menon, and M. Bennewitz. Fruit mapping with shape completion for autonomous crop monitoring. In *Proc. of the Intl. Conf. on Automation Science and Engineering (CASE)*, 2022.

[31] E. Marks, F. Magistri, and C. Stachniss. Precise 3D Reconstruction of Plants from UAV Imagery Combining Bundle Adjustment and Template Matching. In *Proc. of the IEEE Intl. Conf. on Robotics & Automation (ICRA)*, 2022.

[32] R. Menon, T. Zaenker, N. Dengler, and M. Bennewitz. Nbv-sc: Next best view planning based on shape completion for fruit mapping and reconstruction. In *Proc. of the IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems (IROS)*, 2023.

[33] T. Mueller-Sim, M. Jenkins, J. Abel, and G. Kantor. The robotanist: A ground-based agricultural robot for high-throughput crop phenotyping. In *Proc. of the IEEE Intl. Conf. on Robotics & Automation (ICRA)*, 2017.

[34] Y. Pan, F. Magistri, T. Läbe, E. Marks, C. Smitt, C. McCool, J. Behley, and C. Stachniss. Panoptic Mapping with Fruit Completion and Pose Estimation for Horticultural Robots. In *Proc. of the IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems (IROS)*, 2023.

[35] J.J. Park, P. Florence, J. Straub, R. Newcombe, and S. Lovegrove. DeepSDF: Learning Continuous Signed Distance Functions for Shape Representation. In *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[36] I. Pérez-Borrero, D. Marín-Santos, M.E. Gegúndez-Arias, and E. Cortés-Ancos. A fast and accurate deep learning method for strawberry instance segmentation. *Computers and Electronics in Agriculture*, 178:105736, 2020.

[37] G. Ren, T. Wu, T. Lin, L. Yang, G. Chowdhary, K. Ting, and Y. Ying. Mobile robotics platform for strawberry sensing and harvesting within precision indoor farming systems. *Journal of Field Robotics (JFR)*, 2023.

[38] J. Rong, P. Wang, T. Wang, L. Hu, and T. Yuan. Fruit pose recognition and directional orderly grasping strategies for tomato harvesting robots. *Computers and Electronics in Agriculture*, 202:107430, 2022.

[39] M. Ryan. Labour and skills shortages in the agro-food sector. *OECD Food, Agriculture and Fisheries Papers*, 189, 2023.

[40] I. Sa, C. Lehnert, A. English, C. McCool, F. Dayoub, B. Upcroft, and T. Perez. Peduncle Detection of Sweet Pepper for Autonomous Crop Harvesting - Combined Colour and 3D Information. *IEEE Robotics and Automation Letters (RA-L)*, 2(2):765–772, 2017.

[41] D. Schunck, F. Magistri, R. Rosu, A. Cornelißen, N. Chebrolu, S. Paulus, J. Léon, S. Behnke, C. Stachniss, H. Kuhlmann, and L. Klingbeil. Pheno4D: A spatio-temporal dataset of maize and tomato plant point clouds for phenotyping and advanced plant analysis. *PLOS ONE*, 16(8):1–18, 2021.

[42] Y. Shi, J.A. Thomasson, S.C. Murray, N.A. Pugh, W.L. Rooney, S. Shafian, N. Rajan, G. Rouze, C.L. Morgan, H.L. Neely, et al. Unmanned aerial vehicles for high-throughput phenotyping and agronomic research. *PLOS ONE*, 11(7):e0159781, 2016.

[43] A. Silwal, J.R. Davidson, M. Karkee, C. Mo, Q. Zhang, and K. Lewis. Design, integration, and field evaluation of a robotic apple harvester. *Journal of Field Robotics (JFR)*, 34(6):1140–1159, 2017.

[44] A. Tafuro, A. Adewumi, S. Parsa, G.E. Amir, and B. Debnath. Strawberry picking point localization ripeness and weight estimation. In *Proc. of the IEEE Intl. Conf. on Robotics & Automation (ICRA)*, 2022.

[45] N.K. Uppalapati, B. Walt, A.J. Havens, A. Mahdian, G. Chowdhary, and G. Krishnan. A berry picking robot with a hybrid soft-rigid arm: Design and task space control. In *Proc. of Robotics: Science and Systems (RSS)*, 2020.

[46] I. Vizzo, T. Guadagnino, J. Behley, and C. Stachniss. VDBFusion: Flexible and Efficient TSDF Integration of Range Sensor Data. *Sensors*, 22(3):1296, 2022.

[47] I. Vizzo, T. Guadagnino, B. Mersch, L. Wiesmann, J. Behley, and C. Stachniss. KISS-ICP: In Defense of Point-to-Point ICP – Simple, Accurate, and Robust Registration If Done the Right Way. *IEEE Robotics and Automation Letters (RA-L)*, 8(2):1029–1036, 2023.

[48] E. Vrochidou, V.N. Tsakalidou, I. Kalathas, T. Gkrimpizis, T. Pachidis, and V.G. Kaburlasos. An overview of end effectors in agricultural robotic harvesting systems. *Agriculture*, 12(8):1240, 2022.

[49] N. Wagner, R. Kirk, M. Hanheide, G. Cielniak, et al. Efficient and Robust Orientation Estimation of Strawberries for Fruit Picking Applications. In *Proc. of the IEEE Intl. Conf. on Robotics & Automation (ICRA)*, 2021.

[50] J. Wang, M. Rünz, and L. Agapito. Dsp-slam: Object oriented slam with deep shape priors. In *Proc. of the Intl. Conf. on 3D Vision (3DV)*, 2021.

[51] Y. Xiong, Y. Ge, L. Grimstad, and P.J. From. An autonomous strawberry-harvesting robot: Design, development, integration, and field evaluation. *Journal of Field Robotics*, 37(2):202–224, 2020.

[52] Y. Xiong, Y. Ge, Y. Liang, and S. Blackmore. Development of a prototype robot and fast path-planning algorithm for static laser weeding. *Computers and Electronics in Agriculture*, 142:494–503, 2017.

[53] T. Zaenker, C. Smitt, C. McCool, and M. Bennewitz. Viewpoint planning for fruit size and position estimation. In *Proc. of the IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems (IROS)*, 2021.

[54] S. Zakharov, W. Kehl, A. Bhargava, and A. Gaidon. Autolabeling 3d objects with differentiable rendering of sdf shape priors. In *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[55] F. Zhang, J. Gao, C. Song, H. Zhou, K. Zou, J. Xie, T. Yuan, and J. Zhang. Tpmv2: An end-to-end tomato pose method based on 3d key points detection. *Computers and Electronics in Agriculture*, 210:107878, 2023.