Strawberry picking point localization ripeness and weight estimation

Alessandra Tafuro, Adeayo Adewumi, Soran Parsa, Amir Ghalamzan E. and Bappaditya Debnath

***Abstract*— Labour shortage, difficulties in labour management, the digitalization of fruit production pipeline to reduce the fruit production costs have made robotic systems for selective harvesting of strawberries an important industry and academic research. One of the important components of such technologies yet to be developed is fruit picking perception. For picking strawberries, a robot needs to infer the location of picking points from the images of strawberries. Moreover, the size and weight of strawberries to be picked can help the robot to place the picked strawberries in proper punnets directly to be delivered to customers in supermarkets. This can save significant time and packing costs in packhouses. Geometry-based approaches are the most common approach to determine the picking point but they suffer from inaccuracies due to noise, occlusion, and varying shape and orientation of the berries. In contrast, we present two novel datasets of strawberries annotated with picking points, key-points (such as the shoulder points, the contact point between the calyx and flesh, and the point on the flesh farthest from the calyx), and the weight and size of the berries. We performed experiments with Detectron-2, which is an extended version of Mask-RCNN with key-points detection capability. The results show that the key-points detection approach works well for picking and grasping point localization. The second dataset also presents the dimensions and weight of strawberries. Our novel baseline model for weight estimation outperforms many state-of-the-art deep networks. The datasets and annotations are available at *https://github.com/imanlab/strawberry-pp-w-r-dataset*.**

I. Introduction

Developing robotic technologies for selective harvesting of high-value crops, such as strawberries, has been highly demanded because of different social, political and economical factors [1], such as labour shortage, recent COVID-19 pandemic. As such, private and public sectors have invested across the globe to develop robotic harvesting and commercialise the corresponding technologies in the last few years. Only EU invested above €5 million in BACHUS [2] project to develop robotic technology for grape and olive picking and above €4 million in SWEEPER [3] for sweet pepper harvesting robotic technology. Strawberry is among high-value crops with considerable selective harvesting cost in production. With a total retail value of $17 billions globally, above $1 billion is only the picking cost [4].

Nevertheless, there is not yet a commercially viable robotic technology available for selective harvesting of strawberries. Strawberries usually grow in dense clusters. This makes the robot perception challenging as some target fruits may be occluded by non-target fruits and leaves. Commercially available depth sensors, e.g, Realsense *D435i*, also make the perception challenging as they are designed for large objects 3-D perception and controlled lighting conditions. For small fruits under outdoor lighting, the depth maps are not precise. Detecting, segmenting, and localizing a ripe fruit to be picked in a complex cluster geometry, under outdoor lighting conditions make strawberry perception a very challenging problem.

Fig. 1. We introduce two novel datasets targeted towards robotic selective harvesting of strawberries. The datasets provide instance segmentation, "pluckability", key-points and weight information about the strawberries.

Moreover, weight estimation of the strawberries before picking can help sort the fruits in the correct punnet on the robot right after picking, that meets the weight legislation for immediate packaging which results in minimum touch to reduce the likelihood of bruising and reducing the added traditional packaging costs during post-harvest processing. The total cost of harvesting and post-harvesting processing is above 50% of strawberry production cost where almost half of this cost relates to the post-harvest processing and packaging.

Fruits weight estimation while on the plant and picking points localization (namely position and orientation) are needed perception components for a successful selective harvesting robot. Failure in these perception components yields degeneration in the performance of the entire robotic harvesting system. The contribution of this paper is manifold: (1) we present new datasets of strawberries which contain key-points useful for picking point localization and weight estimation; (2) we present an approach to picking point/orientation detection; (3) we present a weight estimation method. Our experimental results demonstrate the effectiveness of the proposed approach of determining picking point through strawberry segmentation and key-points detection (Figure 1). Also, the proposed novel baseline method for weight estimation outperforms several state-of-the-art (SOTA) deep networks.

Lincoln Institute for Agri-food Technology, University of Lincoln, LN25TS, Lincoln, United Kingdom AGhalamzanEsfahani@lincoln.ac.uk

II. Related Work

Different fruit detection and localisation approaches exist, for example classic computer vision (CV)-based operations and modern SOTA Convolutional Neural Networks (CNNs). While classical methods (including morphology, colour-based, thresholding and geometrical approaches) provide good performances, other SOTA approaches in CV and deep learning (DL) have been able to yield improved performance. For instance, Regional CNNs (RCNNs) yield a high degree of accuracy to perform a pixel-wise segmentation of fruit pixels [5]. Researchers have used information extracted from RCNNs with their own algorithm to improve the estimation of picking points [5], [6].

Rajendra et al. [7] converted the colour from RGB to HSI colour map to manually set a threshold for ripe strawberries. The authors also proposed diameter thresholding for peduncle detection of strawberries. Zhuang et al. [8] used automatic thresholding algorithms based on Otsu's thresholding method for more robust thresholding. Arefi et al. [9] used colour-based segmentation to remove background and keep the fruit blob. Instead of directly using colour, colour information can also be used with other features for a more robust approach. Tao et al. [10] used colour with geometric features for apple classification with GA-SVM. Lehner et al. [11] used colour-based segmentation with 3D parametric model-fitting for localisation of sweet peppers. Zhuang et al. [8] improved colour-based segmentation through iterative-retinax algorithm followed by Otsu's thresholding. Arefi et al. [9] used the well-known watershed algorithm to extract the morphology of tomatoes from binary images obtained by colour-thresholding. Huang et al. [12] used erosion and dilatation operations to refine strawberry from the colour segmented binary mask. Li et al. [13] use morphological operations for twig detection to prune fruitless twigs for litchi harvesting. Duran et al. [14] used the connected component algorithm to identify strawberry blobs. Hayashi et al. [15] used geometry of strawberry for calculating peduncle angle with respect to the vertical line for picking point localisation. [10] used a parameterised query of the spatial differences between a point and its adjacent area to form a Fast Point Feature Histogram (FPFH) descriptor. The FPFH descriptor formed a multidimensional histogram to describe the geometric properties within the K-neighborhood of a point. This offered the advantages of rotation invariance and good robustness under different sampling densities and noise levels [10].

Inability to generalise and being prone to noise are among the weaknesses of colour thresholding, geometry, morphology-based and other traditional approaches. CV engineers need to handcraft features and as the variation in data increases it becomes a cumbersome and infeasible task [16]. In recent years DL has been demonstrated to be superior for tasks such as segmentation [17] and key-points detection [18]. Thus, authors in selective harvesting have begun to adapt some of the DL techniques for fruit perception. Lamb et al. [19] used CNN for strawberry detection by optimising the network through input compression, image tiling, colour masking and network compression. Liu et al. [20] used CNN in combination with depth data to calculate the relative 3-D location of fruit. Similarly, Zhang et al. [21] used CNN for tomato classification. Spectral features with CNN is used for strawberry quality or ripeness detection [22]. CNN model is also used for pear bruise detection based on thermal images [23]. While CNN models perform well in image-specific tasks such as classification, for the pixel-wise understanding of images (semantic segmentation) RCNNs are more successful. Sa et al. [24] used bounding box detecting fruit using the fusion of faster RCNN, RGB and Infrared (IR) images. Liu et al. [25] used both YOLOv3 and mask RCNN (MRCNN) with ResNet-52 and ResNet-150 as backbone for bounding box detection for citrus fruit harvesting. Among the three models tested, MRCNN with ResNet-150 as backbone provided the best performance. Yu et al. [26] presented another example of MRCNN for selective harvesting. Moreover, the authors used MRCNN to determine the strawberry shapes. Then a geometrical algorithm was used to localise the picking point. Similarly, Ge et al. [5] used MRCNN to extract strawberry pixels. Then, the extracted strawberry pixels were combined with depth data, density-based clustering and Hough transformation to develop a more robust scene understanding. Similarly, Perez et al. [27] used MRCNN for strawberry segmentation for harvesting. Researchers have also used RCNN in combination with other methods to improve the overall accuracy. For instance, Liu et al. [6] used MRCNN with the logical green operator to improve the overall performance for Cucumber detection. Ganesh et al. [28] used a combination of HSV and RGB images to improve the overall performance for MRCNN for Orange detection. Yu et al. [26] first used a MRCNN to segment strawberry images and then used geometrical calculations to localise the picking point. The authors used two extreme points of a strawberry pixel to draw a horizontal line on the image frame. Then, the fruit axis was calculated based on the similarity of regions that would be divided by the fruit axis. The picking point was localised on the fruit axis based on statistics of strawberry shape.

We argue that, due to the widely varying shape of strawberries, such approaches for picking point localisation may fail in some cases and lacks the generalisation across different strawberries. Varying orientations of strawberry (upside-down, parallel to the gravity vector, or with an angle smaller than 90 degrees with respect to the gravity vector), occlusion and other factors may limit the success of the approach presented in [26]. DL-based key-points detection which has been very successfully applied in other domains, e.g. face landmark detection [29] or human-pose estimation [18], can be successfully applied to determine strawberry picking points, grasping points and orientation. However, this approach requires a strawberry dataset with key-points annotations. As shown in Table I, the existing datasets do not present strawberry key-points or picking points. In contrast, we present two new datasets containing key-points

and picking points information, which are two essential attributes for selective harvesting. Moreover, our first dataset present information that is useful for weight estimation of strawberries on the plant.

III. Datasets

In contrast to the existing works in which classic CV methods are used to determine picking points and suitability for picking, our approach includes SOTA MRCNN models. We collected two datasets to train our models: Dataset-1 is collected at new *15-acre* table-top strawberry glasshouse in Carrington, Lincolnshire, which is the latest addition to SOTA Dyson Farming's circular farming system [32]; Dataset-2 has been derived from the Strawberry Digital Images (SDI) [30].

Dataset-1 is a novel dataset that presents strawberries dimensions, weights, suitability for picking, instance segmentation, and key-points for grasping and picking action. The main purpose of this dataset is to facilitate autonomous robotic strawberry picking.

For each strawberry, the dataset presents five different key-points: picking point (PP), top, bottom points of fruit, left grasping point (LGP), and the right grasping point (LGP). While the PP indicates the position on the stem where the cutting action has to be performed, the left and right grasping points can provide reference to the end effector for grasping action. In addition, the dataset also contains annotation for instance segmentation for each of the strawberries. To determine the suitability of strawberries for harvesting each strawberry is labeled as 'pluckable'–ready to be picked– or '"unpluckable"'–not to be picked–. "unpluckable" strawberries include unripe, semi-, and over-ripe or rotten berries. The 'pluckable' category includes strawberries that are nearly ripe and perfectly ripe. The dataset contains a set of 532 strawberry sets (Table II). Each set has three colors, depth, and point cloud data of the same strawberry cluster from different distances. The farthest image captures the entire cluster whereas the nearest image focuses on one target strawberry in the cluster. In total, this dataset includes 1588 strawberries images (Table II). All the images have been captured with Intel Realsense RGB-D sensor *D435i*.

Dataset-2 is an enhancement of the SDI dataset [30]. SDI dataset contains a total of 3100 images. These are dense strawberry clusters that contain an average of 5.8 strawberries per image. We carefully annotated 10999 berries each with 5 different key-points. Moreover, we labeled "pluckability" (i.e. suitability to be picked) of all the strawberries. The strawberries not annotated for key-points are either severely occluded or are in an early flowering stage where a meaningful annotation is not possible.

IV. Proposed Approach

As seen from Table I, there are very few publicly available strawberry datasets. To our knowledge, these are the first strawberry datasets that present different aspects of strawberry data under farm conditions for selective harvesting. The first aspect is localization and orientation of strawberries for selective harvesting and grasping. Existing approaches use smaller datasets and use traditional algorithms such as geometry-based approaches [26], [27]. These algorithms often make several assumptions about the orientation of the strawberry or positioning of the peduncle or stem. However, strawberries can grow with different positions and orientations and can be of any shape as presented in the SDI dataset [30]. Thus, such assumptions are often not true and lead to inaccurate results. In contrast, our proposed approach relies on DL-based key-points estimation which has not yet been explored for picking point localization and determining orientation of strawberries. Currently available strawberry datasets (see Table I) are not suitable for picking point localization and determination of the orientation through key-points estimation. We demonstrate through our experimental results that the key-points regression approach based on MRCNN [17] works well for localizing the picking points of strawberries. Apart from localization of cutting points, the key-points are also pivotal in determining the orientation of strawberries which will be also used by the robot to approach the ripe strawberry. Similar to human pose estimation, we mark the key-points as visible or invisible. Thus, if the peduncle of a strawberry is occluded or facing away from the camera perspective, the MRCNN algorithm can be trained to estimate the point as invisible.

Dataset-1 presents the dimensions and weight of each strawberry in addition to the key-points. These novel features are not present in available datasets as seen in Table I. These novel features are unique and very much needed for future robotic systems that pick strawberries and place them directly into punnets according to their weights and sizes for delivering to supermarkets. Dryad et al. [31] presents the weights of strawberries only under controlled laboratory conditions with the calyx separated from the strawberries. In contrast, we present the weights of strawberries under farm conditions where it is difficult to obtain very accurate and consistent depth information, especially under sunlight. Our novel dataset presents more challenging and realistic perception scenario for weight estimation.

Normally, harvested strawberries are placed into punnets where the total weights of strawberries in a punnet are roughly required to be the same. Harvesters rely on their experience and scales to estimate the strawberry weights which are not accurate enough. Subsequently, a lot of effort is required to re-sort the strawberries into punnets which increases the production cost. Weighing strawberries after harvesting requires additional handling and maneuvering of strawberries which run the risk of bruising the delicate fruits. AI-based weight estimation has the potential to reduce production cost by estimating the weight of the strawberry before picking and detaching it from its plant [33], [34].

A. Segmentation and key-point detection

Our proposed approach includes Detectron-2 [35] for segmentation and key-points estimation. The Detectron-2 model is based on MRCNN [17] and has become the de facto standard for instance segmentation. It also has an

TABLE I

UNLIKE EXISTING DATASETS, THE PROPOSED DATASETS CONTAIN KEY-POINTS INFORMATION FOR LOCALIZATION OF STRAWBERRY PICKING AND GRASPING POSE. INSPIRED BY HUMAN POSE ESTIMATION, WE LOCALIZE PICKING POINT DURING APPROACHING PHASE OF END-EFFECTOR (EE) USING DEEP NETWORK-BASED KEY-POINT DETECTION. THIS IS VERY DIFFERENT FROM EXISTING APPROACHES WHICH RELY ON GEOMETRICAL OR COLOR FEATURES PRONE TO NOISE, OCCLUSION, ETC.

Datasets	#Images	#Berries	Depth	Segmentation	Farm	"Pluckability"	Key-points	Weight	Dimensions
SDI [30]	3100	17938	✗	✓	✓	✗	✗	✗	✗
Dryad [31]	35442	1611	✓	✗	✗	✗	✗	✓	✓
Ours (Dataset-1)	1588	2413	✓	✓	✓	✓	✓	✓	✓
Ours (Dataset-2)	3100	17938	✗	✓	✓	✓	✓	✗	✗

TABLE II

DETAILS OF THE PROPOSED DATASETS. KPS:KEY-POINTS

Dataset	#Images	#Pluck	#UnPluck	#KPS	#Weight
Dataset-1	1588	1757	656	2413	1910
Dataset-2	3100	3659	14279	10999	NA

added capability of key-points detection for human pose estimation. We adapted this key-points detection method integrated within Detectron-2 to estimate the strawberry key-points. The datasets' key-points, segmentation masks, strawberry categories ('pluckable' and '"unpluckable"') are converted to MSCOCO JSON format [36]. This MSCOCO JSON is the default format for feeding data into Detectron-2. It is also essential to recalculate the bounding box. Without the key-points, the bounding box aligns to the extremities of the segmentation mask. However, the PP key-point lies outside the segmentation mask of the strawberries and thus outside the bounding box. Because of the nature of MRCNN, a key-point outside bounding boxes is not detectable. Thus, the bounding boxes are expanded to accommodate all the key-points. We performed experiments with three backbone networks for Detectron-2, R50-FPN, X101, and X101-FPN. ResNeXt [37] (X101-FPN and X101) is more recent network which was introduced as an improvement to ResNet-50 [38] (R50-FPN). Section V-B discusses the results in details.

B. Weight Estimation

The weight of strawberries is proportional to size, shape, and density.

The instantaneous density of strawberries is also dependent on the sugar and water content of the strawberry which may not be feasible to be inferred from images alone. The higher the accuracy, the better the performance of a robot picker in terms of meeting the weight and packaging legislation. Nonetheless, the visual weight estimation is in addition to the weight measurement of punnets in post-harvesting handling. While minimizing handling of individual strawberries is the goal here, weight measurement of punnets does not impact the quality of individual strawberries by bruising them. Thus, estimating the weights of strawberries to a decent accuracy (80-90%) is the goal here so that they can be properly sorted into punnets while harvesting without the need for further manual processing [33].

a) Metrics: While there are standard protocols like top-1 and top-5 accuracy for classification, there exist various protocols like PCKh [39] and OKS [35] depending on the requirements for regression. To measure the accuracy of weight estimation, we propose the Percentage of Correct Weights (PCW) protocol. We measure the regression error in percentage with respect to the ground truth. If the error is within the desired tolerance, the inference is marked as accurate; otherwise, it is marked as false. The percentage of predictions within the tolerance values gives us the accuracy of a model. In the experiments, the tolerance levels are set at 0.05, 0.1, 0.15, 0.20, 0.25 and 0.30 which correspond to a range of 70-95% accuracy. The protocol is inspired by the Percentage of correct key-points (PCK) protocol used for human pose estimation [39].

b) NN Architectures: For strawberry weight estimation, we adapt several SOTA neural networks. The vision-based weight estimation is dependent on apparent size and images and thus it is essential to use the depth information. First, we extract all the instances of strawberries from each image with the help of segmentation through Detectron-2 [35]. The same segmentation mask is also applied to depth image. These two segmented images (color and depth) are then combined with the help of intrinsic camera calibration parameters of the Intel Real-sense *D435i* to reconstruct the point cloud. The result is a segmented point cloud that only contains the strawberry points. This is fed into PointConv [40], PointNet [41] and PointNet++ [42], which are well-known point cloud-based deep networks. For each of these networks, the final layer in the network has been replaced with a dense layer containing only one unit with a linear activation function for regressing the strawberry weights. Recently there has been an increased interest in graph-based neural networks since the introduction of graph neural networks by Kipf et al. [43]. Thus, we also experiment with graph-based networks, namely DGCN [44], GCN [43] and HGNN [45]. The graph dataset is obtained by firstly applying surface mesh reconstruction to the point clouds. Then, a fixed number of points have been sampled from the reconstructed surfaces. Finally, the k-nearest neighbor graph generation function has been exploited to get the final graphs, composed of nodes and edges. As nodes features, only a unitary number has been added to every node since every point has been considered with the same importance for weight estimation purposes. We also adapt the well-known CNN model, namely EfficientNet [46], for strawberry weight estimation. A two-stream architecture is used, where the first stream takes an RGB image of a single strawberry obtained through segmentation based on Detectron-2. The second stream is fed with the raw-depth image obtained after applying the same segmentation mask to the depth image. Thus in both streams pixels belonging to the same strawberry retain their original value, whereas

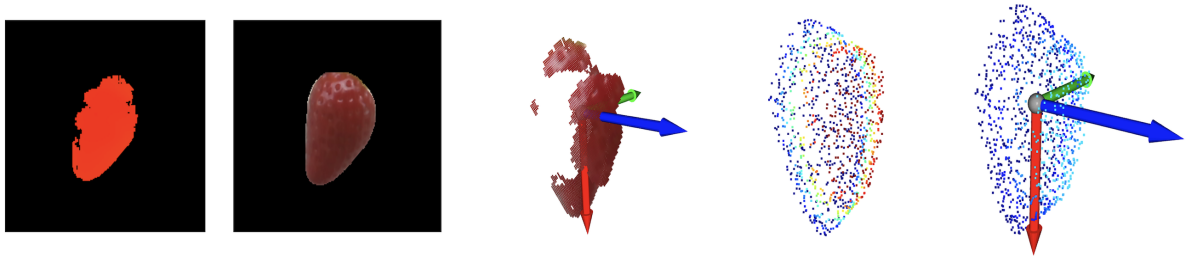

Fig. 2. From left (a): segmented depth (b): segmented color (c): reconstructed point cloud (d): convex hull (e): PCA. The weight estimation pipeline extracts point cloud through segmented color and depth images. The additional orientation information obtained through PCA improved the results.

non-strawberry pixels are zero. The final dense layer of the EfficientNet is removed and the outputs of both streams are concatenated. This concatenated layer is passed through a dense layer with linear activation.

c) Proposed approach: Along with the above approaches, we also present a novel, simple and effective baseline that outperforms the other SOTA networks. As shown in Figure 2, first the segmented images of individual strawberries are obtained through Detectron-2 [35]. The same segmentation mask is applied to the raw-depth image. As done for point cloud-based deep networks these segmented images are combined with the help of Realsense intrinsic camera calibration parameters to create a segmented point cloud (Figure 2c). The point cloud is not continuous and represents two depth layers. This is the case with original point clouds captured with Realsense. In realistic farm or outdoor scenes, the depth could range up to infinity. Thus, the resultant depth sensitivity (resolution) is low and unable to capture the relatively small depth gradient on the surface of a strawberry in a continuous manner. We also obtain the convex hull of the reconstructed point clouds (Figure 2d), which provides a more continuous surface. A principal component analysis (PCA) is performed on the convex hull to determine the direction of the largest extent (Figure 2e). A feature vector is created which contains the strawberry bounding box area obtained from the segmentation mask, the segmentation mask, the histogram of depth values, and the primary principal component. This bounding box area helps the model to learn the apparent size of strawberries. If a strawberry comes in different shapes and sizes, the area of the segmentation mask with respect to the bounding box helps the model to fine-tune the apparent size of the strawberry. Since the depth information is not reliable unlike color pixels, we take the histogram of the strawberry with 12 bins to help the model learn a more accurate depth representation. The bounding box and the apparent area of the strawberry can vary due to the different orientations of the strawberry. Thus, we argue that if the model is aware of the 3-D orientation of the strawberry, it can learn the relationship of the weight and the variation of apparent size caused by the orientation of the berry. Thus, the inclusion of the primary principal component improves the performance of the weight estimation model. The feature vector is used to train a Random Forest model [47] with Decision Trees which performs better than the deep models discussed earlier.

V. Results and Analysis

A. Implementation

The models and algorithms have been implemented in Python 3.7. The point cloud and graph-based neural networks have been implemented in Pytorch 1.8. For graph networks, Pytorch geometric version 1.7 has been used. The ImageNet pre-trained EfficientNets have been implemented through Tensorflow 2.1, with Keras wrapper. In order to help our neural networks generalize better, we performed data augmentation on the strawberry point clouds. The point cloud has been both translated and rotated randomly (the variations are within a region) before being used as input to the model. A random translation has been applied in both dimensions perpendicular to the depth value that is kept fixed as the strawberry depth location is a fundamental parameter for a correct weight regression. On the other hand, a random rotation has been applied around all three axes of the strawberry. After that, the coordinate values have been normalized and finally given as input to the different models for training. Also, for EfficientNets the color and depth images have been augmented through random translations and rotations. All models have been trained through an initial learning rate of 0.001 with Adam optimizer. For our baseline weight estimation model, Scikit Learn version 0.24.2 has been used for implementing the Random Forest, Support Vector Regression algorithms, and calculating PCA. After carefully fine-tuning the models the Random Forest model was found to be more accurate than the Support Vector model. Open3D has been used for manipulating point clouds. For segmentation and key-points detection, Detectron-2 [35] based on Pytorch 1.8 has been used. For Dataset-1, the Detectron-2 model is trained for 2200 iterations, and for the larger Dataset-2 the model is trained for 20000 iterations.

B. Segmentation and Key-point Detection

Table III summarises the results for segmentation and key-points detection of strawberries for both the datasets with Detectron-2 [35]. Although, our results demonstrate different backbones used in our experiments can produce consistent results across the dataset, ResNeXt based model performs better than ResNet-50 based model. The first two columns of Table III show segmentation Average Precision (AP) values for pluckable and ”unpluckable” berries separately. The sub-columns show AP for Intersection over Union (IoU), and thresholds of 0.5, 0.7, 0.9. The standard practice is to

TABLE III

SEGMENTATION AND KEY-POINTS DETECTION RESULTS

Dataset	Backbone	Segn Pluckable			Segn "unpluckable"			Key-points Pluckable			Key-points "unpluckable"		
		0.5	0.7	0.9	0.5	0.7	0.9	0.5	0.7	0.9	0.5	0.7	0.9
Dataset-1	R50-FPN	93.32	90.97	83.55	59.46	53.61	42.91	91.27	89.10	81.90	51.36	46.20	37.30
	X101	94.19	92.83	88.70	61.12	56.22	45.64	92.71	91.40	87.74	61.26	56.52	46.84
Dataset-2	X101-FPN	71.12	64.70	43.24	76.83	74.52	68.79	64.32	58.93	39.92	73.26	71.39	66.46
	X101	72.12	66.84	47.86	78.09	76.65	70.30	59.29	54.40	42.12	74.67	71.45	65.30

consider IoU threshold 0.5 [17], however, we also show up to IoU threshold 0.9. Using Dataset-2, our proposed models yield decent AP values for both pluckable and "unpluckable" strawberries at IoU threshold 0.5. However, as shown in Table III, the '"unpluckable"' berries in Dataset-2 significantly outnumber the 'pluckable' berries. This results in better segmentation performance of '"unpluckable"' berries. The performance drops significantly for stricter thresholds 0.7 and 0.9. This dataset represents berries in very dense clusters and thus the Dataset-2 is a very challenging dataset and has the potential to further advance the research in selective harvesting. On the other hand, Dataset-1 shows very reliable AP values for pluckable strawberries for both the backbones across IoU thresholds. With IOU threshold of 0.5, the Detectron-2 produces 93.32 (R50-FPN) and (X101-FPN) 94.19 AP values, while with a very strict IoU threshold of 0.9 the Detectron-2 provides AP of 83.55 and 88.70 AP with R50-FPN and X101-FPN, respectively. This shows that for selective harvesting the dataset can be reliably used. For Dataset-1, the performance of our models on '"unpluckable"' berries is comparatively less reliable as there are fewer samples of '"unpluckable"' berries in this dataset. However, from a selective harvesting perspective instance segmentation of 'pluckable' berries is more essential.

The results of the key-points detection expressed in terms of AP at different IoU thresholds are similar to segmentation. At each IoU threshold, we take the average results from 0.5, 0.3 and 0.1 OKS. OKS [35] is the standard performance metric used by Detectron-2 [35] and MSCOCO [36] for key-point detection. While the OKS threshold normally used is 0.5, 0.1 is a stricter threshold. The experimental results show that similarly to segmentation, the results are consistent across the two backbones although X101-FPN preforms slightly better. Also, the key-points detection for 'pluckable' berries is much better than '"unpluckable"' berries for Dataset-1. The results for Dataset-2 obtained comparing X101-FPN and X101 networks, provide a good baseline for future research.

C. Weight Estimation results

Figure 3 illustrates the result of strawberry weight estimation experiments. As discussed in earlier (Sec. IV-B) the goal here is to achieve 80-90% accuracy for selective harvesting. We use point cloud, graph and rgb+depth-based SOTA neural networks. The point cloud-based network, PointConv [40] provides only 15.5% (PCW @0.1) to 28.90% (PCW @0.2). Similarly, PointNet [41] gives us only around only 23.00% (PCW @0.1) to 41.90% (PCW @0.2) accuracy while PointNet++ [42] gives us only 27.23% (PCW @0.1) to

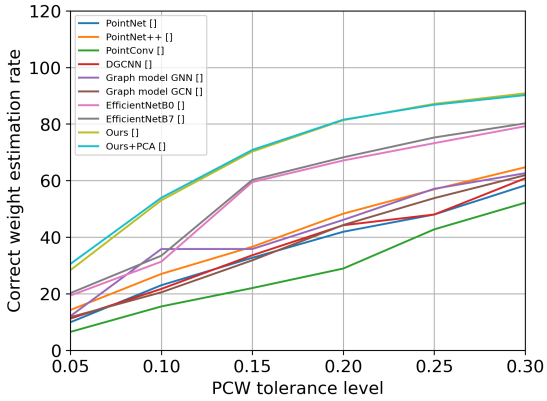

Fig. 3. Results of weight estimation at different PCW levels. The proposed method performs better than existing SOTA deep networks.

48.30% (PCW @0.2) accuracy. From Figure 3, one can also conclude that the graph-based networks achieve performance similar to point cloud-based networks. DGCN [44], GCN [43] and HGNN [45], provides 44.20, 44.30 and 46.10 % accuracy at PCW @0.20. We also used EfficientNet [46] which are a class of SOTA of CNN models that achieved best results on the ImageNet [48] dataset.

Figure 3 shows that the performance of the two-stream EfficientNets (Sec. IV-B) is better than point cloud and graph-based networks. EfficientNet BO and B7 give us much better results with 67.10 and 68.21 % accuracy at PCW @0.2, but still far from suitability for selective harvesting. This motivates us to propose our baseline method which performs much better than SOTA deep methods. The main novelty of the proposed algorithm is that by including the largest principal component using PCA, the method learns the orientation of the strawberry that helps with slightly more accurate weight estimation. The model performs 81.36 and 81.99% PCW @0.2 without and with PCA, respectively. By inclusion of PCA, the performance improves from 28.32% to 29.63% PCW @0.05. The graph (Fig 3), shows that PCA improves the performance across PCW levels.

VI. CONCLUSION

The paper presents two new datasets (including the RGB-D images from different perspectives, instance segmentation, key-points detection, dimension, and weights) useful for selective harvesting of strawberries. We also performed a series of experiments by deep NN models estimating the size, weight, pose, and picking point of strawberries. The results of our experiments show how our datasets facilitate key-points detection and the effectiveness of our approach to localize the picking point for selective harvesting. We also contribute a novel deep model for strawberry weight estimation that outperforms SOTA methods.

REFERENCES

[1] T. Duckett, S. Pearson, S. Blackmore, B. Grieve, P. Wilson, H. Gill, A. Hunter, and I. Georgilas, "Agricultural robotics: the future of robotic agriculture," in *"UK-RAS White Papers"*, 2018.

[2] "Bachus:mobile robotic platforms for active inspection and harvesting in agricultural areas," in *https://cordis.europa.eu/project/id/871704*, [accessed on 07.08.2021].

[3] "Sweeper: Sweet pepper harvesting robot," in *https://cordis.europa.eu/project/id/644313*, [accessed on 07.08.2021].

[4] "Strawberry picking cost ," in *https://www.producebluebook.com/2020/11/25/strawberries-in-2020*, [accessed on 07.08.2021].

[5] Y. Ge, Y. Xiong, G. L. Tenorio, and P. J. From, "Fruit localization and environment perception for strawberry harvesting robots," *IEEE Access*, vol. 7, pp. 147 642–147 652, 2019.

[6] X. Liu, D. Zhao, W. Jia, W. Ji, C. Ruan, and Y. Sun, "Cucumber fruits detection in greenhouses based on instance segmentation," *IEEE Access*, vol. 7, pp. 139 635–139 642, 2019.

[7] P. Rajendra, N. Kondo, K. Ninomiya, J. Kamata, M. Kurita, T. Shiigi, S. Hayashi, H. Yoshida, and Y. Kohno, "Machine vision algorithm for robots to harvest strawberries in tabletop culture greenhouses," *Engineering in Agriculture, Environment and Food*, vol. 2, no. 1, pp. 24–30, 2009.

[8] J. Zhuang, C. Hou, Y. Tang, Y. He, Q. Guo, Z. Zhong, and S. Luo, "Computer vision-based localisation of picking points for automatic litchi harvesting applications towards natural scenarios," *Biosystems Engineering*, vol. 187, pp. 1–20, 2019.

[9] A. Arefi, A. M. Motlagh, K. Mollazade, R. F. Teimourlou, *et al.*, "Recognition and localization of ripen tomato based on machine vision," *Australian Journal of Crop Science*, vol. 5, no. 10, p. 1144, 2011.

[10] Y. Tao and J. Zhou, "Automatic apple recognition based on the fusion of color and 3d feature for robotic fruit picking," *Computers and electronics in agriculture*, vol. 142, pp. 388–396, 2017.

[11] C. Lehnert, A. English, C. McCool, A. W. Tow, and T. Perez, "Autonomous sweet pepper harvesting for protected cropping systems," *IEEE Robotics and Automation Letters*, vol. 2, no. 2, pp. 872–879, 2017.

[12] Z. Huang, S. Wane, and S. Parsons, "Towards automated strawberry harvesting: Identifying the picking point," in *Annual Conference Towards Autonomous Robotic Systems*. Springer, 2017, pp. 222–236.

[13] J. Li, Y. Tang, X. Zou, G. Lin, and H. Wang, "Detection of fruit-bearing branches and localization of litchi clusters for vision-based harvesting robots," *IEEE Access*, vol. 8, pp. 117 746–117 758, 2020.

[14] A. Durand-Petiteville, S. Vougioukas, and D. C. Slaughter, "Real-time segmentation of strawberry flesh and calyx from images of singulated strawberries during postharvest processing," *Computers and electronics in agriculture*, vol. 142, pp. 298–313, 2017.

[15] S. Hayashi, K. Shigematsu, S. Yamamoto, K. Kobayashi, Y. Kohno, J. Kamata, and M. Kurita, "Evaluation of a strawberry-harvesting robot in a field test," *Biosystems engineering*, vol. 105, no. 2, pp. 160–171, 2010.

[16] N. O'Mahony, S. Campbell, A. Carvalho, S. Harapanahalli, G. V. Hernandez, L. Krpalkova, D. Riordan, and J. Walsh, "Deep learning vs. traditional computer vision," in *Science and Information Conference*. Springer, 2019, pp. 128–144.

[17] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2961–2969.

[18] Z. Cao, G. Hidalgo, T. Simon, S.-E. Wei, and Y. Sheikh, "Openpose: realtime multi-person 2d pose estimation using part affinity fields," *IEEE transactions on pattern analysis and machine intelligence*, vol. 43, no. 1, pp. 172–186, 2019.

[19] N. Lamb and M. C. Chuah, "A strawberry detection system using convolutional neural networks," in *2018 IEEE International Conference on Big Data (Big Data)*. IEEE, 2018, pp. 2515–2520.

[20] X. Liu, S. W. Chen, S. Aditya, N. Sivakumar, S. Dcunha, C. Qu, C. J. Taylor, J. Das, and V. Kumar, "Robust fruit counting: Combining deep learning, tracking, and structure from motion," in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2018, pp. 1045–1052.

[21] L. Zhang, J. Jia, G. Gui, X. Hao, W. Gao, and M. Wang, "Deep learning based improved classification system for designing tomato harvesting robot," *IEEE Access*, vol. 6, pp. 67 940–67 950, 2018.

[22] Z. Gao, Y. Shao, G. Xuan, Y. Wang, Y. Liu, and X. Han, "Real-time hyperspectral imaging for the in-field estimation of strawberry ripeness with deep learning," *Artificial Intelligence in Agriculture*, 2020.

[23] X. Zeng, Y. Miao, S. Ubaid, X. Gao, and S. Zhuang, "Detection and classification of bruises of pears based on thermal images," *Postharvest Biology and Technology*, vol. 161, p. 111090, 2020.

[24] I. Sa, Z. Ge, F. Dayoub, B. Upcroft, T. Perez, and C. McCool, "Deepfruits: A fruit detection system using deep neural networks," *Sensors*, vol. 16, no. 8, p. 1222, 2016.

[25] Y.-P. Liu, C.-H. Yang, H. Ling, S. Mabu, and T. Kuremoto, "A visual system of citrus picking robot using convolutional neural networks," in *2018 5th international conference on systems and informatics (ICSAI)*. IEEE, 2018, pp. 344–349.

[26] Y. Yu, K. Zhang, L. Yang, and D. Zhang, "Fruit detection for strawberry harvesting robot in non-structural environment based on mask-rcnn," *Computers and Electronics in Agriculture*, vol. 163, p. 104846, 2019.

[27] 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*, vol. 178, p. 105736, 2020.

[28] P. Ganesh, K. Volle, T. Burks, and S. Mehta, "Deep orange: Mask r-cnn based orange detection and segmentation," *IFAC-PapersOnLine*, vol. 52, no. 30, pp. 70–75, 2019.

[29] Z. Zhang, P. Luo, C. C. Loy, and X. Tang, "Facial landmark detection by deep multi-task learning," in *European conference on computer vision*. Springer, 2014, pp. 94–108.

[30] 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*, vol. 178, p. 105736, 2020. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0168169920300624

[31] A. Durand-Petiteville, D. Sadowski, and S. Vougioukas, "A strawberry database: Geometric properties, images and 3d scans," https://datadryad.org/stash/dataset/doi:10.25338/B8V308, 2018.

[32] "Dyson farming and state of the art strawberry glasshouse," in *https://dysonfarming.com/strawberries/*, [accessed on 07.08.2021].

[33] M. Mousavi and V. Surya Prasath, "On the feasibility of estimating fruits weights using depth sensors," in *4 th International Congress of Developing Agriculture, Natural Resources, Environment and Tourism of IranAt: Tabriz Islamic Art University In cooperation with Shiraz University and Yasouj University, Iran*, 2019.

[34] B. Zhang, N. Guo, J. Huang, B. Gu, and J. Zhou, "Computer vision estimation of the volume and weight of apples by using 3d reconstruction and noncontact measuring methods," *Journal of Sensors*, vol. 2020, 2020.

[35] Y. Wu, A. Kirillov, F. Massa, W.-Y. Lo, and R. Girshick, "Detectron2," https://github.com/facebookresearch/detectron2, 2019.

[36] T.-Y. Lin, M. Maire, S. J. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *ECCV*, 2015.

[37] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1492–1500.

[38] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*, 2016, pp. 770–778.

[39] A. Newell, K. Yang, and J. Deng, "Stacked hourglass networks for human pose estimation," in *European conference on computer vision*. Springer, 2016, pp. 483–499.

[40] W. Wu, Z. Qi, and L. Fuxin, "Pointconv: Deep convolutional networks on 3d point clouds," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 9621–9630.

[41] C. R. Qi, H. Su, K. Mo, and L. J. Guibas, "Pointnet: Deep learning on point sets for 3d classification and segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 652–660.

[42] C. Qi, L. Yi, H. Su, and L. J. Guibas, "Pointnet++: Deep hierarchical feature learning on point sets in a metric space," 2017.

[43] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," in *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*, 2017.

[44] Y. Wang, Y. Sun, Z. Liu, S. E. Sarma, M. M. Bronstein, and J. M.

Solomon, "Dynamic graph cnn for learning on point clouds," *Acm Transactions On Graphics (tog)*, vol. 38, no. 5, pp. 1–12, 2019.

[45] C. Morris, M. Ritzert, M. Fey, W. L. Hamilton, J. E. Lenssen, G. Rattan, and M. Grohe, "Weisfeiler and leman go neural: Higher-order graph neural networks," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, 2019, pp. 4602–4609.

[46] M. Tan and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *International Conference on Machine Learning*. PMLR, 2019, pp. 6105–6114.

[47] T. K. Ho, "Random decision forests," in *Proceedings of 3rd International Conference on Document Analysis and Recognition*, vol. 1, 1995, pp. 278–282 vol.1.

[48] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, *et al.*, "Imagenet large scale visual recognition challenge," *IJCV*, vol. 115, no. 3, 2015.