arXiv:2309.13669v1 [cs.RO] 24 Sep 2023

Autonomous Apple Fruitlet Sizing with Next Best View Planning

Harry Freeman and George Kantor

Abstract— **In this paper, we present a next-best-view planning approach to autonomously size apple fruitlets. State-of-the-art viewpoint planners in agriculture are designed to size large and more sparsely populated fruit. They rely on lower resolution maps and sizing methods that do not generalize to smaller fruit sizes. To overcome these limitations, our method combines viewpoint sampling around semantically labeled regions of interest, along with an attention-guided information gain mechanism to more strategically select viewpoints that target the small fruits' volume. Additionally, we integrate a dual-map representation of the environment that is able to both speed up expensive ray casting operations and maintain the high occupancy resolution required to informatively plan around the fruit. When sizing, a robust estimation and graph clustering approach is introduced to associate fruit detections across images. Through simulated experiments, we demonstrate that our viewpoint planner improves sizing accuracy compared to state of the art and ablations. We also provide quantitative results on data collected by a real robotic system in the field.**

I. Introduction

With population growth, climate change, and labor shortages on the rise, the global food supply is continuously facing increasing pressure. Agriculturalists are actively looking to integrate robotics-based solutions to reduce the time and effort required to perform labor-intensive tasks. Among these solutions is autonomous crop inspection, which is integral for precision agriculture. This is because it enables farmers to make smarter and more informed decisions for a variety of downstream processes, including quality control, disease and pest management, pruning, and harvesting. Apple growers, for example, size apple fruitlets[1] in order to determine when to apply chemical thinners to their crops in order to maintain a more consistent yearly harvest [1]. The current practice of manually obtaining sizes is labor-intensive, as calipers are used to record the diameters of approximately 500 fruitlets per varietal at a minimum of twice per thinning application. As a result, apple growers are looking to adopt autonomous solutions to improve efficiency and reduce cost.

Sizing fruit requires accurate spatial reasoning about the plants. This is often difficult to obtain, as many fruits grow in occluded environments. In the case of apple fruitlets (Fig. 1), the fruits are only 7mm-14mm in diameter, they are surrounded by branches and leaves, and often fruitlets in the same cluster[2] will occlude one another. Therefore, it is insufficient to use fixed camera systems attached to either stationary [2] or mobile [3] bases, as images of the fruit cannot be consistently captured. It is much more reliable to use active perception-based methods, where mobile cameras attached to robotic arms are free to move around, avoid occlusions, and capture more semantically meaningful images. This is commonly achieved using next-best-view (NBV) planning, where information stored inside a volumetric map is used to determine the next best camera pose that maximizes information gain.

Fig. 1. Left: Stereo camera attached to a 7 DoF robotic arm. Right: Example of a fruitlet cluster with AprilTag hung in close proximity for identification.

However, state-of-the-art NBV planners in agriculture are dedicated towards sizing larger fruit that are sparsely populated. They rely on expensive ray casting operations and shape completion methods that do not generalize to smaller fruit sizes. To overcome these limitations, we present an NBV planning approach that utilizes regions of interest for viewpoint sampling and an attention-guided mechanism for calculating information gain. In addition, we integrate a dual-map representation of the environment that significantly speeds up ray casting while maintaining finer occupancy information. To address the challenges of data association in the presence of wind and sensor error, we introduce a robust estimation and graph clustering approach. We demonstrate that our planning method improves sizing accuracy compared to a state-of-the-art viewpoint planner used in agriculture. Although our target application is sizing apple fruitlets, our method could be adapted to size other fruits and plants of comparable size. Our main contributions are

- A novel next-best-view planner that uses coarse and fine occupancy maps, region of interest targeted viewpoint sampling, and an attention-guided information gain metric to capture images of smaller fruit.
- A robust estimation and graph clustering approach to associate fruit detections across images in the presence of wind and sensor error.
- Quantitative evaluation on data collected by a real robotic system in a commercial apple orchard.

H. Freeman and G. Kantor are with Carnegie Mellon University, Pittsburgh PA, USA {hfreeman, kantor}@cs.cmu.edu

[1] A fruitlet is a young apple formed on the tree shortly after pollination

[2] A cluster is a group of fruitlets that grow out of the same bud

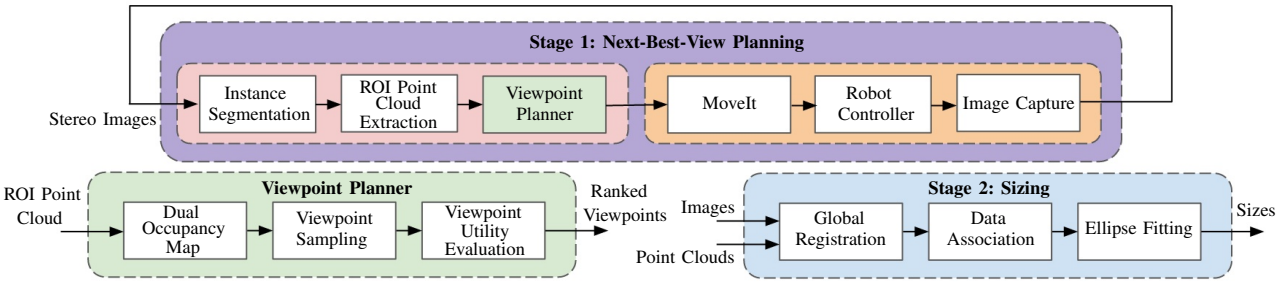

Fig. 2. Overview of our next-best-view planning and sizing pipeline.

II. Related Work

There have been previous works dedicated towards viewpoint planning in agriculture. Few approaches rely on sensor information without maintaining a volumetric map of the environment. Lehnert *et al.* [4] use a 3D camera array to size a target fruit of interest. They find the next best view by computing a local gradient and determining the direction in which the visibility of the target increases. While this method is effective at avoiding local occlusions, it does not extend globally when multiple fruit need to be sized. Roy *et al.* [5] present a viewpoint planning approach to count the number of apples on a tree branch. They determine the next camera pose by maximizing entropy over a set of world hypotheses informed by previous images. However, they make simplified assumptions about the relative positions and occlusion levels of the apples, which does not extend to smaller fruit that are more easily displaced and obscured by leaves.

More commonly, volumetric-based planning approaches have been adopted in agriculture, where occupancy information stored inside a volumetric map is used to guide the viewpoint planning process. The information gain for viewpoint candidates is often calculated by casting rays through the volumetric map. Wu *et al.* [6] use the volumetric map as input into a deep learning-based Point Completion Network [7] to estimate the unobserved points of a plant. The points are then used to sample viewpoint candidates. While deep learning-based approaches work for phenotyping plants, the absence of real and synthetic datasets makes them unadaptable to sizing uncommon fruit, such as apple fruitlets. Non-deep learning-based based methods often utilize spatial knowledge of semantically labeled regions of interest (ROIs). In the work of [8], detected ROIs in the form of segmented apples are used to estimate information gain. The information gain is used as input into a Decentralized Monte Carlo Tree Search [9] planner that outputs a sequence of viewpoints for multiple robot arms. However, computing a sequence of viewpoints is computationally expensive. As well, viewpoints are locally sampled without utilizing the global map. Zaenker *et al.* [10] address these issues by introducing ROI targeted sampling for NBV planning. Viewpoint targets are sampled from frontiers in the vicinity of known ROIs and are used to determine the next best view. They later integrate the work of [4] to combine local and global viewpoint planning to size capsicums [11]. While the computational cost is less, these methods still rely on expensive ray casting operations when calculating information gain, which are slow when performed at the resolution required to map millimeter-sized fruit. The authors of [12] introduce a viewpoint dissimilarity metric based on autonomous shape completion to eliminate ray casting from the information gain formulation. However, shape completion methods tend to perform poorly on small fruit. This is because it is difficult to capture enough of the fruit's surface. As well, wind and sensor error have a more significant effect, making the tasks of point cloud registration between frames and fruit clustering much more challenging.

Burusa *et al.* [13] propose an attention-guided NBV planning approach. Attention regions in the form of 3D bounding boxes are defined around different plant components and are used to restrict which voxels contribute to the information gain. The authors demonstrate that this attention mechanism leads to improved results with regards to both reconstruction accuracy and speed. However, the location and size of the attention regions were assumed to be prior knowledge, which is not the case when sizing fruit in unknown environments.

III. System Overview

An overview of our system can be seen in Fig. 2. Our pipeline is composed of two stages. In the first stage, next-best-view planning is used to determine camera poses that improve spatial knowledge of the scene. An AprilTag [14] is hung in close proximity to each cluster for identification. Images are captured using an illumination-invariant flash stereo camera [15] attached to the end of a 7 DoF robotic arm consisting of a UR5 and linear slider [16]. Examples of our robotic system and tagged fruitlet cluster can be seen in Fig. 1. When planning, the fruitlets are first segmented using a Mask-RCNN [17] instance segmentation network, and disparities are extracted using RAFT-Stereo [18]. The segmentations and disparities are used to build a point cloud of the scene with ROI information. The viewpoint planner updates a dual-map representation of the environment, which is used to sample candidate viewpoints and determine the next best end-effector pose based on expected utility. A path is planned to the pose using the MoveIt framework [19], which is executed by the robot controller. The process repeats until the specified planning duration is exceeded.

Sizing is performed once viewpoint planning is complete. To account for re-projection error resulting from sensor noise and wind, all point clouds are globally registered using a robust estimation method. Data association between images is then performed using Highly Connected Subgraph Clustering [20] to account for outliers and spurious detections. Lastly, fruitlets are sized using a 2D photogrammetric method.

IV. Viewpoint Planning

Our viewpoint planner combines ROI targeted sampling and attention mechanisms from previous works [10], [13]. We also utilize a dual-map representation of the environment in the form of coarse and fine octrees. This representation allows us to reduce the cost of ray casting while maintaining finer map resolutions within attention boundaries.

A. Workspace and Sampling Tree Generation

Our viewpoint planner uses both workspace and sampling trees to generate viewpoint candidates. The workspace tree defines the valid end-effector positions the robot is able to reach, while the sampling tree identifies the regions of space that viewpoint targets should be sampled from. To generate the workspace tree, ten million randomly generated joint configurations were sampled in simulation.

Because fruitlets are sized in clusters, we need a sampling tree large enough to encompass every fruitlet in the cluster, but not so large such that it includes additional fruitlets. To extract the sampling tree, we use a density-based clustering approach inspired from [21]. First, an initial image is captured to detect the AprilTag and nearby fruitlets. We lift the fruitlet centers to 3D and create a density map by treating each centroid as a unit-impulse and applying a 3D Gaussian filter to smooth the space around it. Cluster centroids are found by extracting local maxima in the density map, and each fruitlet is assigned to the closest cluster based on Euclidean distance. The cluster whose local maximum is closest to the AprilTag is determined to be the target cluster, and the sampling tree is constructed by fitting a sphere around the corresponding fruitlets. An example of the process can be seen in Fig. 3. We define the volume spanned by the sampling as the Attention Region. This is used to calculate information gain as described in Section IV-D.

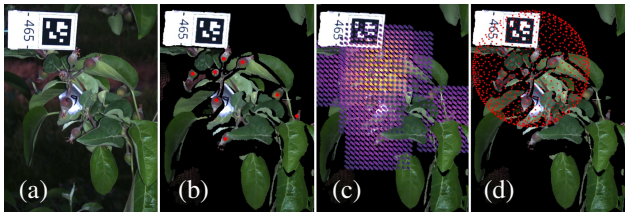

Fig. 3. (a) Original RGB image with AprilTag near cluster. (b) Extracted point cloud with fruitlet centroids. (c) Density map created by smoothing volume around centroids. (d) Sampling tree created from local maxima.

B. Dual-Map Representation

While libraries such as Octomap [22] have been designed to speed up expensive ray casting operations with optimized logic and multi-threading, these operations are still slow when performed at millimeter-level resolutions required to map small fruit. To overcome this, we take advantage of the fact that ROIs are usually confined to certain regions of space. We do this by maintaining two maps of the environment: a coarse octree that stores occupancy information at lower resolution, and a fine octree that stores both occupancy and ROI information at higher resolution. The coarse octree spans the entire observation space and is used to approximate the occupancy of voxels and detect occlusions outside the Attention Region. In contrast, the fine octree is restricted to within the Attention Region, and is used to identify which voxels are in the vicinity of ROIs. This allows the planner to more efficiently evaluate occupancy and occlusions when viewpoint planning, while providing sufficient resolution to map the fruit. We use the OctoMap framework [22] and the ROI-extended implementation presented by Zaenker *et al.* [10] for the coarse and fine octrees respectively.

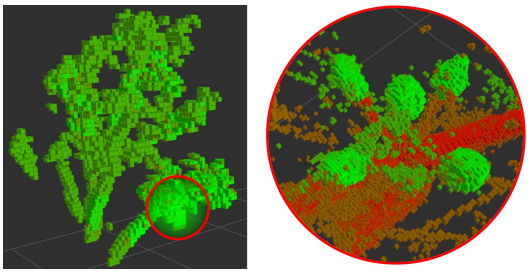

Fig. 4. Dual-map representation. Left: Coarse octree that stores occupancy information and spans the entire observation space. Right: Fine octree that stores occupancy and ROI information (green) within the Attention Region.

When ray casting, for both updating the environment model and calculating information gain, rays are cast through nodes in the coarse octree outside the Attention Region, and the fine octree inside the Attention Region, as demonstrated in Fig. 5. This significantly improves the computational cost of ray casting by reducing the number of traversed voxels when using higher map resolutions.

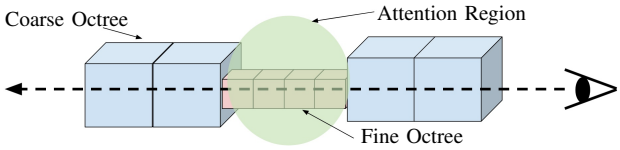

Fig. 5. Dual-map ray casting implementation. When ray casting, the coarse and fine maps are used outside and inside the Attention Region respectively.

While this approach could be implemented using a single octree, maintaining two maps provides more fine-grained control over the planning resolutions, while adding trivial additional overhead. Our implementation also easily scales to include multiple fine octrees for future work on more complete global coverage, as discussed in Section VII.

C. Viewpoint Sampling

Our viewpoint sampling method is inspired by ROI targeted sampling presented in [10]. First, ROI frontier voxels within the sampling tree are identified and used as target candidates. A frontier is the region between an empty voxel and unknown space, and an ROI frontier is a frontier in the vicinity of a known ROI. To determine ROI frontiers, the 6-neighborhoods of all ROIs are checked for free nodes. For all resulting free nodes, their 6-neighborhoods are checked for unknown neighbors. All free nodes that have at least one unknown neighbor are used as viewpoint targets. For each viewpoint target, candidate viewpoints are sampled from a partial Fibonacci sphere centered around the target at a specified sensor distance. Viewpoints that lie outside the workspace tree are discarded.

D. Viewpoint Evaluation

The information gain (IG) is estimated for all candidate viewpoints. For each viewpoint, multiple rays spanning the field of view of the camera are cast from the viewpoint along the direction towards the target using the method from Section IV-B. Rays terminate when they encounter an occupied voxel or exceed a maximum distance.

Our IG metric is an attention-guided version of the Unobserved Voxel IG presented by [10]. The IG for ray $r \in R$ is the number of unknown voxels along the ray that lie in the Attention Region $N_{A,u,r}$ divided by the total number of nodes along the ray that are inside the Attention Region $N_{A,r}$. The IG is averaged across all rays (Eq. 1 and 2). This attention-based approach ensures that only unknown voxels that lie within our sizing area of interest contribute to the information gain, which results in more informative planning.

$$IG_r = \begin{cases} \frac{N_{A,u,r}}{N_{A,r}}, & \text{if } N_{A,r} > 0 \\ 0, & \text{otherwise} \end{cases} \tag{1}$$

$$IG = \frac{1}{|R|} \sum_{r \in R} IG_r \tag{2}$$

We also compute a cost C to move to the viewpoint. This is the Euclidean distance between the position of the viewpoint and the current camera position. The Euclidean distance is used because computing the joint trajectory for every viewpoint is time-consuming. The cost is scaled by a constant α, and the final utility of the viewpoint is

$$U = IG - \alpha \cdot C \tag{3}$$

A max-heap is created from the viewpoints with order determined by their utility. The planner iterates through the heap until a viewpoint is found with a utility greater than a pre-determined threshold and that the motion planner can find a successful path to. If no viewpoints are left in the heap, new viewpoints are sampled.

V. Sizing

Prior methods size fruit using either occupancy information from the volumetric map [10] or automated shape completion methods [12], [23]. Neither of these approaches adapt well to smaller fruit. The former has limited precision, and the second is unreliable as it is difficult to capture enough of the fruit's surface. Additionally, both use simple Euclidean clustering methods that do not extend to fruits that are close to each another. Instead, we look to 2D photogrammetric approaches, which have been used to size small fruit in agriculture [24], [25], [26], [27]. These methods directly estimate the widths and heights of fruit by fitting simple 2D geometric shapes. The challenge is associating fruitlet detections across each image so that the least occluded image can be used to size each fruit. We tackle the data association problem by first globally registering the point clouds using a robust estimation method, and solving for associations using Highly Connected Subgraph Clustering [20].

A. Global Registration

In each frame, there is an offset in the re-projected point cloud as a result of both sensor noise and wind, as visualized in Fig. 6. Qualitatively, we have observed translational shifts in the point clouds typically within the range of 3mm-5mm, but up to as large as 1cm. While this error may be negligible for larger fruit, is is non-trivial when sizing apple fruitlets, which are often only a few millimeters apart.

Fig. 6. Images (a) and (b) with similar camera poses have a noticeable offset in the point clouds (c) as a result of sensor noise and wind.

Using local methods for point cloud registration, such as Iterative Closest Point, will often fail as a result of the sparse viewpoints and inconsistent structure in the point clouds. Similarly, directly solving for partial assignment using methods such as the Hungarian Algorithm [28] will not work because occlusions prevent every fruitlet from being detected in every image.

To globally register the point clouds, we adopt a robust estimation approach. Robust estimation is a popular method used in Simultaneous Localization and Mapping to reduce the effect of outliers in the assumed model. In our case, outliers represent inconsistent detections across images as a result of occlusions and false positive detections. We use robust estimation to globally register all point clouds simultaneously. This is advantageous as fruitlet detections across the span of all images are more globally consistent than detections between individual frames.

We formulate the task as a nonlinear least-squares problem. The objective is to find a transformation $\mathbf{T_i}$ for each frame $i \in \{1, ..., N\}$ that correctly aligns the point clouds. For each image $\mathbf{I}_i$, the centroid $c_{ij} \in \mathbb{R}^3$ of every fruitlet $s_{ij} \in \mathbf{S_i}$, $j \in \{1, ..., M_i\}$ is extracted, and we solve for the set of transformations $\mathbf{T} = \{\mathbf{T_1}, ..., \mathbf{T_N}\}$ that minimizes

$$\underset{\mathbf{T}}{\text{minimize}} \frac{1}{2} \sum_{i=1}^{N} \sum_{j=i+1}^{N} \sum_{k=1}^{M_i} \sum_{l=1}^{M_j} \rho(||\mathbf{T_i} c_{ik} - \mathbf{T_j} c_{jl}||^2) \tag{4}$$

where ρ is a scaled arctan function. The optimization problem is solved using a trust-region reflective algorithm [29], [30]. After solving for the refined transformations $\mathbf{T}$, the point clouds correctly align, as shown in Fig. 7.

B. HCS Clustering

After globally registering the point clouds, we can associate the fruitlet detections across images. We need the association method to be robust to spurious and missed detections. To accomplish this, we utilize Highly Connected

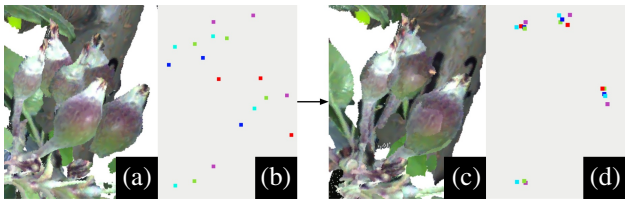

Fig. 7. (a) and (c) Point clouds for two frames before and after global registration. (b) and (d) Fruitlet centroids before and after global registration for all frames where each color represents detections in a single image.

Subgraph (HCS) clustering [20]. HCS clustering is a recursive graph-clustering method that partitions a graph into subgraphs by taking the minimum cut. If the number of edges in the minimum cut is greater than half the number of nodes in the graph, the graph is determined to be highly connected and is a cluster. Otherwise, the edges in the minimum cut are removed, and the process repeats for each new subgraph.

We construct a graph $G(V, E)$ consisting of nodes V and edges E. We treat the corrected world-transformed centroid of each detected fruitlet in each image as a node.

$$V = \{\mathbf{T_i} c_{ij} \forall\ i \in \{1, ..., N\} \text{ and } j \in \{1, ..., M_i\}\} \quad (5)$$

An edge e_{ij} is created between nodes v_i and v_j if i) the detections corresponding to the centroids are from different images, and ii) the Euclidean distances between the nodes is less than a threshold τ.

$$E = \{e_{ij} \forall\ i, j \text{ s.t. } \mathrm{Im}(i) \neq \mathrm{Im}(j) \wedge \|v_i - v_j\| < \tau\} \quad (6)$$

$\mathrm{Im}(i)$ represents the image corresponding to node v_i. The first edge requirement enforces that two detections in the same image cannot associate to the same fruitlet. The second edge requirement limits only reasonable associations. The graph is clustered using HCS clustering, and each resulting cluster represents a fruitlet association. A visualization of the process can be seen in Fig. 8.

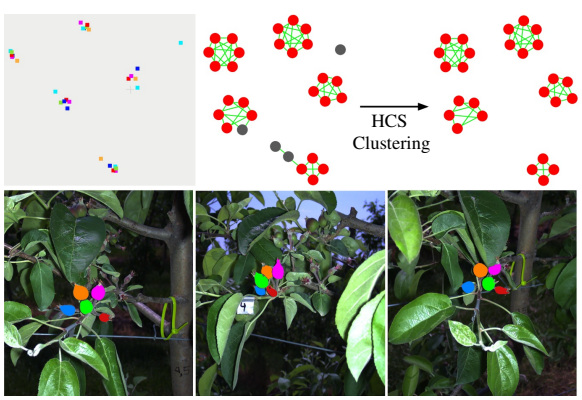

Fig. 8. Top: A point cloud of centroids is used to build a graph which consists of false detections (grey). HCS clustering removes the false detections and each subgraph represents an individual fruitlet. Bottom: Each color represents the same fruitlet associated in different images.

C. Ellipse Fitting and Sizing

For every fruitlet, an ellipse is fit using the OpenCV [31] fitEllipse function (Fig. 9). The size is calculated as $\frac{b \times a_{\text{minor}}}{d_{\text{med}}}$, where b is the baseline of the stereo camera, d_{med} is the median disparity value of the segmented pixels, and a_{minor} is the length of the minor axis of the fit ellipse. The derivation for this equation can be found in [24]. For each fruitlet, the final size is determined to be largest among its associations.

Fig. 9. Examples of ellipses fit to fruitlets.

VI. Experiments

A. Simulated Experiments

Experiments in simulation were performed on synthetic apple trees in a Gazebo [32] simulated environment. Fruitlets are modelled as ellipsoids, and each fruitlet in a cluster is assigned a unique color. Three different structural topologies with varying levels of foliage were created using SpeedTree[3]. For each tree, 16 trials were run with randomly generated clusters for a total of 48 trials. Clusters were randomly assigned 3-6 fruitlets with varying poses. The sizes of the fruitlets were also random with diameters ranging from 7mm-14mm. For each trial, a single cluster is selected and AprilTag hung in close proximity to it, as shown in Fig. 10.

The different coloring of the fruitlets is used only for instance segmentation and has no effect on data association. Fruitlets are segmented by HSV thresholding followed by Density-Based Spatial Clustering [33] to separate different instances of the same color belonging to different clusters.

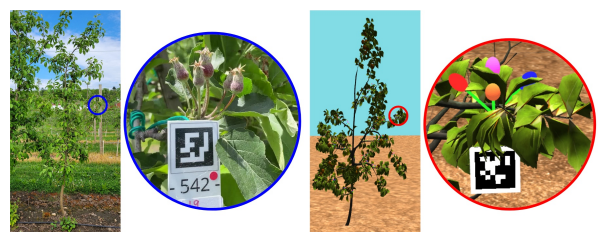

Fig. 10. Example real and simulated apple trees and clusters.

We assess the effectiveness of our approach by running ablations, comparing to two naive approaches, and comparing the planner presented by Zaenker *et al.* [10]. All planners were given a combined planning and execution time of two minutes to move around the world and capture images. For each experiment, coarse and fine octree resolutions of 1cm and 3mm were used. The methods evaluated include

1) *Fruitlet Viewpoint Planner (FVP)*: Our method as described in Section IV.
2) *Single Map Fruitlet Viewpoint Planner*: The dual-map representation of the environment from Section IV-B was removed and replaced with a single coarse (*SM-FVP-LR*) or fine (*SM-FVP-HR*) octree.
3) *Non-Attention-Guided Fruitlet Viewpoint Planner (NAG-FVP)*: The attention mechanism used to calculate information gain is removed.

[3] www.speedtree.com

TABLE I

EXPERIMENTAL RESULTS

Metric	Simulated								Real-World
	FVP	NAG-FVP	SM-FVP-LR	SM-FVP-HR	RVP-LR	RVP-HR	RS	EDS	FVP
MP (%)	**95.8**	87.5	90.7	88.0	95.4	80.0	88.4	85.8	88.1
MAE (mm)	**0.613**	0.983	0.834	1.43	0.841	1.44	1.40	1.44	0.811
MAPE (%)	**6.17**	9.92	8.33	12.9	8.36	13.1	13.2	13.1	7.24
R^2	**0.907**	0.746	0.824	0.534	0.767	0.575	0.519	0.524	0.909

4) *ROI Viewpoint Planner*: The planner presented by [10] with one coarse (*RVP-LR*) or fine (*RVP-HR*) octree.
5) *Random Sampler (RS)*: Naive approach where viewpoints are randomly selected from a partial sphere surrounding the Attention Region.
6) *Evenly Distributed Sampler (EDS)*: Naive approach where viewpoints are evenly sampled on a partial sphere surrounding the Attention Region.

We report the match percent (MP), which is the percent of fruitlets that can be matched with ground truth. We also compare the mean absolute error (MAE) and mean absolute percentage error (MAPE) between the measured and ground truth sizes. The results can be seen in Table I.

$$\text{MAE} = \frac{1}{N}\sum_{n=1}^{N} |\text{measured} - \text{gt}| \tag{7}$$

$$\text{MAPE} = \frac{100}{N}\sum_{n=1}^{N} \frac{|\text{measured} - \text{gt}|}{\text{gt}} \tag{8}$$

Our FVP achieves the highest MP, the lowest MAE, and the lowest MAPE compared to all other evaluated methods. Our planner matches 95.8% of fruitlets and achieves a MAPE of 6.17%. Additionally, we obtain the best linear fit with ground truth, with an R^2 score of 0.907 (Fig. 11).

Analyzing the ablations and state of the art, SM-FVP-HR and RVP-HR performed the worst with regards to MAE, MAPE, and R^2 scores. The results were comparable to the naive approaches. This is because of the computational costs associated with using higher resolution octrees which result in a fewer number of evaluated viewpoints. NAG-FVP also did not perform well, again because of the additional costs required to update the fine octree without the benefits of attention-guided information gain.

There is a trade-off between RVP-LR and SM-FVP-LR. RVP-LR is able to size more fruitlets compared to SM-FVP-LR, but has slightly larger MAE and MAPE. This could be explained by RVP-LR being encouraged to explore more, as voxels outside the Attention Region contribute the information gain. Because SM-FVP-LR applies the attention mechanism, it is incentivized to more accurately explore previously detected ROIs inside the Attention Region.

Based on these results, our FVP effectively demonstrates that combining our dual-map representation with attention improves sizing performance across the presented metrics.

B. Real-World Experiments

Real-world experiments were conducted at the University of Massachusetts Amherst Cold Spring Orchard. The robotic system from Section III was used to size 30 McIntosh apple clusters on 05/22/2023 using our Fruitlet Viewpoint Planner. Caliper measurements were also recorded to be used as ground truth. The Mask-RCNN network was trained using 600 images of Fuji, Gala, and Honeycrisp clusters.

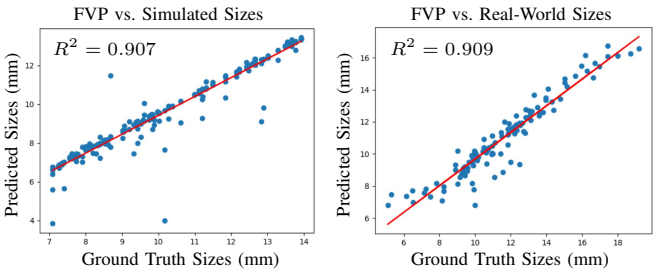

Fig. 11. Linear fit between predicted and ground truth sizes for our FVP in both simulation (left) and real-world (right) experiments.

The results can be seen in Table I. Our Fruitlet Viewpoint Planner demonstrates impressive performance on the real-world experiments. The planner achieves a mean average percentage error of 7.24% and an R^2 score of 0.909 between the predicted and caliper-measured sizes (Fig. 11). In addition, 88.1% percent of fruitlets were able to be matched. While this is a strong result, there is room for improvement. Through qualitative inspection, many missed instances were a result of the Mask R-CNN network missing detections. This is likely because no McIntosh apples were used in training. With additional training data, the performance of our planner with regards to this metric would likely improve.

VII. CONCLUSION

Our next-best-view planner outperforms previous state of the art in simulation and demonstrates strong results on a real-world dataset. Although initially designed to size apple fruitlets, the presented methods could be used to size other small fruits and plants of comparable size. For future work, we hope to integrate more global viewpoint coverage to size multiple clusters at a time. Instead of using an AprilTag to identify a cluster, global broad scans of trees could be used to find high-density fruit regions using the methods presented in Section IV-A. Our map representation could be extended to include multiple fine octrees and Attention Regions to capture finer coverage of each cluster.

VIII. ACKNOWLEDGEMENTS

This work was partially supported by NSF/USDA NIFA 2020-01469-1022394, NSF Robust Intelligence 1956163, and NSF/USDA NIFA AIIRA AI Research Institute 2021-67021-35329. The authors would like to thank Abhi Silwal and Zack Rubinstein for their support throughout this project.

REFERENCES

[1] D. W. Greene, A. N. Lakso, T. L. Robinson, and P. Schwallier, "Development of a fruitlet growth model to predict thinner response on apples," *HortScience*, vol. 48, no. 5, pp. 584–587, 2013.

[2] T. Hondo, K. Kobayashi, and Y. Aoyagi, "Real-time prediction of growth characteristics for individual fruits using deep learning," *Sensors*, vol. 22, no. 17, 2022. [Online]. Available: https://www.mdpi.com/1424-8220/22/17/6473

[3] S. Kelly, A. Riccardi, E. Marks, F. Magistri, T. Guadagnino, M. Chli, and C. Stachniss, "Target-aware implicit mapping for agricultural crop inspection," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*, 2023, pp. 9608–9614.

[4] C. Lehnert, D. Tsai, A. Eriksson, and C. McCool, "3d move to see: Multi-perspective visual servoing towards the next best view within unstructured and occluded environments," in *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2019, pp. 3890–3897.

[5] P. Roy and V. Isler, "Active view planning for counting apples in orchards," in *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2017, pp. 6027–6032.

[6] C. Wu, R. Zeng, J. Pan, C. C. L. Wang, and Y.-J. Liu, "Plant phenotyping by deep-learning-based planner for multi-robots," *IEEE Robotics and Automation Letters*, vol. 4, no. 4, pp. 3113–3120, 2019.

[7] W. Yuan, T. Khot, D. Held, C. Mertz, and M. Hebert, "Pcn: Point completion network," in *2018 International Conference on 3D Vision (3DV)*, 2018, pp. 728–737.

[8] F. Sukkar, G. Best, C. Yoo, and R. Fitch, "Multi-robot region-of-interest reconstruction with dec-mcts," in *2019 International Conference on Robotics and Automation (ICRA)*, 2019, pp. 9101–9107.

[9] G. Best, O. M. Cliff, T. Patten, R. R. Mettu, and R. Fitch, "Dec-mcts: Decentralized planning for multi-robot active perception," *The International Journal of Robotics Research*, vol. 38, no. 2-3, pp. 316–337, 2019. [Online]. Available: https://doi.org/10.1177/0278364918755924

[10] T. Zaenker, C. Smitt, C. McCool, and M. Bennewitz, "Viewpoint planning for fruit size and position estimation," *CoRR*, vol. abs/2011.00275, 2020. [Online]. Available: https://arxiv.org/abs/2011.00275

[11] T. Zaenker, C. Lehnert, C. McCool, and M. Bennewitz, "Combining local and global viewpoint planning for fruit coverage," in *2021 European Conference on Mobile Robots (ECMR)*, 2021, pp. 1–7.

[12] R. Menon, T. Zaenker, N. Dengler, and M. Bennewitz, "Nbv-sc: Next best view planning based on shape completion for fruit mapping and reconstruction," 2023.

[13] A. K. Burusa, E. J. van Henten, and G. Kootstra, "Attention-driven active vision for efficient reconstruction of plants and targeted plant parts," 2022.

[14] E. Olson, "Apriltag: A robust and flexible visual fiducial system," in *2011 IEEE International Conference on Robotics and Automation*, 2011, pp. 3400–3407.

[15] A. Silwal, T. Parhar, F. Yandun, H. Baweja, and G. Kantor, "A robust illumination-invariant camera system for agricultural applications," in *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2021, pp. 3292–3298.

[16] A. Silwal, F. Yandún, A. K. Nellithimaru, T. Bates, and G. Kantor, "Bumblebee: A path towards fully autonomous robotic vine pruning," *CoRR*, vol. abs/2112.00291, 2021. [Online]. Available: https://arxiv.org/abs/2112.00291

[17] K. He, G. Gkioxari, P. Dollár, and R. B. Girshick, "Mask R-CNN," *CoRR*, vol. abs/1703.06870, 2017. [Online]. Available: http://arxiv.org/abs/1703.06870

[18] L. Lipson, Z. Teed, and J. Deng, "RAFT-Stereo: Multilevel Recurrent Field Transforms for Stereo Matching," *arXiv preprint arXiv:2109.07547*, 2021.

[19] D. Coleman, I. A. Sucan, S. Chitta, and N. Correll, "Reducing the barrier to entry of complex robotic software: a moveit! case study," *CoRR*, vol. abs/1404.3785, 2014. [Online]. Available: http://arxiv.org/abs/1404.3785

[20] E. Hartuv and R. Shamir, "A clustering algorithm based on graph connectivity," *Information Processing Letters*, vol. 76, pp. 175–181, 12 2000.

[21] H. Freeman, E. Schneider, C. H. Kim, M. Lee, and G. Kantor, "3d reconstruction-based seed counting of sorghum panicles for agricultural inspection," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*, 2023, pp. 9594–9600.

[22] A. Hornung, K. M. Wurm, M. Bennewitz, C. Stachniss, and W. Burgard, "OctoMap: An efficient probabilistic 3D mapping framework based on octrees," *Autonomous Robots*, 2013, software available at https://octomap.github.io. [Online]. Available: https://octomap.github.io

[23] S. Marangoz, T. Zaenker, R. Menon, and M. Bennewitz, "Fruit mapping with shape completion for autonomous crop monitoring," in *2022 IEEE 18th International Conference on Automation Science and Engineering (CASE)*, 2022, pp. 471–476.

[24] M. Qadri, "Robotic vision for 3d modeling and sizing in agriculture," Master's thesis, Carnegie Mellon University, Pittsburgh, PA, August 2021.

[25] A. Gongal, M. Karkee, and S. Amatya, "Apple fruit size estimation using a 3d machine vision system," *Information Processing in Agriculture*, vol. 5, no. 4, pp. 498–503, 2018. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2214317317302408

[26] J. M. Ponce, A. Aquino, B. Millan, and J. M. Andújar, "Automatic counting and individual size and mass estimation of olive-fruits through computer vision techniques," *IEEE Access*, vol. 7, pp. 59 451–59 465, 2019.

[27] Z. Wang, K. B. Walsh, and B. Verma, "On-tree mango fruit size estimation using rgb-d images," *Sensors*, vol. 17, no. 12, 2017. [Online]. Available: https://www.mdpi.com/1424-8220/17/12/2738

[28] H. W. Kuhn, "The Hungarian Method for the Assignment Problem," *Naval Research Logistics Quarterly*, vol. 2, no. 1–2, pp. 83–97, March 1955.

[29] M. A. Branch, T. F. Coleman, and Y. Li, "A subspace, interior, and conjugate gradient method for large-scale bound-constrained minimization problems," *SIAM Journal on Scientific Computing*, vol. 21, no. 1, pp. 1–23, 1999. [Online]. Available: https://doi.org/10.1137/S1064827595289108

[30] R. H. Byrd, R. B. Schnabel, and G. A. Shultz, "Approximate solution of the trust region problem by minimization over two-dimensional subspaces," *Math. Program.*, vol. 40, no. 1-3, pp. 247–263, 1988. [Online]. Available: https://doi.org/10.1007/BF01580735

[31] G. Bradski, "The OpenCV Library," *Dr. Dobb's Journal of Software Tools*, 2000.

[32] N. Koenig and A. Howard, "Design and use paradigms for gazebo, an open-source multi-robot simulator," in *2004 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) (IEEE Cat. No.04CH37566)*, vol. 3, 2004, pp. 2149–2154 vol.3.

[33] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," in *Proc. of 2nd International Conference on Knowledge Discovery and*, 1996, pp. 226–231.