Active Semantic Mapping with Mobile Manipulator in Horticultural Environments

Jose Cuaran[1], Kulbir Singh Ahluwalia[1], Kendall Koe[1], Naveen Kumar Uppalapati[3], and Girish Chowdhary[1,2]

***Abstract*—Semantic maps are fundamental for robotics tasks such as navigation and manipulation. They also enable yield prediction and phenotyping in agricultural settings. In this paper, we introduce an efficient and scalable approach for active semantic mapping in horticultural environments, employing a mobile robot manipulator equipped with an RGB-D camera. Our method leverages probabilistic semantic maps to detect semantic targets, generate candidate viewpoints, and compute corresponding information gain. We present an efficient ray-casting strategy and a novel information utility function that accounts for both semantics and occlusions. The proposed approach reduces total runtime by 8% compared to previous baselines. Furthermore, our information metric surpasses other metrics in reducing multi-class entropy and improving surface coverage, particularly in the presence of segmentation noise. Real-world experiments validate our method's effectiveness but also reveal challenges such as depth sensor noise and varying environmental conditions, requiring further research.**

Code, video and supplementary material.

***Index Terms*—Active Mapping, Agricultural Robotics**

I. Introduction

Semantic maps in agricultural environments provide robots with crucial information to guide actions, like navigating rows or harvesting fruit. It also supplies data to farmers or management systems for tasks such as predicting yields and monitoring growing rates [1]. However, building these maps presents several challenges, such as variations in environmental conditions, wind disturbances, and incomplete observations caused by occlusions.

Various works address the problem of mapping in agricultural environments [2]–[4]. Most of these works use fixed cameras to collect images, which are post-processed with approaches like Structure from Motion or Neural Radiance Fields to achieve a 3D reconstruction. Such approaches often suffer from limited viewpoints and self-occlusions that are common in agricultural scenes. Moreover, for tasks like yield prediction, semantics corresponding to fruits are more relevant than semantics corresponding to leaves or other parts, making target-aware mapping necessary.

Thus, active mapping approaches, common in other environments, have been adopted to improve the reconstruction quality of agricultural environments. A popular method is Next Best View (NBV) planning, which relies on computing a utility value for potential viewpoint candidates [5]–[7]. Although prior works show promising results in building target-aware reconstructions, some works still rely on the assumption that bounding boxes of targets are given [5]. In addition, binary occupancy maps are commonly used, which only provide information about the occupancy but not the probability of semantic classes. Finally, most of these studies rely on ray-casting, a technique that involves tracing the paths of multiple rays from a sensor to gather occupancy information for each voxel. This method can be computationally expensive if not implemented efficiently.

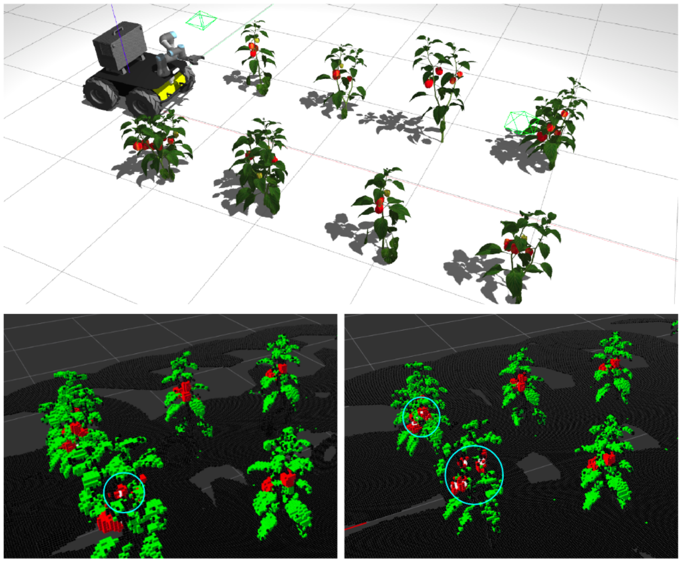

Fig. 1. Top: Simulation environment with capsicum plants and Husky Robot. Bottom left: Reconstructions done using our approach. Bottom right: Reconstruction done using predefined dense scanning. Blue circles highlight incomplete fruit areas.

To address these limitations, we present an efficient and scalable NBV approach for active semantic mapping in horticultural environments. Unlike prior works based on binary occupancy maps [5]–[7], we leverage semantic probabilistic maps [8] to directly detect the semantic targets without the need for predefined bounding boxes. We use these targets to sample multiple viewpoint candidates around them. An information utility value is computed for each candidate for which an efficient ray-tracing strategy is implemented. Finally, we evaluate the performance of our method on different Information Gain (IG) metrics, and propose a new metric that leverages the multi-class probability map.

In summary, the main contributions of this paper are:

- An approach for target-aware semantic mapping in horticultural environments leveraging state-of-the-art multi-

The authors are with (1) the Department of Computer Science, (2) the Department of Agricultural and Biological Engineering and (3) National Center for Supercomputing Applications at University of Illinois, Urbana-Champaign.

Correspondence to {jrc9,girishc}@illinois.edu

class probabilistic maps.
- A simple but effective strategy to make ray casting faster and enable the efficient evaluation of viewpoint candidates.
- A novel information utility function that takes into account occlusions, proximity to targets, and semantics.
- An evaluation of the performance of our method and utility function considering the segmentation noise, which is common in agricultural environments because of environment variations, but rarely considered in previous works.

II. Related works

Numerous studies have addressed the problem of mapping and 3D reconstruction in agricultural environments [4], [9]–[11]. However, the majority of these approaches are passive, where sensors, such as cameras or LiDAR, are fixed on a robot or handheld while capturing a sequence of images or scans. The reconstructed 3D point clouds are subsequently used to estimate parameters such as canopy volume, trunk diameter, tree height, and fruit count. While these methods are effective in capturing details at the plant level, they are less suitable for tasks requiring the estimation of the shape, volume, or pose of individual fruits, as they are often affected by occlusions and limited viewpoints. Although fruit counting is performed in [4] and [11], these analyses are conducted in image space, thereby neglecting occluded fruits that can only be captured using 3D models.

Active mapping approaches have been increasingly proposed to enhance the quality of reconstructions in agricultural environments. Among these approaches, Next Best View (NBV) planning is one of the most common techniques. NBV planning involves identifying sensor poses or trajectories that maximize the information gain (IG), thereby reducing map uncertainty while minimizing associated costs, such as path length or time [5]–[7]. For example, Burusa *et al.* [5] proposes a method for the active 3D reconstruction of tomato plants using a robotic manipulator. In this approach, the selection of the next view is based on the IG calculated for random viewpoint candidates, leveraging a probabilistic occupancy map of the environment. However, a notable limitation of this work is its reliance on predefined 3D bounding boxes for various parts of the plants, which guides the active planning algorithm toward the desired semantics. In contrast, our method addresses this limitation by maintaining a semantic map of the environment, which eliminates the need for predefined bounding boxes.

Zaenker *et al.* [6] introduce an NBV planning approach aimed at mapping sweet pepper plants and estimating the size and position of fruits. Their method employs a probabilistic octree representation of the environment, where each cell encodes occupancy and region of interest (ROI) probabilities. To direct the algorithm's focus toward fruit areas, they generate random viewpoint candidates by leveraging frontier nodes near ROI regions. General frontier nodes are also utilized to encourage exploration and the discovery of new ROIs. The authors report improvements in volume coverage and accuracy compared to pure frontier-based exploration that does not account for semantic ROIs. Building upon this work, our approach introduces several enhancements. Unlike their method, which maintains a single ROI probability for each cell, we maintain a multi-class probability distribution, allowing us to handle multiple semantic classes. Additionally, while their approach samples random candidate viewpoints for each frontier node near the ROIs, potentially leading to redundant viewpoints or the omission of critical ones, our method uniformly samples potential viewpoints around fruit clusters. This ensures uniform coverage and also more efficiency, as the number of clusters is less than the number of frontier voxels.

Recent studies have proposed NBV planning approaches that do not rely exclusively on ray casting, which is known to be effective but computationally intensive. For instance, [7] presents a method that utilizes shape completion with superellipsoids to identify optimal viewpoints. By leveraging the predicted surfaces of fruits, this approach guides the sensor toward viewpoints targeting missing surface regions. While this method significantly reduces planning time compared to ray-casting-based techniques, its application is limited to certain types of crops. It is also dependent on the accuracy of the shape completion algorithm, which continues to be an area of active research. Another notable contribution in this line of research is 3D Move-to-see [12]. This method employs a custom 3D camera array that optimizes a utility function through gradient ascent. A single shot is sufficient to guide the camera toward a viewpoint that maximizes the visible fruit area while avoiding occlusions. However, this method addresses only the local path planning problem and must be integrated with a global path planning strategy to achieve effective active mapping [13].

Octomap [14] has been the primary representation used in most prior works due to its scalability and ability to encode probabilistic information. These maps, which contain binary occupancy probabilities for each cell, enable the computation of various information gain metrics by counting the number of unknown voxels or calculating entropy. Some proposed metrics also consider factors such as visibility and proximity to specific targets [15]. Recently, [8] extended this framework to multi-class probability maps, where each voxel, in addition to the binary occupancy probability, includes a categorical distribution over multiple semantic classes. Building on this framework, we propose an information gain metric that incorporates occlusions, proximity, and class probability to guide the algorithm in prioritizing certain targets.

III. Methods

We aim to create a semantic and geometric representation of plants in a horticultural environment, focusing on specific targets such as fruits with an active mapping approach. We assume that the plants are distributed along rows so that a mobile manipulator can go through the middle of them collecting RGBD images while building the map. The plants' height is also assumed to be known.

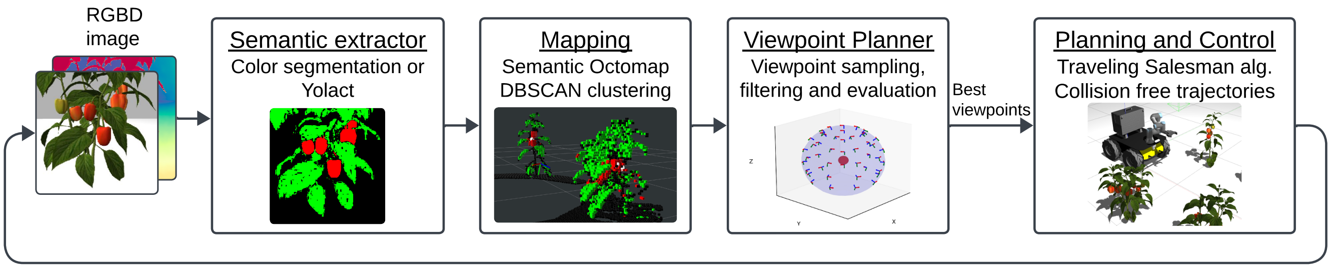

Fig. 2. System overview. Semantic extractor: semantic segmentation is used to extract fruits, leaves, and background in the scene. Mapping: a 3D semantic and probabilistic representation of the environment is generated. Viewpoint planner: It generates potential viewpoint candidates and chooses those with the highest information gain. Planning and control: collision-free trajectories to the desired goals and control signals are generated.

Fig. 2 presents an overview of our system, which is composed of four modules: (i) Semantic Extractor, (ii) Mapping, (iii) Viewpoint Planner, (iv) Planning and Control. Given an RGBD observation of the scene, the semantic extractor generates masks for K semantic classes. The depth images and masks are merged into a semantic point cloud, which is then used by the mapping module. This module creates a probabilistic semantic map encoded as an octree data structure, where each voxel stores occupancy and semantic probabilities. Voxels with a high probability of belonging to a desired target class k_d are clustered into different groups. Each cluster is then utilized by the viewpoint planning module to generate a set of candidate viewpoints, which are subsequently filtered and evaluated using a utility function. The viewpoints with the highest utility values are finally executed. Further details are given below.

(i) Semantic extractor: We define fruits, leaves, and background as our three semantic classes, with fruits as the main semantic class. Given an RGB image, semantic segmentation is applied to extract pixel labels corresponding to those classes. We perform color segmentation based on HSI color values for simulation, whereas for real-world experiments we use the Yolact model [16].

(ii) Mapping: We use Semantic Octomap [8] to obtain an efficient, compact, and scalable map representation. Given a semantic point cloud and camera pose, Semantic Octomap creates a probabilistic semantic octree where each voxel x encodes a categorical distribution $p_c(x)$ over different classes as well as a binary occupancy probability value $p_o(x)$. As we know the height of the plants and the robot location with respect to the crop rows, we execute a sparse scanning with predefined camera poses around the plants. With these images, Semantic Octomap creates an initial and incomplete semantic octree. Subsequently, the voxels classified as fruits according to the semantic probability values $p_c(x)$ are used to compute fruit clusters applying the DBSCAN clustering algorithm [17]. We set the neighborhood threshold ϵ equal to the expected average fruit radius, ensuring that fruits close to each other are merged into a single cluster, thus preventing redundant viewpoints when sampling viewpoint candidates. This parameter, along with other parameters described in this paper are summarized in Table I.

(iii) Viewpoint Planner: We present an NBV planner that refines an incomplete map by selecting the best viewpoints to improve reconstruction, focusing on specific semantic classes. For each fruit cluster centroid, we sample viewpoints on a sphere with radius r, using N_θ elevation and N_ϕ azimuth angles in the ranges $[30°, 150°]$ and $[0°, 360°]$, respectively. Viewpoints near the manipulator's workspace are filtered. Subsequently, a utility value is computed for each candidate viewpoint based on the IG for selecting the top-k utility viewpoints. These selected high IG viewpoints are then passed onto the MoveIt motion planner.

Filtering out candidate viewpoints. Sampling viewpoints uniformly around each cluster centroid generates many candidates, necessitating a filtering strategy to reduce computational overhead before viewpoint evaluation. We compute points defining the manipulator's workspace using forward kinematics for multiple joint combinations and move candidate viewpoints into the arm workspace. Viewpoints whose viewing direction does not intersect the workspace are discarded, with intersections computed via the Nearest Neighbor algorithm.

OSAMCEP - A novel IG metric. Leveraging the multi-class probability information of our semantic map, we propose an information metric called Occlusion and Semantic Aware Multi-Class Entropy with Proximity Count (OSAMCEP), which considers occlusions and focuses on desired semantic targets. For a viewpoint v with a set of cast rays $\mathcal{R}_v$, a set $\mathcal{X}$ of observed voxels per ray and the information utility function $I(x)$, we compute the IG, $\mathcal{G}_v$, as:

$$\mathcal{G}_v = \textstyle\sum_{r \in \mathcal{R}_v} \sum_{x \in \mathcal{X}} I(x) \tag{1}$$

Given a voxel x with a probability of occupancy $p_o(x)$, and $p_c(x)$ the categorical distribution over K semantic classes, the corresponding information utility function $I(x)$ is computed as seen in Equation 2:

$$I(x) = \begin{cases} P_v(x)H(x) & \text{,if semantic class } l_x \text{ is unknown or} \\ & \text{semantic target and } dist(x) < max\_dist \\ 0 & \text{,otherwise} \end{cases} \tag{2}$$

$$P_v(x_n) = \textstyle\prod_{i=1}^{n-1} (1 - p_o(x_i)) \tag{3}$$

$$H(x) = -\textstyle\sum_{k=0}^{K} p_c(l_x = k) \ln p_c(l_x = k) \tag{4}$$

Where $P_v(x_n)$ is the probability of voxel x_n being visible considering the previous voxels x_i traversed on the ray before reaching voxel x_n as seen in Equation 3. $H(x)$ is the multi-class entropy defined in Equation 4. $dist(x)$ is the distance between a semantic target and voxel x, and max_dist is a threshold distance value for determining the relevant voxels close to the semantic target.

Unlike other metrics proposed in previous works like [5], [6] which rely only on unknown voxels in proximities to the targets to compute the IG, we consider both unknown and semantic voxels. This encourages the algorithm to reduce the entropy of specific semantic nodes, making target-aware mapping more effective. In addition, while previous works [5], [6] determine the presence of occlusions based on occupancy thresholds, we employ the probability of visibility $P_v(x)$ suggested in [15], computed along each ray, resulting in a more consistent estimation for occlusions.

Increasing Ray Tracing Efficiency. Ray-tracing all image pixels to compute each viewpoint's utility is computationally expensive, so uniform downsampling is often used to decrease runtime. However, relevant semantic regions can be missed depending on the distance from the scene because of downsampling. To address this, we propose two strategies. First, downsample pixels with a step size ds inversely proportional to the target distance z:

$$ds = \frac{\delta S * F_x}{z} \tag{5}$$

where δS is the octomap resolution and F_x the camera focal length. The second strategy assumes distant rays do not impact the current cluster's IG and discards them during ray tracing. Using the typical fruit cluster size L, we define a bounding box with size b in the image space for ray casting boundaries:

$$b = \frac{L * F_x}{z} \tag{6}$$

(iv) Planning and Control: Given a set of best viewpoints for the current scene, we compute the order of execution following the traveling salesman planning algorithm. The MoveIt ROS package is then used to execute these viewpoints following collision-free trajectories.

IV. Experiments

Our approach is evaluated in Gazebo simulation using 8 sweet pepper plant models from [6]. The three semantic classes used in all simulation experiments consist of the background, leaves, and the target semantic class, fruit. The experiments run on a laptop with an Intel Core i7-11800H (2.30 GHz) and 16 GB RAM.

Metrics We evaluate the evolution of multi-class entropy and surface coverage defined as in [8] and [15] respectively, only for fruits as they are the main targets. While surface coverage is an indicator of reconstruction completeness, the entropy serves as a measure of uncertainty in this reconstruction. We compute the total entropy as the sum of entropy values (equation 4) for all voxels inside a 3D bounding box enclosing each fruit cluster. The surface coverage SC is computed as follows:

$$SC = \frac{\text{Observed surface points}}{\text{Total points in ground truth model}} \tag{7}$$

A surface point in the ground truth model is considered observed if the closest point in the reconstruction point cloud is within a distance threshold equal to the map resolution. See some evaluation parameters in Table I. Additional parameters can be found in the supplementary material.

TABLE I
Parameters used during evaluation

Category	Parameter	Value	Description
Mapping	δS [m]	0.015	Map resolution
	max_range [m]	1.0	Max depth range for mapping
Segmentation	P_{gt}	0.7	Probability of correct classification during semantic segmentation
Downsampling	L [m]	0.1	Typical size of a fruit cluster
NBV-planner	r [m]	0.4	Radius of sphere for viewpoint sampling
	N_ϕ	10	Number of azimuth samples
	N_θ	5	Number of elevation samples
OSAMCEP	max_dist [m]	0.1	Max distance from the semantic target for relevant voxels
DBSCAN	ϵ [m]	0.05	Radius of a neighborhood

In addition, we simulate segmentation noise to account for the limitations of common segmentation models, especially in agricultural environments [1]. To this end, for every mask we assign the ground truth class with probability P_{gt} and a wrong class with probability $(1 - P_{gt})$.

A. Overall performance

In this part, we aim to evaluate our whole pipeline for active mapping using a 6 DOF UR3 manipulator on a Clearpath Husky robot [18] and an RGBD camera on the end-effector. We create a simulation environment with 8 sweet pepper plants distributed in rows as shown in Fig. 1. The mobile robot is given a sequence of waypoints to take the manipulator close to each pair of plants. At each position, the manipulator starts the initial scanning with 12 predefined and sparse viewpoints. Then, the next best viewpoints for the fruit clusters near the robot are computed and executed. Since some viewpoints are not reachable, we sample, evaluate, and execute new viewpoints if necessary until successfully executing 12, before moving to the next plants. We compare our approach with two baselines: (i) the frontier-based approach proposed by Zaenker *et al.* [6] described in section II. It is adapted to start with the initial predefined scanning as our method to ensure the same initialization, followed by 12 additional NVB successful viewpoint executions based on ROI frontiers as in the original implementation; (ii) a dense scanning approach consisting of the same number of viewpoints (24) for fair comparison. The results of 10 trials with different random seeds are averaged for each method.

Fig. 3 depicts the evolution of entropy and surface coverage as the robot navigates through the crop row, collecting views. Fig. 4 shows the corresponding runtime for each method. It is evident that the two active mapping approaches result in higher fruit surface coverage (approximately 11% higher by the end of the experiment) and lower entropy compared to the predefined scanning method. This improvement can be attributed to the richer viewpoints computed by these methods, as both are information-based and target-aware. However, this enhancement comes at a cost, as the runtime for the active methods is more than 45% higher than that of the predefined scanning approach (Fig. 4). Although all three methods execute the same number of successful viewpoints, the active mapping approaches require the evaluation of the utility of each candidate viewpoint. Additionally, some of the best viewpoints are not reachable by the manipulator arm, resulting in increased computation time for motion planning and control.

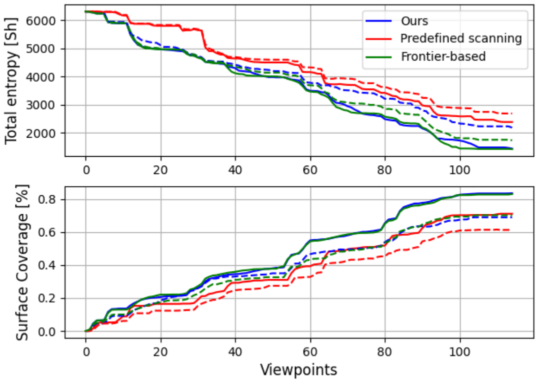

Fig. 3. Total entropy and surface coverage with our active mapping approach, a frontier-based approach, and predefined scanning, with (dashed lines) and without (solid lines) segmentation noise. All the methods execute 120 viewpoints along the path between two crop rows with 8 plants in total.

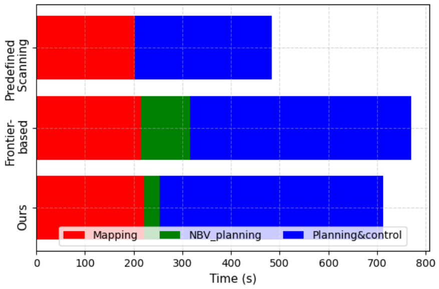

Fig. 4. Runtime comparison between our approach and two baselines. The total time is divided into three tasks: mapping, NBV planning, and motion planning and control. Note that our approach significantly reduces the NBV planning time compared to the frontier-based method.

Between the two active mapping methods, the frontier-based approach and our cluster-centric method show similar performance in surface coverage and entropy. However, our method improves runtime by 8% due to a more efficient NBV planning strategy. Specifically, we observed that most of the random viewpoint candidates generated by the frontier-based method are not reachable by the robot arm. Consequently, multiple rounds of sampling and evaluation are required to achieve successful viewpoints, which increases the NBV planning time.This issue becomes more pronounced over time, as the number of ROI frontier nodes decreases, affecting the diversity of viewpoint candidates. In contrast, our approach uniformly samples candidate viewpoints from fruit cluster centroids, ensuring diversity and complete coverage.

It is worth noting in Fig. 3 that both surface coverage and entropy are significantly impacted by segmentation noise (indicated by dashed lines). Specifically, segmentation noise results in a 17% reduction in surface coverage by the end of the experiment, which could pose a critical limitation in real-world applications. Finally, Fig. 1 shows that the final reconstruction built by our active mapping approach is more complete than the one of the predefined scanning method. The baseline exhibits several blank centroids, which requires additional viewpoints for a complete reconstruction.

B. Ablation studies

To minimize the influence of the constrained manipulator workspace, control and motion planning algorithm accuracy on the comparison of our active mapping approach with other mapping algorithms, we perform ablation studies using a free-moving camera. We separately assess the performance of our downsampling strategy for ray casting and the information utility function. Using 6 plant models, we run the active mapping algorithms for 30 viewpoints per plant across 10 trials with varying initializations. We then compute the average entropy and surface coverage across all plants.

Evaluation of our downsampling strategy for ray casting The downsampling strategy in Section III aims to accelerate ray casting while preserving viewpoint quality. We compared its performance with two baselines: dense sampling (28x28 grid) and sparse sampling (6x6 grid) from [6]. Both baselines perform uniform sampling across the images. Since our downsampling approach is dependent on the distance between cluster centroids and viewpoints, we tested two distance values, 0.4m and 0.6m. Fig. 5 shows our downsampling method achieves comparable performance to the dense sampling baseline by prioritizing image regions corresponding to semantic targets. Note that the sparse sampling baseline resulted in higher entropy and lower surface coverage over time, especially for larger distance values. This outcome is expected, as uniform sampling across the entire image can miss semantic areas depending on the distance from the scene. Our strategy mitigates this by parameterizing the downsampling process based on distance from the clusters and expected cluster size. The average ray tracing times were 2.58 ms for our method, 45.21 ms for dense sampling, and 2.50 ms for sparse sampling, indicating that our approach retains the efficiency of sparse downsampling while maintaining the viewpoint quality of dense sampling.

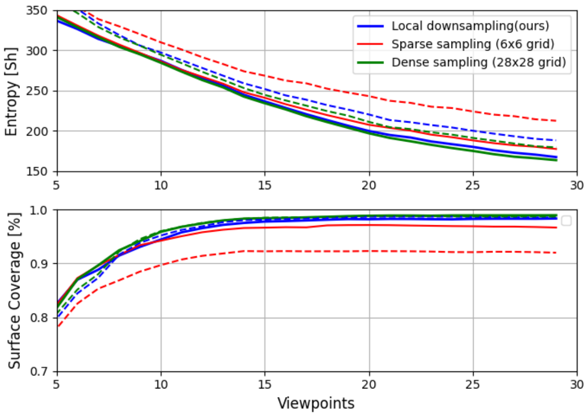

Fig. 5. Entropy and surface coverage for different downsampling strategies. Solid lines and dashed lines indicate viewpoints taken at 0.4 m and 0.6 m from cluster centroids, respectively. Note that despite the distance from the targets, our downsampling strategy maintains close performance to dense sampling.

Evaluation of the Information Gain Metric We consider several information metrics as baselines, including Average Entropy (AE) [15], Unknown Voxels Count (UVC) [6], [15], Unknown Voxels with Proximity Count [6], Occlusion Aware Entropy (OAE) [15], and Mutual Information (MI) [8]. We also consider the case of random sampling (RS). Figure 6 presents the results averaged across six plants, demonstrating that our proposed IG metric outperforms all baselines in reducing scene entropy and maximizing surface coverage with the fewest viewpoints. Interestingly, the majority of the IG metrics perform closely to our metric when segmentation noise is not considered, with even the random sampling strategy yielding competitive results. This suggests that active mapping can be effectively performed as long as the viewpoints are directed toward target clusters. This observation aligns with the findings reported in [6], where no significant difference in fruit coverage was observed when using different metrics. However, when segmentation noise is introduced, our utility function shows clear advantages over other metrics. For instance, to achieve 80% of surface coverage, our metric requires an average of 10 viewpoints per plant, whereas other metrics require between 11 and 17 viewpoints. These results would translate into substantial efficiency gains in real-world mapping scenarios.

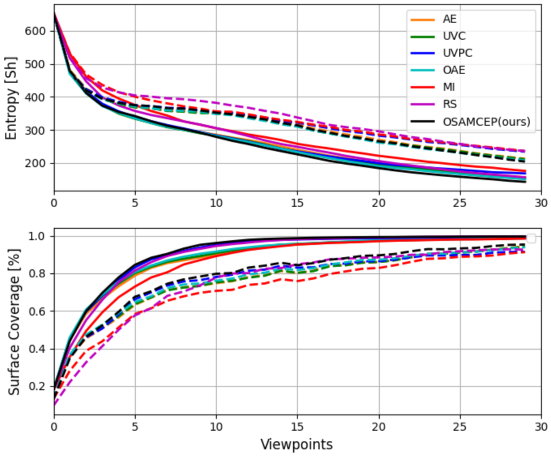

Fig. 6. Entropy and surface coverage averaged over six plants vs the number of executed viewpoints for different information gain metrics (RS: Random sampling; AE: Average entropy; UVC: Unknown voxels count; UVPC: Unknown voxels with proximity count; OAE: Occlusion aware entropy; MI: Mutual information; OSAMCEP: Occlusion and semantic Aware Multi-Class Entropy with Proximity Count), with (dashed lines) and without (solid lines) segmentation noise.

C. Real-world experiments

We perform real-world experiments with a custom 6-DOF robotic arm mounted on a wheeled platform. An Intel RealSense D405 camera (400x400 resolution) is attached to the end effector for RGBD data acquisition. The experiments take place in an urban high tunnel with tomato plants, using AprilTag markers [19] mounted on the ceiling for global localization.

Our mapping approach involves an initial predefined scanning phase, followed by NBV planning. The semantic classes of interest are fruits and background, with semantic segmentation performed using the Yolact model [16].

Fig. 7 shows the reconstructed tomato row scene. Our method effectively identifies complex viewpoints, exposing typically obscured fruit areas in fixed camera setups. However, challenges such as noisy depth maps caused by illumination variations, erratic plant motion due to variable wind speeds that violate the static-environment assumption critical to our approach, and segmentation errors, including missed detections or incorrect label assignments from crop variability, reduce reconstruction accuracy, as seen in the supplementary video.

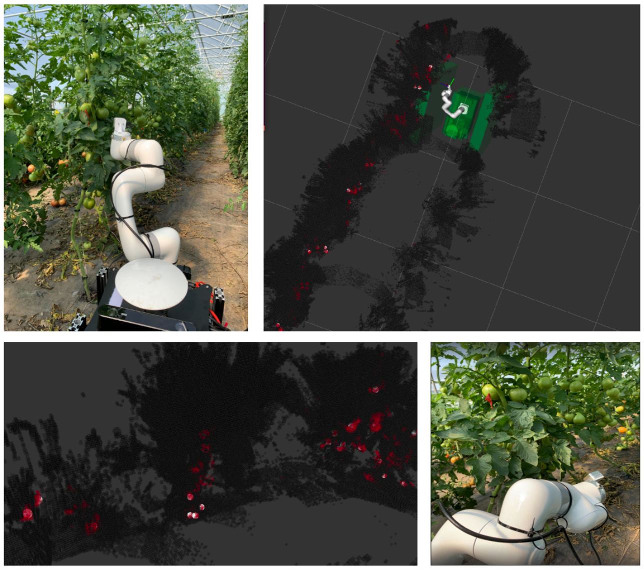

Fig. 7. Top left: Mobile manipulator in a high tunnel with tomato plants. Top right: Reconstructed scene along a tomato row. Bottom left: An amplified view of the reconstruction (white spheres are cluster centroids; red voxels are fruit nodes; black voxels are background). Bottom right: Sample best viewpoints, in which the camera looks upwards to reveal fruit areas.

V. Conclusion

This paper presented an approach for target-aware active semantic mapping in horticultural environments. By leveraging semantic octrees, semantic targets were effectively found. We introduced a cluster-centric approach for NBV planning which enhances surface coverage and reduces the entropy of the reconstructions. In addition, a downsampling strategy was proposed to improve the efficiency of ray casting, achieving significant improvements compared to traditional methods. Finally, we presented an information utility function that considers visibility, proximity, and semantics to evaluate viewpoints, outperforming other metrics especially when segmentation noise is considered.

However, some limitations still remain. For example, the accuracy of the geometric reconstruction depends significantly on the map resolution, which directly impacts the computational demand. A future direction to address this issue could be combining our NBV planning method with state-of-the-art reconstruction methods (e.g. NeRF, or Gaussian Splatting) to achieve a more accurate reconstruction of agricultural scenes.

References

[1] G. Kootstra, "Advances in visual perception for agricultural robotics," in *Advances in agri-food robotics*. Burleigh Dodds Science Publishing, 2024, pp. 3–36.

[2] C. Smitt, M. Halstead, P. Zimmer, T. Läbe, E. Guclu, C. Stachniss, and C. McCool, "Pag-nerf: Towards fast and efficient end-to-end panoptic 3d representations for agricultural robotics," *IEEE Robotics and Automation Letters*, vol. 9, no. 1, pp. 907–914, 2023.

[3] 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 *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2023, pp. 4226–4233.

[4] W. Dong, P. Roy, and V. Isler, "Semantic mapping for orchard environments by merging two-sides reconstructions of tree rows," *Journal of Field Robotics*, vol. 37, no. 1, pp. 97–121, 2020.

[5] A. K. Burusa, E. J. van Henten, and G. Kootstra, "Attention-driven active vision for efficient reconstruction of plants and targeted plant parts," *arXiv preprint arXiv:2206.10274*, 2022.

[6] T. Zaenker, C. Smitt, C. McCool, and M. Bennewitz, "Viewpoint planning for fruit size and position estimation," in *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2021, pp. 3271–3277.

[7] 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 *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2023, pp. 4197–4203.

[8] A. Asgharivaskasi and N. Atanasov, "Semantic octree mapping and shannon mutual information computation for robot exploration," *IEEE Transactions on Robotics*, 2023.

[9] P. Gao, J. Jiang, J. Song, F. Xie, Y. Bai, Y. Fu, Z. Wang, X. Zheng, S. Xie, and B. Li, "Canopy volume measurement of fruit trees using robotic platform loaded lidar data," *IEEE Access*, vol. 9, pp. 156 246–156 259, 2021.

[10] J. Dong, J. G. Burnham, B. Boots, G. Rains, and F. Dellaert, "4d crop monitoring: Spatio-temporal reconstruction for agriculture," in *2017 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2017, pp. 3878–3885.

[11] A. K. Nellithimaru and G. A. Kantor, "Rols: Robust object-level slam for grape counting," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, 2019, pp. 0–0.

[12] 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)*. IEEE, 2019, pp. 3890–3897.

[13] 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)*. IEEE, 2021, pp. 1–7.

[14] A. Hornung, K. M. Wurm, M. Bennewitz, C. Stachniss, and W. Burgard, "Octomap: An efficient probabilistic 3d mapping framework based on octrees," *Autonomous robots*, vol. 34, pp. 189–206, 2013.

[15] J. Delmerico, S. Isler, R. Sabzevari, and D. Scaramuzza, "A comparison of volumetric information gain metrics for active 3d object reconstruction," *Autonomous Robots*, vol. 42, no. 2, pp. 197–208, 2018.

[16] D. Bolya, C. Zhou, F. Xiao, and Y. J. Lee, "Yolact: Real-time instance segmentation," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 9157–9166.

[17] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," in *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, ser. KDD'96. AAAI Press, 1996, p. 226–231.

[18] QualiaT, "Husky ur3 simulator," 2023, https://github.com/QualiaT/husky_ur3_simulator.

[19] E. Olson, "AprilTag: A robust and flexible visual fiducial system," in *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, May 2011, pp. 3400–3407.