Overlapped Coffee Beans Detection and Localization Using a Low-Cost 3D Monocular Point Cloud Clustering Method

Isack Farady[1], Alifya Febriana[1], Chih-Yang Lin[2]
[1]Department of Electrical Engineering, Yuan Ze University, Taiwan
[2]Department of Mechanical Engineering, National Central University, Taiwan
andrewlin@ncu.edu.tw

***Abstract*—Coffee beans present significant challenges for automated analysis due to their small size, overlapping arrangements, and distinctive morphological features. Most existing approaches rely on 2D image classification, typically distinguishing beans as either good or defective, but they do not adequately address occlusion and overlap—critical issues in real-world processing. Furthermore, many current methods depend on costly equipment such as LiDAR and depth cameras, limiting their practical adoption in agricultural contexts. This study proposes a low-cost monocular image–based framework for overlapped coffee bean detection and 3D localization. The approach integrates: (1) instance segmentation neural networks with monocular depth estimation to generate Pseudo-LiDAR point clouds; (2) projecting each 2D bounding box into a 3D pyramidal region of interest and clustering the enclosed points to isolate beans, followed by fitting 3D bounding boxes; (3) automatic pseudo-label generation from clustered point clouds to reduce manual annotation; and (4) tone-mapping techniques to enhance color contrast and improve detection robustness under challenging lighting conditions.**

I. INTRODUCTION

Coffee is one of the world's most widely consumed beverages, cultivated in over 70 tropical countries and supported by an estimated 12 million small-scale growers. Despite its global importance, coffee bean sorting remains largely manual in many regions. Manual sorting often leads to price fluctuations and inconsistencies, particularly since not all farmers can visually distinguish between bean varieties without errors [4]. As coffee evolves beyond a commodity into a subject of scientific research and sensory exploration [1, 8, 9], there is growing demand for automated and reliable methods to support quality control.

Deep learning offers a promising solution for rapid and accurate coffee bean analysis, reducing reliance on manual labor. Advances in 2D object detection [7, 8] and convolutional neural networks (CNNs) [10, 12] have enabled tasks previously deemed unfeasible. However, most existing research has concentrated on binary classification (e.g., good vs. defective beans), whereas real-world applications require differentiation between multiple varieties (e.g., peaberry, longberry, premium). More critically, current methods struggle with occlusion, overlapping beans, and background similarity, all of which significantly reduce accuracy in practice.

Therefore, rather than focusing solely on classification, the aim of this work is to address the more challenging problem of 3D detection and localization of overlapping coffee beans using a cost-effective approach. Specifically, we propose a monocular image-based framework that integrates: (1) instance segmentation (Mask R-CNN [5]) for 2D detection (2) monocular depth estimation (Pseudo-LiDAR [3, 13]) for 3D point cloud generation, DBSCAN clustering [2] for 3D box pseudo-labeling. This eliminates the need for costly depth or LiDAR sensors while enabling robust handling of occluded beans. Additionally, tone-mapping techniques are introduced to improve low-light visibility and enhance segmentation and depth estimation performance. Our contributions are summarized as follows:

- A novel hybrid 2D-to-3D detection framework combining instance segmentation and 3D point cloud clustering for precise localization of green coffee beans.
- Adaptation of LiDAR-based detection algorithms for agricultural use, enabling detailed 3D modeling of coffee beans without stereo cameras. Our Pseudo-LiDAR representation bridges the performance gap between monocular and LiDAR-based systems.
- Automated 3D pseudo-labeling to reduce manual annotation efforts and a tone-mapping technique to enhance low-light image visibility, improving detection robustness.

The remainder of this paper is organized as follows: Section 2 details the experimental setup, dataset, and methodology; Section 3 evaluates 2D and 3D detection performance; and Section 4 presents conclusions.

979-8-3315-7206-8/25/$31.00 ©2025 IEEE

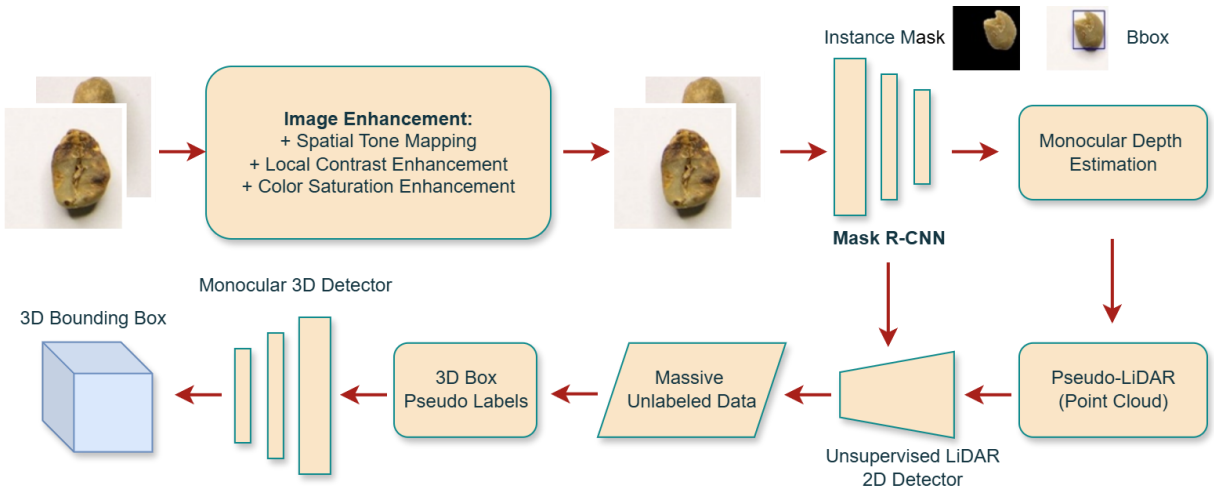

Fig. 1. Overview of the proposed framework for overlapped coffee bean detection and 3D localization.

II. Data Collection and Methodology

2.1. Overall Framework

Based on the flowchart in Fig. 1, our pipeline comprises five sequential stages. First, we enhance each monocular RGB image using spatial tone mapping, local contrast enhancement, and color saturation adjustment. Second, a Mask R-CNN [5] branch performs 2D instance segmentation, exporting per-instance masks (NPZ files) and bounding-box predictions (TXT files). Third, a parallel monocular depth-estimation branch generates a dense Pseudo-LiDAR point cloud. Fourth, these 2D outputs and the point cloud feed into an unsupervised LiDAR-style 3D detector: we carve frustums from each 2D box, cluster the frustum points with DBSCAN, retain the largest cluster as the object, and fit minimal 3D bounding boxes—saving these as pseudo-labels. Finally, we train a Monocular 3D Detector on massive unlabeled RGB data paired with the generated 3D pseudo-labels, yielding end-to-end 3D bounding-box predictions directly from single-view images.

2.2. Data Acquisition

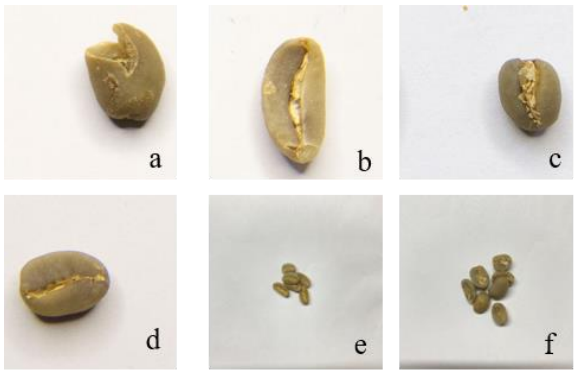

Fig. 2. Example images of the collected dataset: (a) defect, (b) longberry, (c) peaberry, (d) premium, and (e–f) overlapped beans.

The data collected consists of RGB image data of Arabica coffee beans with the types of green bean defect, longberry, peaberry, and premium. Because publicly available images were insufficient, we manually captured the data using a Canon EOS 60D digital camera and an iPhone 11 smartphone [4].

The image data collection is conducted by taking photos of defects, longberry, peaberry, and premium varieties of coffee beans. These coffee beans are acquired by purchasing them directly from coffee farmers in Indonesia, and have been divided into four separate locations according to their types. The data collection process is divided into two parts: capturing data with individual coffee beans in one image and capturing data with multiple coffee beans in one image. For the visual representation of the dataset, refer to the images in Fig. 2. To ensure balanced learning, we split the data into training and test sets, yielding 6399 single-bean and 280 multi-bean training images, plus 1600 and 70 test images, respectively.

2.3. Tone Mapping Techniques

As illustrated in Fig. 3, we apply four complementary tone-mapping techniques to enhance our RGB images before segmentation and depth estimation. First, photometric masking

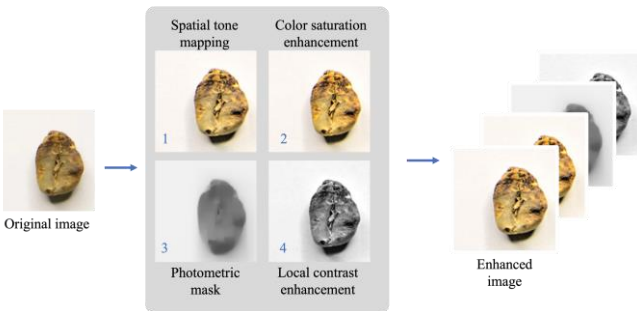

Fig. 3. Tone mapping technique results (1) photometric mask, (2) color saturation, (3) photometric mask and (4) local contrast enhancement.

performs a four-pass, edge-aware blur to emphasize local detail while avoiding halo artifacts. Second, spatial tone mapping adjusts each pixel's brightness and contrast based on its immediate neighborhood, preserving edge sharpness. Third, global brightness & saturation mapping remaps the grayscale guide image to control overall luminance and selectively boost color saturation in darker regions. Finally, local contrast enhancement increases mid-tone contrast via localized gain control, further improving visibility of subtle textures and small objects. By concatenating the outputs of these methods with the original image (see Fig. 3), we provide Mask R-CNN and our monocular depth estimator with rich, high-contrast inputs that facilitate more accurate 2D segmentation and 3D point-cloud reconstruction.

2.4. 2D Instance Segmentation

Fig 4. shows the workflow of the Mask R-CNN. We applied ResNet-50 [6] to extract the features of the image of coffee beans, the preprocessed image is first fed into the trained backbone network. Regions of interest (ROIs) are then created by setting the retrieved eigenvalues to the region proposal network (RPN). Third, the preselected box's position coordinates are used to pool the matching region into a fixed size in the feature map using the ROIAlign layer. In order to achieve coffee identification and segmentation, N-class classification, bounding box regression, and mask generation are employed. After training the data, we generated two files containing mask with NPZ files and bounding box prediction in txt files that will be used later to generate the 3D unsupervised detector

2.5. Monocular Depth Estimation

We chose MiDaS (Multiple-depth Estimation Accuracy with Single Network) [15] for monocular depth estimation. This model has undergone thorough evaluation for its

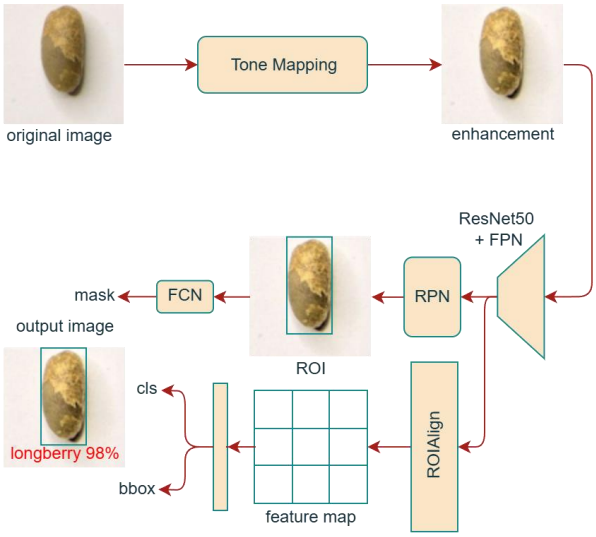

Fig. 4. The structure of Mask R-CNN in our proposed work.

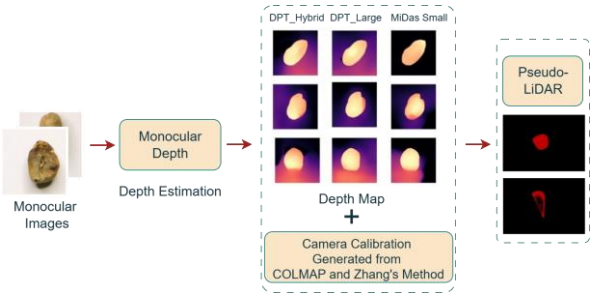

Fig. 5. Workflow of 3D point cloud generation from monocular images using depth estimation and camera calibration.

reliability and versatility through zero-shot cross-dataset transfer, where it performed well on datasets it hadn't encountered during training. MiDaS is a deep learning-based residual model built atop ResNet for monocular depth estimation. We use the most recent versions of the DPT Model transformer models, including DPT_large, DPT_hybrid and MiDaS_small [16].

2.6. 3D Point Cloud Generation

Fig. 5. Shows a comprehensive overview of the process involved in generating a 3D point cloud from monocular images. Before generating the point cloud, it is essential to obtain the calibration matrix. For the known camera, we apply traditional calibration estimation methods from Zhang [14], it uses checkerboard patterns and then captures the checkerboard patterns with different angles. On the other hand, for the unknown camera, we employ the self-calibration method from COLMAP [11] which adopts the SfM (Structure from Motion) strategy and employs self-calibration techniques to estimate camera parameters and 3D structure from scenes.

To compute the 3D coordinates (X, Y, Z) in the camera coordinate system for each pixel (u, v), we combine the estimated depth map with the intrinsic parameters of the camera. The variable Z represents the distance from the camera to the object point. The 3D coordinates are calculated as follows:

$$X = \frac{(u-c_x)\times Z}{f_x} \text{ and } Y = \frac{(v-c_x)\times Z}{f_y} \quad (1)$$

Where, $f_{x,y}$ is the focal lengths of the camera, $c_{x,y}$ is the coordinates of the principal point (optical center) of the image.

2.7. Unsupervised LiDAR Monocular 3D Detector and 3D Box Pseudo-Label Generation

In the unsupervised LiDAR 3D detector pipeline, the input data includes LiDAR point cloud data, camera calibration information, mask estimation, and 2D bounding box detector. Initially, the segmented bounding boxes and binary masks are used to identify objects in the scene. LiDAR points are filtered

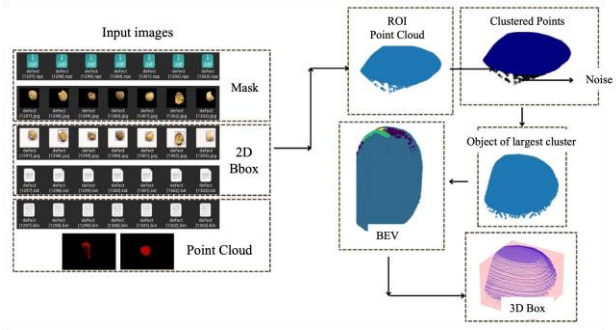

Fig. 6. Framework of unsupervised LiDAR monocular 3d detector.

based on their presence within the field of view and their alignment with the 2D image from the segmented mask. Next, points from the point cloud corresponding to each segmented object in the mask estimation are selected. As shown in Fig. 6, these estimates help build camera frustums to select relevant LiDAR ROI points for each object.

However, the point cloud still contains background, noise, and occluded points within the same frustum as the object points. To eliminate irrelevant points, DBSCAN clusters the ROI points based on density, selecting the largest cluster as the target corresponding to the object and determining the minimum 3D bounding box that covers all target points. The largest cluster in each object, determined by the number of points, is classified as a valid object, while others are labeled as noise and subsequently removed.

For the largest cluster, the minimum bounding box is calculated in the Bird's Eye View (BEV), where the vertical or *Z* component is discarded, focusing solely on *X* and *Y* coordinates. The convex hull, the smallest convex shape enclosing all points in the point cloud, is then applied to identify the outer boundary of each cluster of points. Following this, the minimum bounding box in the BEV view, which encloses the LiDAR points while minimizing its area, is computed. This minimum bounding box is transformed back to the 3D space, considering the *Z* dimensions, to generate 3D label information and visualize the result.

Table 1. Precision, recall, AP, and mAP of the Mask R-CNN.

2D Images	Original	Enhancement
Precision	0.978	0.98
Recall	0.956	0.993
AP	0.922	0.977
mAP	0.95	0.979

Table 2. Quantitative comparison of monocular depth estimation model on MiDaS_small.

Metrics	Original Images	Image Enhancement
Abs_rel	2.817	2.337
Sq_rel	1.131	0.934
Rmse	0.286	0.332
Rmse_log	1.257	1.316
δ<1.25	0.239	0.219
δ<1.25**2	0.434	0.400
δ<1.25**3	0.572	0.533

III. Experiments and Results

3.1. 2D Detection Results

We trained the Mask R-CNN [12] model—fine-tuning its head layers—over 5000 epochs, monitoring both training and validation loss. As shown in Table 1, images enhanced via tone mapping outperform original images across all metrics: precision, recall, average precision (AP), and mean AP (mAP). These gains demonstrate that tone mapping substantially improves 2D detection accuracy. Predicted instance masks were exported to NPZ files, and bounding-box coordinates to TXT files, to guide subsequent 3D region-of-interest selection.

3.2. Monocular Depth Estimation Results

We compare our approach against a DPT_hybrid baseline on the same dataset. We observe that the depth maps from MiDaS_small and DPT_large: both accurately delineate bean contours and preserve surface details, whereas the baseline exhibits greater background noise and a slight bias toward nearer depths. Quantitatively (Table 2), MiDaS_small applied to tone-mapped images yields higher accuracy than when applied to original images, although the improvement is modest. These results confirm that our preprocessing enhances monocular depth estimation for coffee-bean imagery.

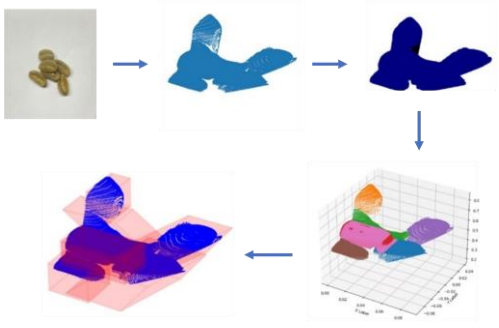

Fig.7. Process of the monocular unsupervised 3D detector.

3.3. 3D Detection and Localization Results

Table 3. Performance of the monocular unsupervised 3D detector.

Model	Precision	Recall	mAP
Original Images	88.7	93.5	84.5
Image Enhancement	89	94.9	84
YOLOv5	86	86.7	93
YOLOv7	86.6	89.3	92.6
YOLOv8	85.5	92	94.9

Our unsupervised LiDAR-style 3D detector converts a single RGB view into accurate 3D bounding boxes by clustering frustum points and fitting minimal enclosures (Fig. 7). Quantitatively (Table 3), on tone-mapped data we achieve 89 % precision and 94.9 % recall—an improvement of +7 pp and +9 pp, respectively, over standard 2D detectors. These gains stem from two key factors: (1) depth-informed separation, where even visually overlapping beans occupy distinct 3D clusters; and (2) noise suppression, since DBSCAN discards isolated or background points that would otherwise trigger false positives in 2D.

Qualitatively, our method excels at disambiguating tight bean groupings and partially occluded specimens. In Fig. 8, several instances that 2D models either merge or miss entirely are correctly identified and localized in 3D space by our pipeline. This robustness reduces the need for costly manual box correction and enables more reliable downstream analytics (e.g., per-variety yield estimation).

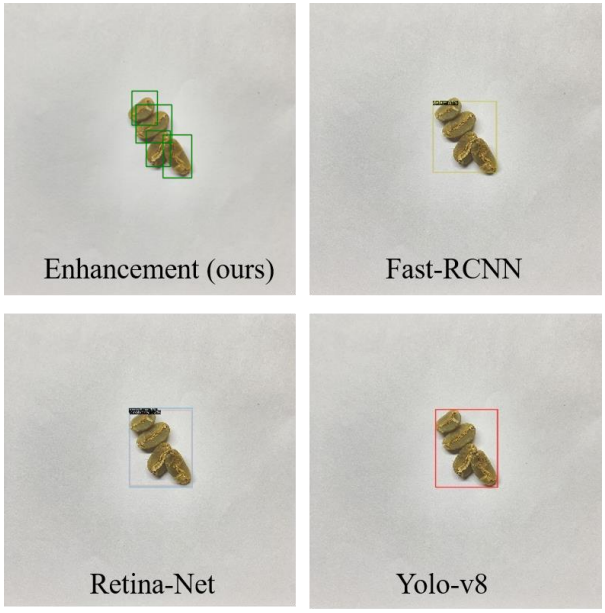

Fig. 8. Qualitative comparison result between the proposed method and 2D object detection models.

Nonetheless, some limitations persist. First, over-segmentation can occur when beans are densely packed: DBSCAN may split a single object into multiple clusters if ε (neighborhood radius) is set too small. Second, projection artefacts occasionally enlarge the re-projected 2D boxes, since the convex-hull fitting in BEV neglects fine-grained object contours. Third, camera calibration errors propagate into the point cloud, introducing slight drift in box placement.

Future work will address these issues by exploring adaptive clustering thresholds, incorporating a refinement stage that aligns 3D boxes with image masks, and integrating uncertainty estimation in the monocular depth branch. Such enhancements will further tighten bounding-box accuracy and extend applicability to other small-object agricultural scenarios.

IV. CONCLUSION

Classification and 2D object detection are two areas where deep learning dominates. However, it frequently fails to identify overlapping small objects, especially when dealing with small objects like coffee beans. We provide a low-cost technique based on monocular images to identify four types of coffee beans: defect beans, longberry beans, peaberry beans, and premium coffee beans, in order to overcome this limitation. Our approach aims to address issues like coffee bean occlusion and overlap. To do this, we eliminate the need for extra costly equipment like depth and LiDAR cameras by combining instance segmentation neural networks with depth map estimation and 3D point cloud data (LiDAR data). We are able to manage occluded and overlapping beans and distinguish their front and back sides by utilizing 3D point cloud data. To improve detection in difficult lighting situations, we also utilize tone mapping techniques. Our work presents a viable solution to the problems related to small objects detection in cluttered situations by effectively addressing the possibility for reliably and efficiently detecting overlapped coffee beans.

ACKNOWLEDGMENT

This work was supported by funding from the National Science and Technology Council (NSTC), Taiwan, under grants NSTC 111-2221-E-008-110-MY3 and NSTC 114-2923-E-008-003-MY3.

REFERENCES

[1] S.-J. Chang and C.-Y. Huang, “Deep learning model for the inspection of coffee bean defects,” *Applied Sciences*, vol. 11, no. 17, pp. 8226, 2021.

[2] D. Deng, "DBSCAN clustering algorithm based on density," *2020 7th international forum on electrical engineering and automation (IFEEA)*, pp. 949-953, 2020.

[3] A. Febriana, I. Farady, P.-C. Lin, C.-Y. Lin, and K. Muchtar, "Pseudo-LiDAR Meets Agriculture: Leveraging 3D Monocular Point Cloud Processing for Coffee Beans," *2023 2nd International Conference on Computer System, Information Technology, and Electrical Engineering (COSITE)*, pp. 84-89, 2023.

[4] A. Febriana, K. Muchtar, R. Dawood, and C.-Y. Lin, "USK-COFFEE dataset: a multi-class green arabica coffee bean dataset for deep learning," *2022 IEEE international conference on cybernetics and computational intelligence (CyberneticsCom)*, pp. 469-473, 2022.

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

[6] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770-778, 2016.

[7] N.-F. Huang, D.-L. Chou, and C.-A. Lee, "Real-time classification of green coffee beans by using a convolutional neural network," *2019 3rd international conference on imaging, signal processing and communication (ICISPC)*, pp. 107-111, 2019.

[8] C.-J. Kuo, D.-C. Wang, T.-T. Chen, Y.-C. Chou, M.-Y. Pai, G.-J. Horng, M.-H. Hung, Y.-C. Lin, T.-H. Hsu, and C.-C. Chen, "Improving defect inspection quality of deep-learning network in dense beans by using hough circle transform for coffee industry," *2019 IEEE International Conference on Systems, Man and Cybernetics (SMC)*, pp. 798-805, 2019.

[9] P. Poltronieri and F. Rossi, "Challenges in specialty coffee processing and quality assurance," *Challenges*, vol. 7, no. 2, pp. 19, 2016.

[10] J. P. Rodríguez, D. C. Corrales, J.-N. Aubertot, and J. C. Corrales, "A computer vision system for automatic cherry beans detection on coffee trees," *Pattern Recognition Letters*, vol. 136, no. 1, pp. 142-153, 2020.

[11] J. L. Schonberger and J.-M. Frahm, "Structure-from-motion revisited," *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4104-4113, 2016.

[12] Y. Unal, Y. S. Taspinar, I. Cinar, R. Kursun, and M. Koklu, "Application of pre-trained deep convolutional neural networks for coffee beans species detection," *Food Analytical Methods*, vol. 15, no. 12, pp. 3232-3243, 2022.

[13] Y. Wang, W.-L. Chao, D. Garg, B. Hariharan, M. Campbell, and K. Q. Weinberger, "Pseudo-lidar from visual depth estimation: Bridging the gap in 3d object detection for autonomous driving," *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8445-8453, 2019.

[14] Z. Zhang, "A flexible new technique for camera calibration," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 22, no. 11, pp. 1330-1334, 2002.