Master Thesis

Thesis Title: Vision-Guided Robotic System for Basil Harvesting with Machine Learning

by

Marin Marian
(mmn519)

First Supervisor: Kevin S. Luck
Daily Supervisor: Tony Currà
Second Reader: Agoston E. Eiben

July 16, 2025

Submitted in partial fulfillment of the requirements for
the VU degree of Master of Science in Artificial Intelligence

Vision-Guided Robotic System for Basil Harvesting with Machine Learning

Marin Marian

Vrije Universiteit Amsterdam, Amsterdam

Abstract. This thesis presents the development of a low-cost, vision-guided robotic system for selectively harvesting basil using only monocular RGB input. The proposed system integrates a fine-tuned Segment Anything Model (SAM), adapted with Low-Rank Adaptation (LoRA) to perform real-time petiole segmentation without prompts. An image-based visual servoing (IBVS) strategy guides a custom-built, 3D-printed robotic arm to perform precise centering, approach, and cutting actions. The full pipeline operates autonomously, without depth sensors, and is powered by an ESP32-S3 microcontroller.
Evaluated in repeated indoor trials on real basil plants, the system achieved reliable petiole segmentation (Dice score: 0.63), consistent alignment using visual servoing, and a cutting success rate of approximately 40%. The results show that prompt-free segmentation models and 2D visual feedback can enable effective robotic manipulation even with limited hardware. With a total hardware cost under €50, this work demonstrates that low-cost, vision-based harvesting is technically feasible and provides a replicable foundation for future research in accessible agricultural robotics.

Keywords: Robotic harvesting · Petiole segmentation · Image-Based Visual Servoing · Segment Anything Model · Low-cost agriculture

1 Introduction

According to a report by Maucorps et al. [24], approximately 2.5 million workers have left the agricultural sector across the European Union over the past decade. Moreover, the same report projects that the agricultural workforce will continue to shrink by around 2% annually until 2030. This significant labour shortage, coupled with the increasing demand for sustainable agricultural practices, has fuelled a growing interest in automating tasks within agriculture [28]. Among these tasks, selective harvesting of delicate crops such as basil presents a unique challenge due to the need for precision. Although industrial automation has made considerable advances in large-scale farming, precision herb harvesting in greenhouses or indoor environments remains largely manual and time-consuming, primarily due to the key challenge of navigating highly cluttered crop environments [15]. Basil is a widely cultivated herb that requires regular harvesting to promote leaf regrowth and maintain quality [31]. Unlike whole-plant harvesting used for

crops like wheat or lettuce, where the entire plant is cut at once, basil harvesting is selective, targeting only mature leaves while preserving the rest of the plant. This demands fine-grained perception and precision.

Recent advances in computer vision and robotic manipulation offer promising tools for tackling the challenges of precision herb harvesting, particularly the problem of accurately identifying where to cut in dense environments. In crops like basil, cutting requires not only detecting the mature leaf to cut but also precisely localising the optimal cutting point to avoid damaging nearby leaves or stems. Deep learning segmentation models, such as the Segment Anything Model (SAM) [14], enable accurate object localisation like detecting and isolating specific plant parts, such as petioles or mature leaves, even in cluttered scenes [3]. However, SAM was originally designed to generate segmentation masks in response to a prompt (e.g., a point, box, or text) [14], which makes it unsuitable for fully autonomous robotics. To overcome this limitation and enable operation without user input, large models like SAM must be adapted to new, highly specific tasks through targeted fine-tuning. Techniques such as Low-Rank Adaptation (LoRA) [9] make it possible to fine-tune large pre-trained models using only a small number of parameters and minimal annotated data, making them particularly useful in agricultural applications where high-quality labels are hard to get. These methods allow for the adaptation of models like SAM into specialised tools that can detect basil petioles and cutting regions.

Detecting where to cut is only part of the challenge. Once a viable cutting point is identified, a robotic arm must be able to move toward the target with high precision. This requires a control strategy that can handle visual variability and uncertainty in depth estimation. Depth cameras can provide explicit 3D information but they are expensive and large, making them less suitable for compact, low-cost robots. They can affect balance and movement precision when mounted on lightweight robotic arms [13]. To overcome this limitation, the project uses a single RGB camera and a method known as Image-Based Visual Servoing (IBVS). IBVS uses real-time visual feedback to guide robot movements, aligning the end effector with the target in the camera frame [8]. Because monocular RGB cameras do not provide explicit depth information, recent studies have proposed heuristic area-based approaches or monocular depth estimation techniques to approximate the distance between the camera and the target object [4,12]. These approaches, when combined with visual servoing, reduce both hardware cost and system complexity, allowing even low-cost robotic arms to perform accurate positioning and cutting.

This thesis presents the development of such a system: a low-cost, vision-guided robotic platform designed to selectively harvest basil by integrating deep learning-based segmentation with image-based visual servoing. The proposed system uses only a monocular RGB camera and avoids expensive depth sensors, making it accessible and efficient for small-scale agricultural applications.

The robotic arm used in this project is a custom-built, lightweight, 3D printed manipulator driven by five servomotors: one for base rotation, three for shoulder, elbow, and wrist movement, and one for operating the end effector, which

is a small cutting mechanism designed to trim basil petioles with precision. A low-resolution camera is mounted just above the cutter, providing a live video stream that is sent to a PC that runs the vision system. The robotic arm is controlled by a microcontroller, which receives continuous updates from the vision module. This setup prioritises affordability, making it accessible for small-scale agriculture and academic research with limited budgets.

This paper presents the design, implementation, and evaluation of a low-cost robotic system for basil harvesting. The system operates in three stages: (1) detection and segmentation of petiole regions using a fine-tuned SAM model; (2) vision-based centering and approach using segmentation area feedback and image-based visual servoing; and (3) execution of the cutting action once the target is within reach. Building on this pipeline, the thesis addresses two central research questions:

1. (How well) Can a fine-tuned SAM model accurately identify petioles and optimal cutting points in basil plants in real time?
2. (To what extent) Can image-based visual servoing combined with segmentation-based depth estimation guide a low-cost robotic arm to position itself to perform precise cutting of delicate plant structures like petioles, without dedicated depth sensors?

2 Related Work

Crop and Leaf Detection Earlier harvesting systems often focused on detecting entire leaves or fruits. Commonly targeted crops include strawberries, tomatoes, sweet peppers and lettuce, where segmentation plays a central role in guiding robotic manipulators to perform picking or cutting actions [5,11,18,39]. However, such approaches often relied on locating the leaf or fruit without a clear strategy for determining a precise and safe cutting point. Similarly, this project began by exploring leaf detection for identifying mature leaves ready to harvest, but it quickly became evident that determining an appropriate cutting point remains difficult due to occlusion by overlapping leaves.

Cutting Point Estimation To address the challenge of locating cutting points more accurately, image processing and geometry-based techniques have been explored. Principal Component Analysis (PCA) has been used to infer leaf orientation, while skeletonization was applied to extract mid-veins or structural lines that could guide cutting decisions [27,38]. Attempts were also made to apply 3D reconstruction techniques from monocular images to recover the spatial structure of the plant [37]. However, due to visual occlusion and ambiguity in leaf attachment points, these methods proved unreliable in practice. As a result, the focus shifted from leaf detection to a more targeted approach: petiole segmentation.

Petiole Segmentation for Precision Harvesting By isolating the petiole, the narrow stem connecting a basil leaf to the main plant, it becomes possible to

directly identify a suitable cutting region with greater precision. Giang et al. [6] used semantic segmentation to identify pepper stems, petioles, and leaves, enabling accurate pruning. Although fewer studies have targeted petiole segmentation specifically, this shift aligns with recent insights that emphasize task-specific structure detection for fine manipulation tasks in agricultural robotics [17].

Deep Learning for Plant Segmentation To implement petiole segmentation robustly, deep learning methods were adopted. Mask R-CNN has been widely used to segment fruits and crops, achieving high instance-level accuracy [23,30]. DeepLab and U-Net models have proven effective in semantic segmentation for plant health monitoring and disease detection [7,10,29]. YOLO models can offer real-time inference tasks on low-power devices [2,30,34], but they struggle with thin petiole regions.

Foundation Models and Adaptation in Agriculture More recently, researchers have explored adapting foundation models such as the Segment Anything Model (SAM) to agriculture [3]. While SAM performs well in zero-shot general segmentation, its accuracy on crop-specific structures such as petioles, leaves, or stems is limited without domain adaptation. To improve this, Li et al. [21] proposed the Agricultural SAM Adapter (ASA), which adds lightweight adapters to SAM and dramatically improved accuracy across all their 12 agricultural segmentation tasks. Similarly, Li et al. [19] introduced EMSAM, which fine-tunes SAM using multi-scale adapters, achieving over 21% points improvement in IoU for fine-grained leaf lesion detection.

In this project, SAM was fine-tuned using Low-Rank Adaptation (LoRA) to perform precise basil petiole segmentation. This method updates only a small fraction of the model's parameters while keeping the core of SAM frozen. Wu et al. [36] used SAM-ViT-huge-LoRA for straw segmentation under few-shot learning conditions and achieved an F1-score of 83.6%, outperforming all other baselines, while requiring just 0.65% of the training data. Their approach proved especially beneficial in environments with scarce labelled data, demonstrating that LoRA-based SAM tuning can achieve high segmentation accuracy even with minimal supervision. Similarly, Song et al. [33] introduced a multistage fine-tuning strategy (MAF-SAM) combining prefix adapters and LoRA layers to adapt SAM for multispectral crop segmentation. Their LoRA stage significantly boosted the model's ability to extract crop-specific semantics, achieving F1 scores above 0.92 for soybean segmentation. These studies show that SAM, when fine-tuned with efficient strategies like LoRA, can outperform traditional models like U-Net or DeepLab in plant segmentation. In line with these findings, the fine-tuned SAM used in this work outperformed other tested models, including DeepLabv3, Mask R-CNN, YOLOv11-seg, in detecting narrow petiole structures, making it particularly suitable for selective basil harvesting.

Visual Servoing and Monocular Depth Estimation Many agricultural robot manipulators are using Image-Based Visual Servoing (IBVS) to reach a

target position [32]. Traditional systems often incorporate stereo vision or RGB-D sensors to obtain range information, since deriving precise distance from a single 2D view can be challenging [32]. However, adding LiDAR or dual-cameras raises cost and complexity, which conflicts with the goal of lightweight, low-cost robots. Recent research has therefore revisited monocular IBVS in agriculture, a single RGB camera approach being simpler and more convenient [22].

A key challenge with monocular guidance is estimating target distance. Researchers have proposed using visual cues like object size in the image to infer depth. For example, Liu et al. [22] developed a pixel-area regression method: as the camera moves, changes in the fruit's segmented area correlate with distance, providing a low-computation ranging solution. Such monocular techniques achieved high accuracy (e.g. within 6% error for fruit distance estimation [20]), validating that segmentation masks or known object scale can replace stereo triangulation. The present work follows this paradigm: it employs IBVS with a single RGB camera (no depth sensor), extracting the basil petiole's pixel area to determine distance.

3 Methodology

This section describes the system architecture and implementation of the proposed vision-guided robotic system for basil harvesting. The system consists of a vision module, a control module, and a physical robotic manipulator. The vision module runs on a PC and processes a real-time video stream captured by an OV2640 RGB camera mounted above the cutter. It performs petiole segmentation, identifies cutting points, and estimates depth. These outputs are sent via serial communication to an ESP32-S3 microcontroller, which controls the five-joint robotic arm. The robot executes centering, approaching, and cutting motions based on the received visual feedback. An overview of the system architecture is shown in Fig. 1.

3.1 Visual Perception

The visual perception module is built around the Segment Anything Model (SAM) [14], originally designed to perform segmentation given a user prompt, such as a point, box, or text. SAM cannot perform segmentation autonomously which makes it incompatible with real-time robotic applications. To remove the prompt dependency, SAM's mask decoder was fine-tuned on a task-specific dataset, and the prompt encoder was disabled. The result is a fully automatic segmentation model that outputs binary masks directly from raw RGB images.

The fine-tuning was performed using Low-Rank Adaptation (LoRA) [9], a parameter-efficient method that inserts trainable low-rank matrices into the attention layers of the transformer blocks in SAM's decoder. This allows the model to learn task-specific features while keeping the vast majority of the original parameters frozen, significantly reducing memory and compute requirements. The training dataset consisted of 97 annotated RGB images of basil plants, where

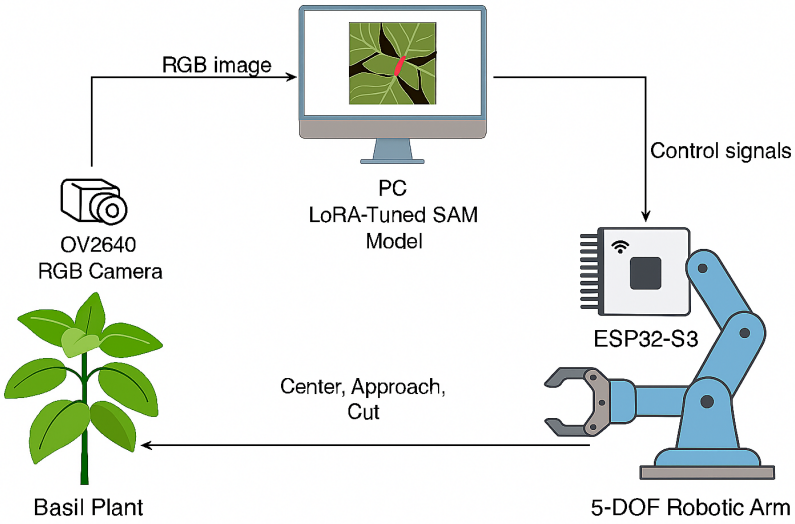


Fig. 1: System overview

the petiole regions were labelled with pixel-accurate binary masks. The model was trained using a combination of Dice loss and binary cross-entropy. Dice loss measures the overlap between the predicted segmentation mask and the ground truth, making it particularly effective when the foreground (petiole) occupies only a small part of the image [25]. Binary cross-entropy evaluates pixel-wise classification accuracy and helps the model learn fine-grained details [1].

During inference, the live RGB stream from the OV2640 camera is sent to the PC, where each frame is processed through the LoRA-tuned SAM. The output is a binary mask indicating the petiole. The system then computes the centroid of the mask, which is used as the fixed 2D target for the cutting action. Both the mask and the cutting point are updated continuously across frames and serve as the visual reference for all downstream motion control.

3.2 Control Strategy

The control module uses an Image-Based Visual Servoing (IBVS) strategy [8] to guide the robotic arm based on real-time image features extracted from the visual perception module. IBVS relies on directly minimizing the error between visual features (in this case, the petiole centroid) and a desired target position in the image plane. This eliminates the need for explicit 3D reconstruction or pose estimation, making it well-suited for systems using only monocular RGB input.

Control is executed in two stages: centering and approach. During centering, the system computes the pixel offset between the centroid of the segmented

petiole mask and the center of the image frame. This 2D error vector is translated into joint commands that adjust the base, shoulder, and elbow angles to align the cutting point with the optical axis of the camera. The robot iteratively updates its pose until the target lies within a ± 10 pixel window around the image center, at which point the robot transitions to the approach phase.

In the approach phase, the robot advances toward the petiole by adjusting the shoulder, elbow, and wrist joints. The intended strategy is to monitor the growth of the segmentation mask area as a proxy for distance, based on the principle that the closer the camera is to the object, the larger the object appears in the image. However, due to the specific physical arrangement of the camera and cutting tool, the petiole begins to exit the field of view when the cutter gets very close. As a result, the system uses a practical stopping condition: the disappearance of the segmentation mask. When the mask is no longer detected in the frame after previously being centered, the robot assumes that the cutter has reached the correct distance and proceeds to execute the cut.

The cutting action is performed by triggering the fifth servo motor, which activates the end-effector blade. The arm then returns to its initial pose, and the system restarts the detection-control loop for the next harvesting cycle.

3.3 Robotic Platform

The robotic platform consists of a custom-built, lightweight manipulator designed for low-cost agricultural automation. The arm features five degrees of freedom: base rotation, shoulder pitch, elbow pitch, wrist pitch, and a fifth joint dedicated to actuating the end-effector. All joints are driven by MG90S servo motors, which are controlled using pulse-width modulation (PWM) signals generated by an ESP32-S3 microcontroller. A full view of the assembled robotic arm is shown in Fig. 2.

All structural components of the robot were 3D-printed using PLA filament, keeping material costs low while allowing rapid prototyping and replacement. The arm was assembled from modular components, making it both lightweight and easily reconfigurable. The manipulator was designed to operate in constrained environments, such as small greenhouse plots or indoor farming racks, where high precision and a compact footprint are required.

An OV2640 RGB camera is mounted on the wrist of the arm, directly above the cutter, providing a live video stream of the region in front of the end-effector. The camera streams RGB images over Wi-Fi to an external PC, which handles all image processing and control logic. The PC communicates with the ESP32-S3 microcontroller over a UART serial interface, sending motion commands in real time based on the output of the perception and control modules.

The cutting mechanism consists of a lightweight blade actuated by the fifth servo motor. The blade is mounted at the end of the arm and aligned with the camera's optical axis to simplify visual targeting. A close-up of the end-effector, showing the cutter and camera, is provided in Fig. 3.

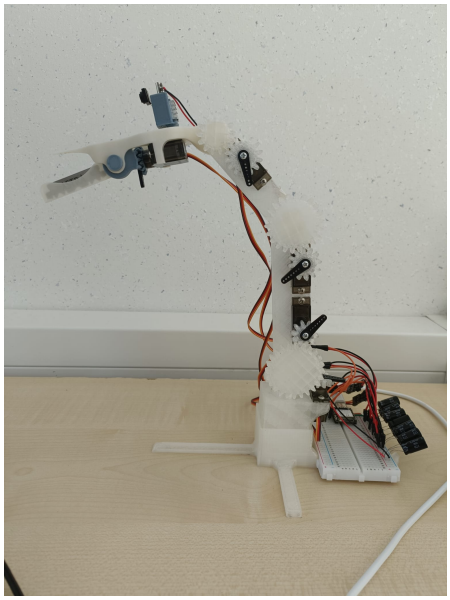


Fig. 2: Robotic arm overview

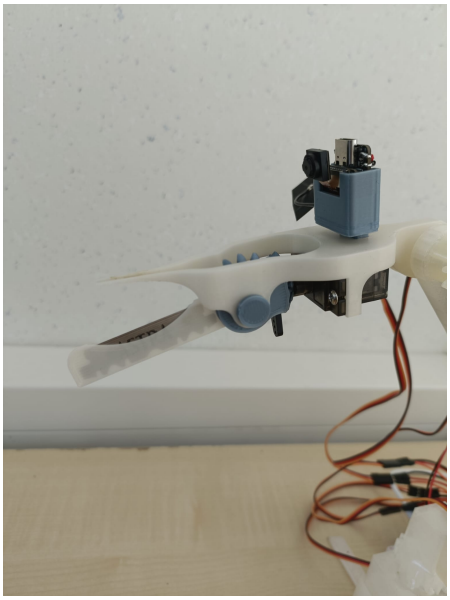


Fig. 3: End-effector and camera

The design emphasizes affordability, modularity, and ease of replication, making it suitable for research environments and small-scale automated harvesting applications.

4 Experiments and Results

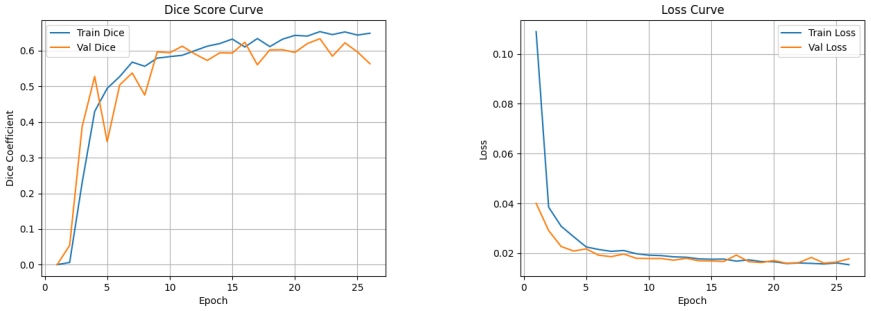
The system was evaluated through repeated trials conducted in a consistent indoor environment, using potted basil plants placed in front of the robotic arm, with natural indoor lighting and a static background. The complete harvesting pipeline was tested, consisting of the three integrated modules: visual perception, visual servoing and cutting execution. Each trial included a full detection–centering–approach–cut cycle. System performance was evaluated across three dimensions: segmentation accuracy, control behaviour, and cutting success rate.

4.1 Segmentation Performance

The segmentation model was evaluated on a held-out test set of annotated basil images. It achieved a Dice score of 0.63 and an Intersection over Union (IoU) of 0.51, indicating moderate overlap between predicted masks and ground truth. One key challenge was the visual similarity between the petiole and its background, particularly when the background was another leaf of the same plant.

In these cases, the model was still able to detect the petiole region, but the predicted mask was not always sharply defined, which affected metric scores. This was primarily due to the low resolution of the OV2640 camera, which made it difficult to distinguish fine structural details in cluttered scenes.

Nevertheless, the segmentation output was consistent enough to provide usable visual features and support downstream control in most trials. Figure 4 shows the evolution of the Dice score and loss over the training epochs, while Fig. 5 presents an example of petiole segmentation during inference, with the computed cutting point (red dot) and frame center (green cross) overlaid.



(a) Dice Score over epochs (b) Loss over epochs

Fig. 4: Training performance of the segmentation model: (a) Dice score and (b) loss curves.

Fig. 5: Petiole segmentation output with cutting point and frame center overlaid

4.2 Servoing and Control Behaviour

The visual servoing module used the 2D centroid of the segmentation mask as input for alignment. During the centering phase, the robot successfully minimized the offset between the centroid and the image center. This part of the control pipeline worked reliably and consistently across trials, with the end-effector almost always aligning accurately with the target petiole.

In the approach phase, the robot advanced toward the petiole based on a heuristic: forward motion continued until the segmentation mask disappeared from the image. This disappearance was used as a proxy for proximity, based on the limited field of view of the camera mounted above the cutter. This method introduced some instability near the end of the approach. The petiole would occasionally exit the frame before the cutter was properly aligned, leading to premature stopping.

To better illustrate how these stages play out during operation, Figure 6 presents a complete execution sequence from a successful trial. The left column shows the robot's camera feed (including overlays for cutting point and alignment) and the right column shows an external view of the robot's position relative to the plant. The four phases captured are: (1) Centering: the robot adjusts its base and joints to bring the petiole centroid into the image center, (2) Centering complete and approaching: alignment is achieved and forward motion begins toward the petiole, (3) Approaching complete: the cutter is close to the target, with forward motion stopping once the segmentation mask is no longer visible in the camera feed, and (4) Cut: the end-effector actuates to trim the petiole.

Although centering was highly reliable, the approach phase remained sensitive to visual occlusion, camera positioning, and the limitations of monocular vision.

4.3 Cutting Success Rate

The final cutting action was executed after the segmentation mask disappeared from view, which was used as an indicator that the petiole was close enough to the cutter. When the petiole was actually within reach at that point, the cutter performed clean, successful cuts. However, in some cases, the mask disappeared too early due to occlusion or camera positioning, causing the robot to stop while still out of range.

Across all trials, the system achieved a cutting success rate of approximately 40%. The majority of failures were due to premature cutting, caused by the lack of reliable depth feedback during the final phase of the approach. Since the system relied entirely on 2D visual cues, it could not verify whether the cutter had reached the correct physical distance before triggering the cut.

Despite this, the robot was able to complete multiple autonomous harvesting cycles using only monocular RGB input and low-cost components. The total hardware cost was less than €50. The results demonstrate that while the current system is limited by its simplistic proximity estimation method, the core pipeline is functional and can be improved significantly with better depth handling.

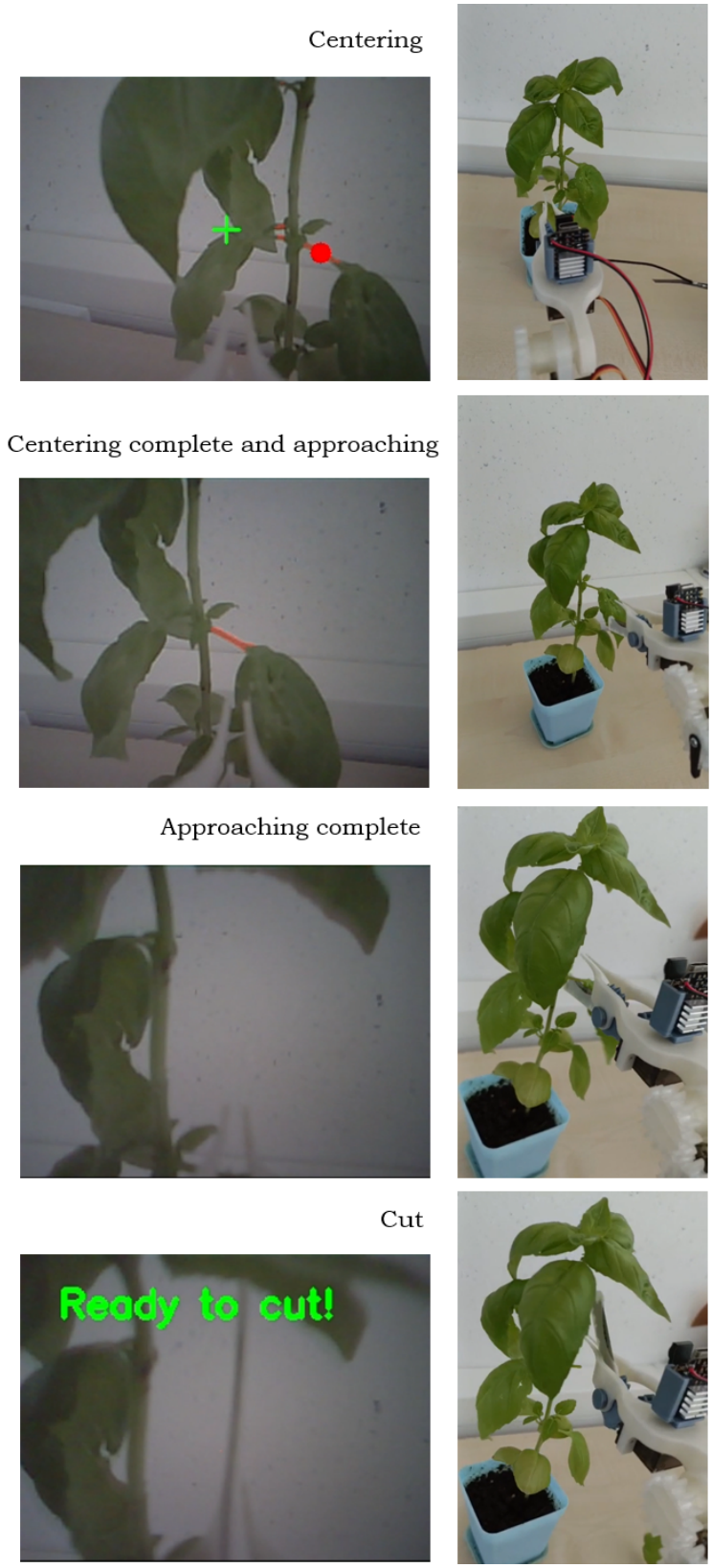


Fig. 6: Execution sequence of the basil harvesting cycle, shown from the robot's camera feed (left) and an external view (right).

5 Discussion and Conclusion

This project demonstrates that it is possible to build a functional, low-cost, vision-guided robotic system for selective basil harvesting using only monocular RGB input and minimal hardware. The integrated pipeline consisting of a LoRA-tuned SAM model, an IBVS control loop, and a lightweight 5-DOF robotic arm successfully performed complete detection-to-cut cycles autonomously in indoor settings. The results highlight both the feasibility and the current boundaries of monocular visual feedback for robotic agricultural tasks.

A core strength of the system was the performance of the fine-tuned segmentation model. Despite being trained on a small dataset of just 97 images, the LoRA-tuned SAM consistently produced accurate binary masks of basil petioles in varied scenes. The model performed well even under visual clutter and in cases where the background closely resembled the petiole. Although the predicted masks were not always pixel-perfect, particularly at the edges where the petiole overlapped with other plant structures, they were reliable and stable enough to drive real-time control. This confirms the potential of lightweight, prompt-free SAM adaptations for high-precision agricultural segmentation tasks.

The visual servoing pipeline also performed reliably during the centering phase. The IBVS controller minimized the pixel offset between the segmentation centroid and the image center with high consistency, ensuring the end-effector was properly aligned with the petiole before each approach. This validates the use of 2D image-based features for real-time robot alignment, even on systems with limited resolution and no depth sensors.

The main source of error in the system was during the approach and cutting phase. The robot advanced based on a simple heuristic: the segmentation mask disappearing from view was taken as a signal that the cutter had reached the petiole. While this approach allowed for operation without explicit depth estimation, it introduced uncertainty. In several cases, the mask disappeared too early due to occlusion or limited field of view, leading to premature cuts before the petiole was fully within reach. This constraint was the primary factor limiting the cutting success rate to approximately 40%.

Despite this, the system was able to complete multiple autonomous harvesting cycles using only RGB input and low-cost components. The total hardware cost was under €50, and the mechanical structure was entirely 3D-printed, making the platform accessible, lightweight, and easy to replicate. These features make it especially suitable for research, prototyping, and small-scale automated agriculture.

Future Work The most critical improvement lies in achieving more reliable proximity estimation during the final approach. Incorporating a second camera to enable stereo vision [16] could provide accurate depth perception and help resolve the main cause of cutting failures. While this would increase system cost slightly, it remains a feasible option for improving performance. Alternatively, training a monocular depth estimator [26] could offer a more cost-effective solution.

Refining the mechanical design, such as adjusting the camera angle or mounting position, would also help prevent occlusion at close range. On the perception side, expanding the training dataset to include more diverse lighting, angles, and occlusion cases would improve segmentation robustness even further.

Finally, a more adaptive control strategy that monitors segmentation area dynamics, confidence scores, or incorporates temporal tracking across frames [35] could improve decision-making during the approach phase. Tracking the petiole mask over time, even when partially occluded, would increase stability and reduce the risk of premature cutting caused by segmentation loss.

Conclusion This work demonstrates that vision-based, low-cost selective harvesting is technically viable using only monocular RGB input. The successful integration of a fine-tuned segmentation model with real-time visual servoing and simple hardware shows that precision robotic manipulation in agriculture does not necessarily require expensive sensors or complex control strategies.

In response to the first research question, the LoRA-tuned SAM model proved capable of accurately identifying petioles and estimating cutting points in real time, even in visually cluttered scenes. For the second question, the system confirmed that image-based visual servoing, combined with segmentation depth heuristics, can effectively guide a low-cost robotic arm to position itself for precise cutting, despite the absence of dedicated depth sensors.

While the cutting success rate was limited by the simplicity of the depth estimation method, the system establishes a clear and accessible baseline. It opens the door for scalable, affordable, and modular robotic harvesting solutions that can be refined and extended in future work.

References

1. Bishop, C.M.: Pattern recognition and machine learning. Springer (2006)
2. Bochkovskiy, A., Wang, C.Y., Liao, H.Y.M.: Yolov4: Optimal speed and accuracy of object detection (2020), https://arxiv.org/abs/2004.10934
3. Carraro, A., Sozzi, M., Marinello, F.: The segment anything model (sam) for accelerating the smart farming revolution. Smart Agricultural Technology **6**, 100367 (2023). https://doi.org/https://doi.org/10.1016/j.atech.2023.100367, https://www.sciencedirect.com/science/article/pii/S2772375523001958
4. Dong, X., Garratt, M.A., Anavatti, S.G., Abbass, H.A.: Towards real-time monocular depth estimation for robotics: A survey (2021), https://arxiv.org/abs/2111.08600
5. Ge, Y., Xiong, Y., From, P.J.: Instance segmentation and localization of strawberries in farm conditions for automatic fruit harvesting. IFAC-PapersOnLine **52**(30), 294–299 (2019)
6. Giang, T.T.H., Ryoo, Y.J.: Pruning points detection of sweet pepper plants using 3d point clouds and semantic segmentation neural network. Sensors **23**(8) (2023). https://doi.org/10.3390/s23084040, https://www.mdpi.com/1424-8220/23/8/4040

7. Gonçalves, J.P., Pinto, F.A., Queiroz, D.M., Villar, F.M., Barbedo, J.G., Del Ponte, E.M.: Deep learning architectures for semantic segmentation and automatic estimation of severity of foliar symptoms caused by diseases or pests. Biosystems Engineering **210**, 129–142 (2021). https://doi.org/https://doi.org/10.1016/j.biosystemseng.2021.08.011, https://www.sciencedirect.com/science/article/pii/S1537511021001951
8. He, L., Sun, Y., Chen, L., Feng, Q., Li, Y., Lin, J., Qiao, Y., Zhao, C.: Advance on agricultural robot hand–eye coordination for agronomic task: A review. Engineering (2025). https://doi.org/https://doi.org/10.1016/j.eng.2025.01.022, https://www.sciencedirect.com/science/article/pii/S2095809925001559
9. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: Lora: Low-rank adaptation of large language models (2021), https://arxiv.org/abs/2106.09685
10. Häni, N., Roy, P., Isler, V.: A comparative study of fruit detection and counting methods for yield mapping in apple orchards. Journal of Field Robotics **37**(2), 263–282 (Aug 2019). https://doi.org/10.1002/rob.21902, http://dx.doi.org/10.1002/rob.21902
11. Islam, S., Reza, M.N., Chowdhury, M., Ahmed, S., Lee, K.H., Ali, M., Cho, Y., Noh, D., Chung, S.O.: Detection and segmentation of lettuce seedlings from seedling-growing tray imagery using an improved mask r-cnn method. Smart Agricultural Technology **8**, 100455 (04 2024). https://doi.org/10.1016/j.atech.2024.100455
12. Jain, V., Anunay, A., Srivastava, A., Gupta, A., Aggarwal, N., Harikesh, H.: Depth estimation using monocular camera for real-world multi-object grasp detection for robotic arm. pp. 7–18 (05 2024). https://doi.org/10.1145/3634865.3634869
13. Kaleem, A., Hussain, S., Mehmood, M.A., Cheema, M.J., Saleem, S., Farooq, U.: Development challenges of fruit-harvesting robotic arms: A critical review. AgriEngineering **5**, 2216–2237 (11 2023). https://doi.org/10.3390/agriengineering5040136
14. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., Dollár, P., Girshick, R.: Segment anything (2023), https://arxiv.org/abs/2304.02643
15. Kootstra, G., Wang, X., Blok, P.M., Hemming, J., Van Henten, E.: Selective harvesting robotics: current research, trends, and future directions. Current Robotics Reports **2**, 95–104 (2021)
16. Laga, H., Jospin, L.V., Boussaid, F., Bennamoun, M.: A survey on deep learning techniques for stereo-based depth estimation. IEEE Transactions on Pattern Analysis and Machine Intelligence **44**(4), 1738–1764 (2022). https://doi.org/10.1109/TPAMI.2020.3032602
17. Lehnert, C., McCool, C., Sa, I., Perez, T.: A sweet pepper harvesting robot for protected cropping environments (2018), https://arxiv.org/abs/1810.11920
18. Lehnert, C., McCool, C., Sa, I., Perez, T.: Performance improvements of a sweet pepper harvesting robot in protected cropping environments. Journal of Field Robotics **37**(7), 1197–1223 (2020)
19. Li, J., Feng, Q., Zhang, J., Yang, S.: Emsam: enhanced multi-scale segment anything model for leaf disease segmentation. Frontiers in Plant Science **16** (03 2025). https://doi.org/10.3389/fpls.2025.1564079
20. Li, Y., Li, Y., Li, D., Zhou, D., Liu, J.: A monocular distance measurement method of orange based on the changes of target pixels. In: 2021 ASABE Annual Inter-

national Virtual Meeting. p. 1. American Society of Agricultural and Biological Engineers (2021)
21. Li, Y., Wang, D., Yuan, C., Li, H., Hu, J.: Enhancing agricultural image segmentation with an agricultural segment anything model adapter. Sensors **23**(18), 7884 (2023)
22. Liu, J., Zhou, D., Wang, Y., Li, Y., Li, W.: A distance measurement approach for large fruit picking with single camera. Horticulturae **9**(5), 537 (2023)
23. Liu, X., Wang, D., Li, Y., Guan, X., Qin, C.: Detection of green asparagus using improved mask r-cnn for automatic harvesting. Sensors **22**(23), 9270 (2022)
24. Maucorps, A., Münch, A., Brkanovic, S., Schuh, B., Dwyer, J.C., Vigani, M., Khafagy, A., Coto Sauras, M., Deschellette, P., Lopez, A., et al.: The eu farming employment: current challenges and future prospects (2019)
25. Milletari, F., Navab, N., Ahmadi, S.A.: V-net: Fully convolutional neural networks for volumetric medical image segmentation. In: 2016 Fourth International Conference on 3D Vision (3DV). pp. 565–571. IEEE (2016)
26. Ming, Y., Meng, X., Fan, C., Yu, H.: Deep learning for monocular depth estimation: A review. Neurocomputing **438**, 14–33 (2021). https://doi.org/https://doi.org/10.1016/j.neucom.2020.12.089, https://www.sciencedirect.com/science/article/pii/S0925231220320014
27. Müller-Linow, M., Pinto-Espinosa, F., Scharr, H., Rascher, U.: The leaf angle distribution of natural plant populations: assessing the canopy with a novel software tool. Plant methods **11**, 1–16 (2015)
28. Ryan, M.: Labour and skills shortages in the agro-food sector. OECD Food, Agriculture and Fisheries Papers (189) (2023)
29. Sahin, H.M., Miftahushudur, T., Grieve, B., Yin, H.: Segmentation of weeds and crops using multispectral imaging and crf-enhanced u-net. Computers and Electronics in Agriculture **211**, 107956 (2023). https://doi.org/https://doi.org/10.1016/j.compag.2023.107956, https://www.sciencedirect.com/science/article/pii/S0168169923003447
30. Sapkota, R., Ahmed, D., Karkee, M.: Comparing yolov8 and mask r-cnn for instance segmentation in complex orchard environments. Artificial Intelligence in Agriculture **13**, 84–99 (2024). https://doi.org/https://doi.org/10.1016/j.aiia.2024.07.001, https://www.sciencedirect.com/science/article/pii/S258972172400028X
31. Schuh, M., Kooyman, S.M.: Growing and harvesting basil (2020), https://extension.umn.edu/vegetables/growing-basil
32. Shamshiri, R.R., Dworak, V., ShokrianZeini, M., Navas, E., Käthner, J., Höfner, N., Weltzien, C.: An overview of visual servoing for robotic manipulators in digital agriculture. 43. GIL-Jahrestagung, Resiliente Agri-Food-Systeme pp. 231–241 (2023)
33. Song, B., Yang, H., Wu, Y., Zhang, P., Wang, B., Han, G.: A multispectral remote sensing crop segmentation method based on segment anything model using multistage adaptation fine-tuning. IEEE Transactions on Geoscience and Remote Sensing **62**, 1–18 (2024). https://doi.org/10.1109/TGRS.2024.3411398
34. Wang, C.Y., Bochkovskiy, A., Liao, H.Y.M.: Yolov7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors (2022), https://arxiv.org/abs/2207.02696
35. Wang, Q., Chang, Y.Y., Cai, R., Li, Z., Hariharan, B., Holynski, A., Snavely, N.: Tracking everything everywhere all at once. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 19795–19806 (October 2023)

36. Wu, D., Liu, C., Liu, X., Gong, Y., Bai, S., Mei, X.: Fine tuning large models for straw detection in harvested fields under few-shot learning scenarios. IEEE Geoscience and Remote Sensing Letters pp. 1–1 (2025). https://doi.org/10.1109/LGRS.2025.3548106
37. Yu, S., Liu, X., Tan, Q., Wang, Z., Zhang, B.: Sensors, systems and algorithms of 3d reconstruction for smart agriculture and precision farming: A review. Computers and Electronics in Agriculture **224**, 109229 (2024)
38. Zhang, L., Weckler, P., Wang, N., Xiao, D., Chai, X.: Individual leaf identification from horticultural crop images based on the leaf skeleton. Computers and Electronics in Agriculture **127**, 184–196 (2016)
39. Zu, L., Zhao, Y., Liu, J., Su, F., Zhang, Y., Liu, P.: Detection and segmentation of mature green tomatoes based on mask r-cnn with automatic image acquisition approach. Sensors **21**(23), 7842 (2021)