Journal of Engineering Science and Technology
Vol. 20, No. 4 (2025) 975 - 989

PRECISION ROBOTICS IN AGRICULTURE: PANOPTIC SEGMENTATION-BASED IMAGE ACQUISITION AND ANALYSIS FRAMEWORK FOR BIRD'S-EYE CHILI PLANTS

K. M. SAIPULLAH[1,2], W. H. MOHD SAAD[1,*], S. MOHD HUSNI[1], M. S. J. ABDUL RAZAK[3], W. A. O. ANWAR[4]

[1]Fakulti Teknologi dan Kejuruteraan Elektronik dan Komputer (FTKEK), Universiti Teknikal Malaysia Melaka, Hang Tuah Jaya, 76100 Durian Tunggal, Melaka, Malaysia
[2]AI Division, REKA Inisiatif Sdn. Bhd, 26A, Off, Jln Kuching, Taman City, 51200 Kuala Lumpur, Federal Territory of Kuala Lumpur, Malaysia
[3]MSJ Perwira Enterprise, 75460 Duyong, Melaka, Malaysia
[4]Faculty of Engineering, Universitas Muhammadiyah Palembang, Indonesia
*Corresponding Author: wira_yugi@utem.edu.my

Abstract

Precision robotics have become essential in streamlining and automating agricultural operations in modern farming practices. This study aims to advance precision robotics in smart agriculture by integrating pixel-wise context-aware information through a proposed panoptic segmentation-based image acquisition and analysis framework powered by Detectron2, specifically applied to a bird's-eye chili farm. The framework comprises stereo image acquisition and advanced image analysis using panoptic segmentation and instance tracking, enabling detailed segmentation, identification, and continuous tracking of each bird's-eye chili plant within a fertigation farm. Evaluation of a customized bird's-eye chili dataset revealed that the proposed framework achieved a panoptic segmentation precision of 77.4% of Average Precision at 50% Intersection over Union (AP50), outperforming the You Only Look Once (YOLO)-based approach, which reached 68.7% accuracy. The proposed target instance tracking algorithm also demonstrated an impressive Average Tracking Accuracy (ATA) of 94.9%, significantly surpassing YOLO's ATA of 73.4%.

Keywords: Agriculture precision robotics, Chili farm, Detectron2, FCNN, Panoptic, Video panoptic segmentation.

1. Introduction

The agriculture sector faces substantial challenges recruiting qualified personnel to manage crops effectively [1]. About 7% of the global population is engaged in agricultural activities on 2,781 million hectares of land [2]. As the sector grows, the demand for skilled and knowledgeable staff to ensure efficient farm management has become critical. To address the persistent workforce shortages, artificial intelligence (AI)-powered robots are increasingly being deployed as part of the broader Smart Agriculture framework, providing integrated solutions to enable green and sustainable large-scale farming [3].

One of the key components of the Smart Agriculture Framework is precision robotics, which significantly alleviates the burdens of labor-intensive tasks. For precision robotics to operate effectively, the system must comprehensively understand the farm environment. Advances in computer vision and AI-driven guidance have greatly enhanced the capabilities of these systems, enabling them to address labor-intensive challenges with remarkable efficiency [4]. Moreover, the development of sophisticated AI algorithms is progressively providing solutions to complex agricultural problems, such as automated damage detection [5, 6], non-destructive food and fruit inspection [7-9], and crop disease detection [10, 11].

To comprehensively understand the farm environment, agricultural robots often rely on key visual navigation information processing technologies, including filtering-based, segmentation-based, and line-detection-based data computation algorithms [4]. This paper presents a panoptic segmentation-based image acquisition and analysis framework for environment understanding designed to enhance autonomous robot mapping by integrating precise and detailed context-aware information. This framework leverages panoptic segmentation combined with inter-frame instance tracking that can be used to improve mapping accuracy and functionality. As a result of this framework, an improved digital life cycle of crops in precision agriculture is also proposed, incorporating the panoptic elements introduced by the framework to enhance environmental understanding and operational efficiency.

Among the various panoptic segmentation libraries available, the two most widely used are Detectron2 and You Only Look Once (YOLO) [12-15]. Detectron2 was selected as the primary library for panoptic image analysis in this research due to its fast-processing capabilities and superior performance despite being more complex than YOLO. The panoptic segmentation model, implemented using the Detectron2 algorithm, is trained and enhanced with a customized single-crop dataset, specifically the bird's-eye chili dataset, following a methodology similar to that described by Xu et al. [16]. The proposed framework aims to facilitate precise object and route recognition within the single-crop bird's-eye chili farm environment as an initial step toward developing an enhanced precision agricultural robotics framework, with plans to expand its application to multi-crop environments.

2. Methods

As part of Smart Agriculture's precision robotics framework for understanding farm environments, various approaches utilizing different types of sensors, including Light Detection and Ranging (LiDAR), image, thermal, and hyperspectral sensors, have

been implemented [17]. This research focuses on image sensors due to their cost-effectiveness and widespread adoption as a preferred choice in precision robotics and agriculture. Figure 1 illustrates the digital life cycle of crops in precision agriculture, highlighting the integration of precision robotics as proposed in [18].

In precision robotics image acquisition, collecting data from various image types, including red, green, and blue (RGB), multispectral, thermal, and hyperspectral, is crucial. This data is then analyzed and stitched into a non-overlapping panoramic image database, serving as the virtual reference of the farm. The processed data is then fed into a decision engine, which determines the necessary operations. Finally, operation commands are delivered to the robots, enabling them to navigate and execute tasks on the target plants autonomously.

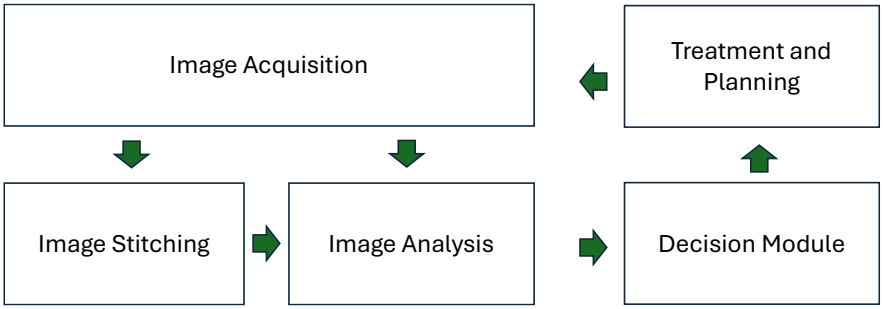

Fig. 1. The digital life cycle of crops in precision agriculture.

The proposed framework focuses on image acquisition and analysis within the entire life cycle. The image acquisition component involves curating a custom dataset of a bird's-eye chili farm, while the image analysis component employs panoptic image analysis to achieve two primary objectives: segmenting all objects in the bird's-eye chili farm environment into distinct classes and unique instances and tracking target instances, specifically bird's-eye chili plants, across consecutive frames. These objectives are critical for achieving precise localization and identification of entire objects within the farm, enhancing the efficiency and reliability of robotic navigation and operations.

The first objective is accomplished through panoptic segmentation on a customized bird's eye chili dataset [19]. In contrast, the second objective is achieved using a tracking algorithm proposed in this paper, which will be discussed later in this section. A summary of the proposed image acquisition and analysis framework is presented in Fig. 2. The first two stages pertain to image acquisition. The initial stage involves data collection, capturing images of objects in and around a bird's-eye chili farm.

In the second stage, these images are filtered according to specific standards to ensure object clarity and suitability for subsequent analysis. Manual labeling and annotation are performed during dataset preparation to create a reference ground truth image for testing. The final two stages focus on panoptic image analysis. The first of these stages involves transfer learning-based panoptic segmentation, where the curated dataset is utilized to train a pre-existing panoptic segmentation model. The next stage involves pixel-wise instance tracking of the chili plants, with the tracking performance evaluated only within the target region.

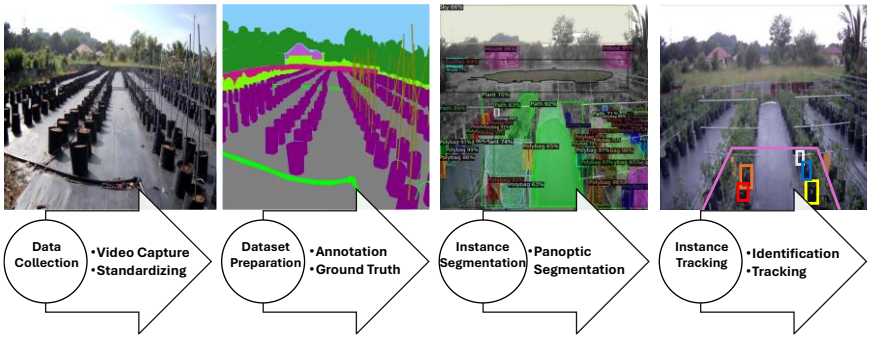

Fig. 2. The image acquisition and analysis framework for instances of segmentation and tracking.

As a result of the proposed image acquisition and analysis framework, an improved digital lifecycle of crops in precision agriculture or robotics is proposed and illustrated in Fig. 3. The key advancement in the proposed digital lifecycle is the integration of three-dimensional context-aware data derived from panoptic image analysis and depth perception images. Depth perception images can be obtained from various sources, including LiDAR, Time-of-Flight (ToF) sensors, and, depth or conventional cameras.

The incorporation of depth perception ensures sufficient information to construct a precise three-dimensional map of the entire environment. With panoptic segmentation, each point in the three-dimensional map is classified based on its category and assigned a unique instance identity (ID). This approach enables the creation of a context-aware map, allowing the robot to identify the object class associated with each point and its unique instance. Such a map enhances the robot's ability to navigate accurately and execute precise operations within the farm environment, significantly improving efficiency and reliability in precision agriculture.

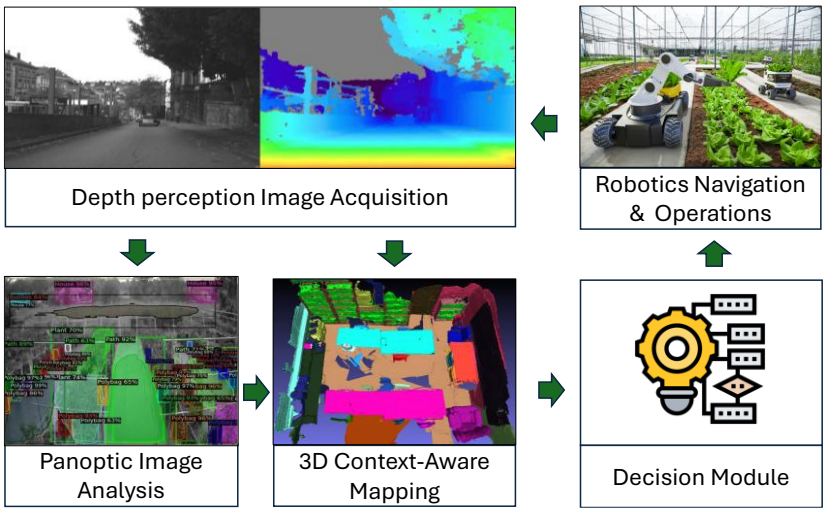

Fig. 3. The proposed digital life cycle of crops in precision agriculture.

2.1. Image acquisition: Bird's-eye chili farm dataset curation

A new dataset specific to the bird's-eye chili farm needs to be curated for the image acquisition process. The data collection begins with capturing ten right-angle and ten left-angle stereo videos comprising fifty frames at the chili fertigation farm. This results in 1,000 frames, which are saved in PNG format. Detailed information regarding the stereo calibration parameters used for this dataset can be found in [19]. Data annotation is crucial in developing the image dataset, providing ground truth information to interpret each class of pixels accurately [20]. Roboflow, a powerful tool that streamlines data labeling, is used to annotate the videos efficiently [21].

The frames are labeled with Roboflow's user-friendly interface and annotation tools by identifying and marking the presence of various objects such as bushes, trees, sky, plants, paths, polybags, and houses, as shown in Fig. 4. The labeled data is used to train the computer vision model as ground truth. A standardized representation is provided by the Common Objects in Context (COCO) format, making it compatible with popular deep-learning frameworks and libraries [22]. The outcome of the image acquisition stage includes continuous frames of images from the bird's-eye chili farm, their corresponding ground truth images, and an info file containing detailed information about those images.

Fig. 4. Sample of the image frame after labeling in Roboflow.

2.2. Panoptic image analysis

The panoptic image analysis involves two primary objectives: first, to perform panoptic segmentation, assigning each pixel in the frame to a specific class and instance, and second, to track the target instances across consecutive frames. The detailed architecture for the panoptic image analysis is illustrated in Fig. 5. For panoptic segmentation, transfer learning is implemented using the Bird's Eye Chili dataset and the Detectron2 segmentation library, leveraging its robust backbone structure built on the Feature Pyramid Network (FPN) [13, 23, 24]. Meanwhile, target tracking is implemented using a simple yet effective approach, which will be detailed later in this section.

Detectron2 provides three pre-trained models for model training, which differ based on their backbone architecture (ResNet-50 or ResNet-101) and training duration (1x or 3x). Adjustments, augmentation, and optimization can be made to the ROI batch size per image, the number of workers, brightness, and exposure, and the test threshold to balance memory usage, training speed, and precision-recall

trade-off [23, 25]. Among the available models, COCO_rcnn_R_50_FPN_1x is experimentally selected and optimized for transfer learning using the bird's-eye chili dataset.

Target tracking aims to accurately assign a unique ID to each target across continuous frames. As illustrated in Fig. 5, the first step is to define a target region for tracking, as the robot's operation is confined to its view angle. Based on the robot's view, a two-by-three-meter square area was selected as the target region. Once the target region is established, target instance tracking is performed on all panoptically segmented instances. However, the panoptic segmentation results do not consistently assign the same instance ID to each detected target across different frames. This inconsistency results in varying IDs for the same target across frames. To address this problem, a multi-frame target depth searching tracking algorithm is employed to ensure that the instance IDs remain consistent and accurately correspond to the number of detected chili plant instances within the target region.

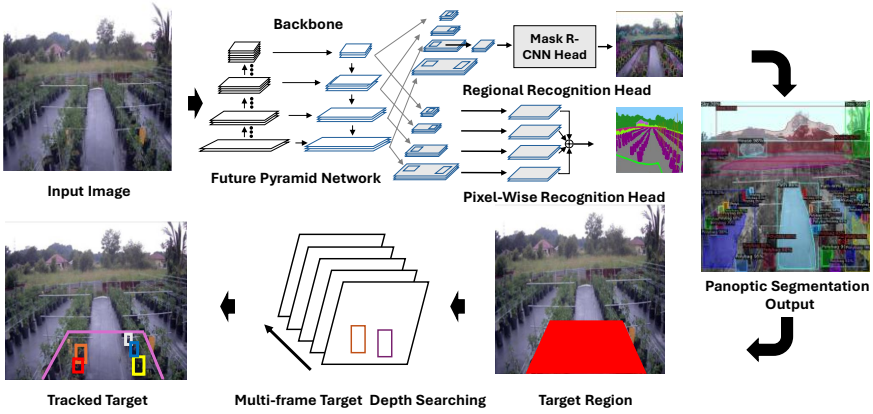

Fig. 5. The transfer learning framework design for the panoptic segmentation in the chili farm.

First, the top-left depth of each polybag instance is recorded based on the depth calculations obtained from the stereo camera calibration data, as detailed in [19]. The top-left point is chosen as the reference because, as the robot moves forward, the upper part of the plants remains visible, enabling more accurate instance tracking. This process is repeated for each subsequent frame to capture the top-left depth of each instance. Next, the instance depths in the current frame are compared with those from the previous frames. The logic outlined in the pseudocode in Fig. 6 is then executed to determine whether an instance in the current frame corresponds to a previous instance.

```
1  Get all instance's top left depth in previous frame ROI
2  Get all instance's top left depth in current frame ROI
3  FOR each instance's top left depth in current frame x:
4     FOR each instance's top left depth in previous frame y:
5        IF IFID(x , y) < IFTD:
6           y instance ID = x instance ID
```

Fig. 6. The pseudocode for the multi-frame target depth searching algorithm.

To accurately determine the true instance distance between subsequent frames, it is essential to calculate the distance of each pixel from the camera, as shown in Fig. 7. By employing the stereo camera calibration values as demonstrated by Saipullah et al. [19], the pixel-to-camera distance can be approximately calculated using the following formula [26]:

$$dZ_c = \frac{Z^2}{fb} dp_x \tag{1}$$

where dZ_c is depth resolution at Z distance, f is focal length, b is the baseline, and dp_x is refer to disparity accuracy.

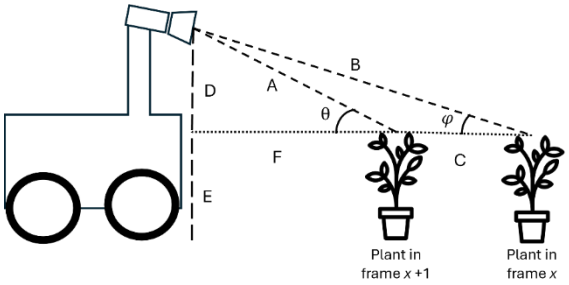

Fig. 7. The diagram for target distance calculation.

The inter-frame travel distance (IFTD) between two consecutive frames must be estimated based on the robot's speed and the frame rate (FR) to assess whether a detected instance corresponds to the same instance from the previous frame, as shown in the following formula [27].

$$\text{IFTD} = \frac{\text{S}_\text{R}}{\text{FR}} + \alpha \tag{2}$$

where s_R is the robot's speed, and α is the distance between two plants. With the robot's speed of 0.5 ms^{-1}, and FR as 1 *fps*, and α as 0.5 *m,* the IFTD can be determined as 1m based on formula (2). The inter-frame instance distance (IFID) can be calculated using the trigonometry concept as shown in Fig. 6 and formalized as follows:

$$IFID = \text{B}\cos(\varphi) - \text{A}\cos(\theta) \tag{3}$$

$$\varphi = \sin^{-1}\left(\frac{F+C}{D}\right) \tag{4}$$

$$\theta = \sin^{-1}\left(\frac{F}{D}\right) \tag{5}$$

Given that the camera height *D* is fixed, the value of *A*, representing the top-left depth or distance from the camera, can be calculated for each frame as the robot moves. Consequently, the IFID can be determined for every frame.

3. The Result and Discussions

The dataset was randomly divided into two equal groups for the experiment: a training group and a testing group. The training group was utilized during the pre-trained model selection process. Only the model generated from the selected pre-trained model will be employed for subsequent experiments.

3.1. Pre-trained model selection

For pre-trained model selection, each model-COCO_rcnn_R_50_FPN_1x, COCO_rcnn_R_50_FPN_3x, and COCO_rcnn_R_101_FPN_3x-was trained on the same training group dataset to ensure a consistent baseline for evaluation. The training was conducted for 3,000 and 5,000 iterations to evaluate performance across different iteration counts. Following training, the models were rigorously validated using a range of evaluation metrics, including Average Precision (AP) with 50% Intersection over Union (IoU) or AP50, precision (P), recall (R), and F1-score (F1), with the highest-scoring model selected.

These metrics offered critical insights into the model's performance under real-world conditions and its generalization ability to previously unseen data [28]. This analysis encompassed the visualization of the model's predictions for each frame, which were then compared with the corresponding ground truth annotations [29]. The selected evaluation metrics are related to the results' confusion matrix, as is explained in Table 1.

These metrics evaluate model performance in accuracy, object categorization, and optimizing true positives while reducing false positives. The detailed formula for precision-recall and F1-score are as follows:

$$p = \frac{TP}{TP + FP} \tag{6}$$

$$r = \frac{TP}{TP + FN} \tag{7}$$

$$F1 = \frac{2 \cdot p \cdot r}{p + r} \tag{8}$$

Table 1. The confusion matrix for true and false positive and negative.

		Actual	
		Positive	**Negative**
Predicted	**Positive**	True Positive (TP)	False Positive (FP)
	Negative	False Negative (FN)	True Negative (TN)

As illustrated in Fig. 8, superior performance in all AP50, P, R, and F1 was demonstrated by the COCO_rcnn_R_50_FPN_1x model compared to the other pre-defined models. The COCO_rcnn_R_50_FPN_1x model demonstrated superior capability in transfer learning with the bird's-eye chili dataset, effectively detecting and classifying objects in the farm environment with high accuracy.

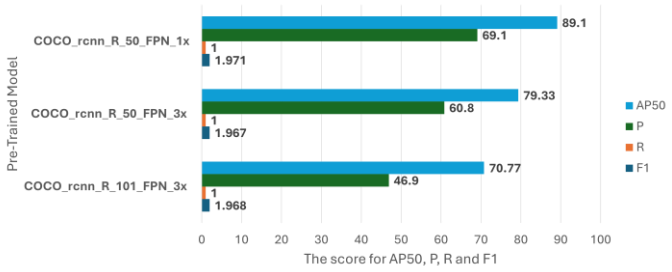

Fig. 8. Pre-train model comparison on 3000 iteration.

An additional 2,000 iterations were applied to the selected model, resulting in 5,000. This extended training period enabled the model to discern complex patterns and enhance its accuracy in detecting and classifying objects within the provided dataset. Figure 9 presents the performance metrics, offering a clear and concise overview of the model's performance across various evaluation criteria with different iterations.

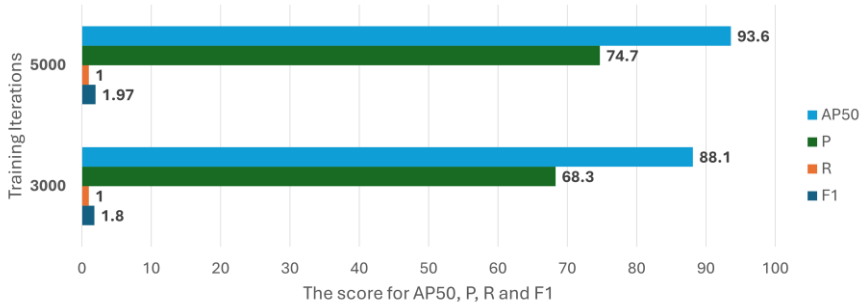

Fig. 9. Comparative results between 3000 and 5000 iterations.

The results indicate that increasing the iterations to 5,000 improves panoptic segmentation performance by nearly 5% compared to lower iteration counts. As shown in Fig. 10, detailed object class performance indicates that iteration impacts each class differently. Low-accuracy classes improved by over 10%, while high-accuracy classes experienced gains of up to 3%. Based on its superior segmentation accuracy, the COCO_rcnn_R_50_FPN_1x pre-trained model was selected for further experiments and extended into the bird's-eye chili model through transfer learning using the Detectron2 panoptic segmentation algorithm, trained with 5,000 iterations.

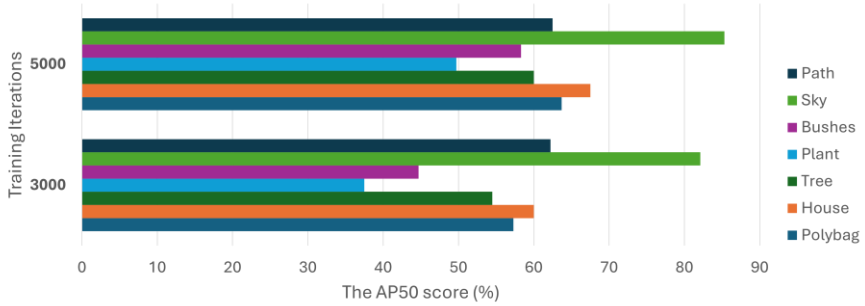

Fig. 10. Comparison of average precision per categories.

3.2. Panoptic segmentation test

To evaluate the panoptic segmentation performance, the AP50 metric of the proposed framework is assessed frame by frame on a separate portion of the dataset, the testing group, which was excluded from the model training process. This detailed frame-level examination has acquired valuable insights into the model's performance, enabling informed decisions regarding its application and potential areas for further enhancement [30]. By evaluating the detection results across

various frames, insights into the model's strengths and limitations can be obtained [31]. This experiment compares the proposed panoptic segmentation (PPS) framework based on Detectron2 to the YOLO-based panoptic segmentation implementation described by Bolya et al. [15].

The results of the panoptic segmentation on the testing group are presented in Table 2. Sample segmentation outcomes on selected frames are illustrated in Fig. 11. Overall, the output of PPS on the video frames demonstrates the ability of the Detectron2 model to detect instances, including polybags, paths, plants, trees, bushes, sky, and houses, achieving an overall accuracy of 77.4%.

In contrast, YOLO detects fewer instances, resulting in a lower overall accuracy of 68.7%. The highest segmentation accuracies achieved by PPS are for the sky (98%), polybag (80%), and path (81%). YOLO, in comparison, performs about 10% lower on these key classes. Although both methods achieved high accuracy in sky detection due to the limited number of instances of that object, accurate segmentation of polybags and paths is crucial for robot operation and navigation. In this regard, PPS outperforms YOLO because of its advanced Detectron2-based framework.

Fig. 11. Panoptic segmentation results: PPS (left) and YOLO (right).

Table 1. The class-based panoptic segmentation AF50 results.

Algorithm	polybag	house	tree	plant	bushes	sky	path
PPS	80%	85%	73%	60%	65%	98%	81%
YOLO	67%	83%	66%	50%	57%	88%	70%

3.3. Instance Tracking performance on a continuous frame's video

Instance tracking in panoptic videos involves identifying and monitoring individual objects or instances across successive frames. By analyzing the variations in these masks between frames, the movement and transformations of individual objects over time can be tracked [32, 33]. The target instance for tracking is identified as the polybag class, serving as a container for chili plants. Future implementations will require unique identifiers for each polybag, emphasizing the need for high-tracking performance. For comparison, PPS and YOLO tracking performance were

evaluated using the Average Tracking Accuracy (ATA) metric on two 100-frame videos with the same tracking algorithm.

The ATA is calculated by averaging the Tracking Accuracy (TA) of all instances across each frame using the following formula:

$$ATA = \frac{1}{F}\sum_{f=1}^{F} TA(f) \tag{9}$$

$$TA(c) = \frac{1}{I_c}\sum_{i=1}^{I_c} P(i) \tag{10}$$

$$P(x) = \begin{cases} 1, & x = ID \\ 0, & x \neq ID \end{cases} \tag{11}$$

where F is the number of frames evaluated, I_c is the number of IDs of each instance in frame c, and P is the mark given if the instance is correctly identified in that particular frame. After applying the TA to each frame, the average tracking accuracy (ATA) was calculated. Figure 12 illustrates sample frames of target tracking results for PPS and YOLO. Each detected polybag instance is assigned a unique ID, highlighted with distinct bounding box colors. PPS accurately tracked all five instances of the same polybag throughout the video sequence. In comparison, YOLO detected only three instances due to its poorer detection performance during panoptic segmentation. Detailed results are presented in Fig. 13, showing that PPS achieved excellent tracking performance with an ATA of 94.9%, while YOLO demonstrated 10% lower performance with an ATA of 73.4%.

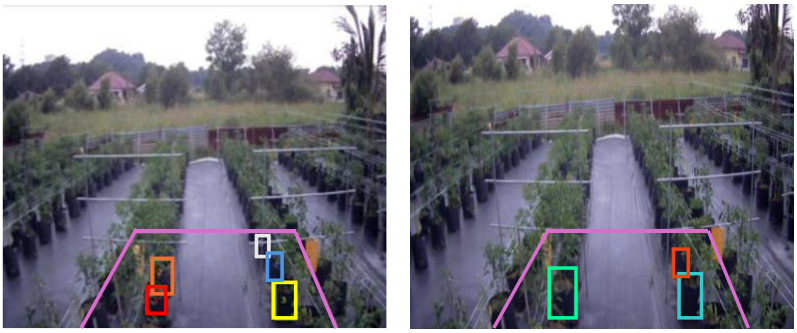

Fig. 12. Target tracking sample frame: PPS vs. YOLO.

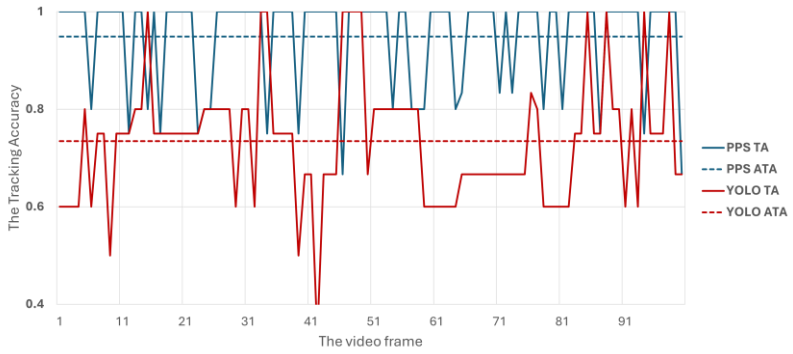

Fig. 13. Frame-by-frame target tracking performance for PPS and YOLO.

4. Conclusion

This study presents a panoptic segmentation-based image acquisition and analysis framework specifically designed for a single crop, the bird's-eye chili. Both objectives of the panoptic image analysis-the panoptic segmentation and target instance tracking-were successfully achieved. The proposed framework, built on the Detectron2 platform using the experimentally selected pre-trained COCO_rcnn_R_50_FPN_1x model, was compared to a YOLO-based panoptic image analysis. The results demonstrate that the proposed framework significantly outperforms YOLO, achieving over 10% higher performance in panoptic segmentation and target tracking.

Additionally, it introduces an improved digital life cycle of crops in precision agriculture, optimized to integrate the proposed image acquisition and analysis framework that relies on depth and three-dimensional data outputs. Future work will focus on completing the remaining components of this enhanced digital life cycle, with detailed exploration of depth image acquisition and three-dimensional data mapping. To extend the digital life cycle to other crops, developing a multi-crop dataset will be a key focus in future work after completing the full digital life cycle for the bird's-eye chili crop. This will involve creating a context-aware three-dimensional map as the primary reference for the farm environment and implementing at least one autonomous robotic operation, such as a chili-picking task, that fully integrates and utilizes all elements of the proposed digital life cycle.

Integrating the proposed digital life cycle as a precision agriculture framework with the broader Smart Agriculture framework will position this work as a significant contributor to advancing Smart Agriculture. This integration will enhance the automation of labor-intensive tasks through AI and robotics, promoting greater efficiency and sustainability in agricultural operations. Upon completing the full life cycle for bird's-eye chili, future work will extend the implementation of this framework to other local crops, such as pineapple and palm trees, similar to existing applications with tomato [34] and grape crops [35].

Acknowledgment

This research was supported by Ministry of Higher Education (MoHE) through Fundamental Research Grant Scheme (FRGS/1/2021/ICT02/UTEM/02/2). The authors would like to thank the Machine Learning & Signal Processing (MLSP) research group of the Faculty of Technology and Engineering Electronics and Computer (FTKEK), Universiti Teknikal Malaysia Melaka (UTeM), for the usage of existing facilities to complete this project. The authors would also like to thank the Solok Fertigasi by MSJ Perwira Enterprise for their collaboration.

References

1. Hemming, S.; de Zwart, F.; Elings, A.; Righini, I.; and Petropoulou, A. (2019). Remote control of greenhouse vegetable production with artificial intelligence-greenhouse climate, irrigation, and crop production. *Sensors*, 19(8), 1807.
2. Banerjee, G.; Sarkar, U.; Das, S.; and Ghosh, I. (2018). Artificial intelligence in agriculture: A literature survey. *Scientific Research in Computer Science Applications and Management Studies*, 7(3).

3. Rehman, A.U.; Alamoudi, Y.; Khalid, H.M.; Morchid, A.; Muyeen, S.M.; and Abdelaziz, A.Y. (2024). Smart agriculture technology: An integrated framework of renewable energy resources, IoT-based energy management, and precision robotics. *Cleaner Energy Systems*, 9, 100132.
4. Bai, Y.; Zhang, B.; Xu, N.; Zhou, J.; Shi, J.; and Diao, Z. (2023). Vision-based navigation and guidance for agricultural autonomous vehicles and robots: A review. *Computers and Electronics in Agriculture*, 205, 107584.
5. Chaiyasarn, K.; and Buatik, A. (2021). Tile damage detection in temple facade via convolutional neural networks. *Journal of Engineering Science and Technology*, 16(4), 3057-3071.
6. Soltani Firouz, M.; and Sardari, H. (2022). Defect detection in fruit and vegetables by using machine vision systems and image processing. *Food Engineering Reviews*, 14(3), 353-379.
7. Pham, V.; Pham, C.; and Dang, T. (2020). Road damage detection and classification with Detectron2 and Faster R-CNN. *Proceedings of the* 2020 *IEEE International Conference on Big Data* (*Big Data*), Atlanta, GA, USA, 5592-5601.
8. Mahanti, N.K.; Pandiselvam, R.; Kothakota, A.; Ishwarya, S.P.; Chakraborty, S.K.; Kumar, M.; and Cozzolino, D. (2022). Emerging non-destructive imaging techniques for fruit damage detection: Image processing and analysis. *Trends in Food Science & Technology*, 120, 418-438.
9. Tripathi, M.K.; and Maktedar, D.D. (2020). A role of computer vision in fruits and vegetables among various horticulture products of agriculture fields: A survey. *Information Processing in Agriculture*, 7(2), 183-203.
10. Jaware, T.H.; Badgujar, R.D.; and Patil, P.G. (2012). Crop disease detection using image segmentation. *World Journal of Science and Technology*, 2(4):190-194.
11. Khakimov, A.; Salakhutdinov, I.; Omolikov, A.; and Utaganov, S. (2022). Traditional and current-prospective methods of agricultural plant diseases detection: A review. *IOP Conference Series*: *Earth and Environmental Science*, 951(1), 012002.
12. de Almeida, G.P.S.; dos Santos, L.N.S.; da Silva Souza, L.R.; da Costa Gontijo, P.; de Oliveira, R.; Teixeira, M.C.; De Oliveira, M.; Teixeira, M.B.; and do Carmo França, H.F. (2024). Performance analysis of YOLO and Detectron2 models for detecting corn and soybean pests employing customized dataset. *Agronomy*, 14(10), 2194.
13. Cao, L.; Li, Y.; Zhang, X.; Ouyang, Z.; Li, Z.; Wang, Y.; Guo, Q.; Han, L.; and Zhang, D. (2024). Comparison of two deep learning model YOLOF and Detectron2 for mesoscale eddies identification in the South China Sea. *International Journal of Remote Sensing*, 45(19-20), 6919-6933.
14. Karthikeya Nalam, V.S.; Amar Koushik Tanniru, V.S.; Posani, A.; and Suneetha, M. (2022). Detection and recognition of drones using deep convolution neural networks. *Proceedings of the* 2022 *IEEE* 6th *Conference on Information and Communication Technology* (*CICT*), Gwalior, India.
15. Bolya, D.; Zhou, C.; Xiao, F.; and Lee, Y.J. (2019). YOLACT: Real-time instance segmentation. *Proceedings of the* 2019 *IEEE/CVF International Conference on Computer Vision* (*ICCV*), Seoul, Korea (South).

16. Xu, Y.-S.; Fu, T.-J.; Yang, H.-K.; and Lee, C.-Y. (2018). Dynamic video segmentation network. *Proceedings of the* 2018 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, 6556-6565.
17. Botta, A.; Cavallone, P.; Baglieri, L.; Colucci, G.; Tagliavini, L.; and Quaglia, G. (2022). A review of robots, perception, and tasks in precision agriculture. *Applied Mechanics*, 3(3), 830-854.
18. Ghazal, S.; Munir, A.; and Qureshi, W.S. (2024). Computer vision in smart agriculture and precision farming: Techniques and applications. *Artificial Intelligence in Agriculture*, 13, 64-83.
19. Saipullah, K.M.; Saad, W.H.M.; Wong, Q.L.; Husni, M.S.M.; Idris, M.I.; and Razak, M.S.J.A. (2023). Dataset of bird's eye chilies farm for stereo image semantic segmentation. *Data in Brief*, 51, 109714.
20. Simpson, A.L. et al. (2019). A large annotated medical image dataset for the development and evaluation of segmentation algorithms. *arXiv*:1902.09063*v*1 [*cs.CV*].
21. Ciaglia, F.; Zuppichini, F.S.; Guerrie, P.; McQuade, M.; and Solawetz, J. (2022). Roboflow 100: A rich, multi-domain object detection benchmark. *arXiv*:2211.13523*v*3 [*cs.CV*].
22. Rostianingsih, S.; Setiawan, A.; and Halim, C.I. (2020). COCO (Creating common object in context) dataset for chemistry apparatus. *Procedia Computer Science*, 171, 2445-2452.
23. Ter-Sarkisov, A. (2021). Lightweight model for the prediction of Covid-19 through the detection and segmentation of lesions in chest CT scans. *International Journal of Automation*, *Artificial Intelligence and Machine Learning*, 2(1), 1-15.
24. Silva, M.O. et al. (2022). Action recognition of industrial workers using Detectron2 and AutoML algorithms. *Proceedings of the* 2022 *IEEE International Conference on Consumer Electronics-Taiwan*, Taipei, Taiwan.
25. Perez, L.; and Wang, J. (2017). The effectiveness of data augmentation in image classification using deep learning. *arXiv*.
26. Kytö, M.; Nuutinen, M.; and Oittinen, P. (2011). Method for measuring stereo camera depth accuracy based on stereoscopic vision. *Proceedings of the IS&T/SPIE Electronic Imaging*, San Francisco Airport, California, USA.
27. Doğan, S.; Temiz, M.S.; and Külür, S. (2010). Real time speed estimation of moving vehicles from side view images from an uncalibrated video camera. *Sensors*, 10(5), 4805-4824.
28. Ding, J.; Zhang, J.; Zhan, Z.; Tang, X.; and Wang, X. (2022). A precision-efficient method for collapsed building detection in post-earthquake UAV images based on the improved NMS algorithm and faster R-CNN. *Remote Sensing*, 14(3), 663.
29. Liu, Z.; Zhu, X.; Hu, G.; Guo, H.; Tang, M.; Lei, Z.; Robertson, N.M.; and Wang, J. (2019). Semantic alignment: Finding semantically consistent ground-truth for facial landmark detection. *Proceedings of the* 2019 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Long Beach, CA, USA.
30. Oh, S.W.; Lee, J.-Y.; Xu, N.; and Kim, S.J. (2019). Video object segmentation using space-time memory networks. *Proceedings of the* 17th *IEEE/CVF*

International Conference on Computer Vision, *ICCV* 2019, Korea, Seoul, 9225-9234.

31. Hemke, R.; Buckless, C.G.; Tsao, A.; Wang, B.; and Torriani, M. (2020). Deep learning for automated segmentation of pelvic muscles, fat, and bone from CT studies for body composition assessment. *Skeletal Radiology*, 49(3) 387-395.
32. Bertasius, G.; and Torresani, L. (2020). Classifying, segmenting, and tracking object instances in video with mask propagation. *Proceedings of the* 2020 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Seattle, WA, USA, 9736-9745.
33. Azimjonov, J.; and Özmen, A. (2021). A real-time vehicle detection and a novel vehicle tracking systems for estimating and monitoring traffic flow on highways. *Advanced Engineering Informatics*, 50, 101393.
34. Lee, J.; Nazki, H.; Baek, J.; Hong, Y.; and Lee, M. (2020). Artificial intelligence approach for tomato detection and mass estimation in precision agriculture. *Sustainability*, 12(21), 9138.
35. Oberti, R. et al. (2016). Selective spraying of grapevines for disease control using a modular agricultural robot. *Biosystems Engineering*, 146, 203-215.