Generative Monocular Perception Pipeline-Based Framework for Accurate Stem Detection in Automated Strawberry Harvest

Kohei Arai[1], Jin Sawada[2], Mariko Oda[3]
Department of Information Science, Saga University, Saga City, Japan[1]
Graduate School, Kurume Institute Technology, Kurume City, Japan[2]
Applied AI Laboratory, Kurume Institute of Technology, Kurume City, Japan[3]

***Abstract*—Horticulture faces a growing labor crisis, driving demand for autonomous harvesting robots, but reliable strawberry peduncle detection remains a critical unsolved challenge due to their fine, millimeter-scale structure and severe intertwining with leaves and stems. Existing single-view RGB imaging struggles with occlusions and ambiguities, while depth sensors falter in reflective greenhouse environments plagued by noise and data gaps. Introducing a generative monocular perception pipeline—the first to reconstruct multi-view cues purely from a single RGB image—this study achieves perceptual consistency through four novel, synergistic innovations: (i) pseudo multi-view synthesis to emulate diverse viewpoints, (ii) monocular depth estimation for precise geometric guidance and background isolation, (iii) line-curve geometric modeling to capture subtle peduncle features, and (iv) occlusion-order reasoning via cross-view consistency analysis. In comparative trials against a YOLO-based detector (85.71% region accuracy vs. 57.14%), our pipeline delivers orientation precision, slashing mean angular error from 18.31° to 13.96°—robust, clutter-resilient cutting cues for next-generation robotic harvesters. Evaluated on farm images, it reduces mean angular error to 13.96° (SD 10.15°) from YOLO's 18.31° (SD 11.27°), with $p<0.05$ (paired t-test, n=14).**

***Keywords*—Monocular depth estimation; marigold; easy wan22; strawberry harvesting robot; stem detection**

I. Introduction

The primary objective of this study is to identify the most effective Generative AI approach for detecting thin stems. We hypothesize that the distinct generation mechanisms, specifically mesh reconstruction versus pixel-level view synthesis, significantly affect the preservation of fine shape details. Accordingly, this paper presents a comparative evaluation of 3D model generation AIs such as CSM and Tripo against state-of-the-art video generation AIs like Easy Wan 22 and Dream Machine. Based on experimental results demonstrating the superiority of video generation in retaining fine structures, we designed a stem detection framework that integrates the selected video generation AI with monocular depth estimation. We demonstrated the feasibility of the proposed method through experiments using real-world environment data provided by INAC System Co., Ltd. [1].

We propose a novel detection framework that utilizes generative AI to recover 3D spatial information from monocular images. We developed a pipeline that integrates the locally operable video generation AI "Easy Wan 22" and the monocular depth estimation AI "Marigold." While 3D model generation AI tends to smooth or lose fine details of thin shapes during mesh reconstruction, video generation AI focuses on preserving fine structures by synthesizing high-resolution multi-view frames, successfully revealing the fine structures of thin shapes. As a key application of this technology, accurate stem detection is a key challenge in automated strawberry harvesting robots. Conventional vision systems using YOLO-based object detection or general-purpose depth cameras often fail to detect thin stalks or distinguish between overlapping stems due to insufficient resolution and occlusion. To address these limitations, this study compared two approaches: 3D model generation with mesh reconstruction and video generation with view synthesis, to determine a suitable generation method for preserving fine structures such as plant stems.

The following section describes research background followed by related works. Then, the procedure, the results, and some discussions of the comparative study are described together with the proposed method. After that, experiments procedure is described with results, followed by a conclusion is described with some discussions and future works.

II. Research Background

A. Background and Challenges

Smart agriculture is rapidly advancing to address global labor shortages in the agricultural sector. The development of automated harvesting robots is particularly urgent for strawberries, a highly labor-intensive crop. These robots require the capability not only to detect fruit but also to accurately identify the cutting position of the peduncle stem. Conventional robotic vision systems typically combine object detection algorithms, such as YOLO, with physical depth sensors like LiDAR or stereo cameras. While these systems are highly effective for detecting large, distinct objects like fruit, they encounter significant difficulties in detecting "thin stems." As stems are only a few millimeters in diameter, general depth cameras often suffer from insufficient spatial resolution, resulting in "depth information loss" (Fig. 1).

Furthermore, relying solely on 2D image recognition makes it difficult to identify the correct cutting target when stems overlap or are occluded by leaves. This is our problem statement: because it is not possible to know the front and rear

of the strawberry stems, it is not possible to know which stem to cut to harvest ripe strawberries.

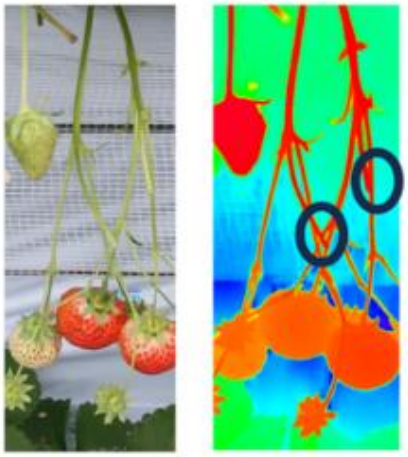

Fig. 1. Example of depth estimation failure. (a) Original image (Left). (b) Stem detected image (Right). The thin stem (Circled) is assimilated into the background due to its fine geometry.

B. Potential of Generative AI

To overcome the limitations of physical sensors, approaches utilizing AI to recover 3D information from standard monocular images are gaining attention. In human vision, depth and 3D structure are perceived not only through binocular parallax but also through "Motion Parallax," where an object is observed from slightly different angles. Recently, Generative AI has evolved to mimic this capability, enabling the generation of 3D models or multi-view videos from a single 2D input. However, in the context of agricultural robotics, it remains unclear which generation mechanism is better suited for preserving "fine structures" like plant stems.

III. Related Works

This section organizes prior research streams relevant to the problem setting addressed in this study: recognizing fine, wire-like structures such as peduncles in complex cultivation environments with occlusion and crossings, and estimating their three-dimensional pose as well as depth ordering. As described in the background, while automating harvesting operations is becoming increasingly important in response to labor shortages, real agricultural sites often exhibit unstable visual recognition compared with laboratory settings due to factors such as illumination changes, complex backgrounds, object-level variability, and frequent occlusions. The aim of this study is to obtain pseudo multi-view information from a single viewpoint image, thereby strengthening the understanding of depth ordering at crossings and the three-dimensional structure of fine targets---cases where conventional monocular recognition tends to struggle. In what follows, we organize the discussion from three perspectives: 1) general frameworks for visual perception and depth estimation in agricultural robots, 2) the difficulty of pose and depth-order estimation under occlusion, and 3) the potential of generative AI for recovering three-dimensional information.

A. Visual Perception and Depth Estimation in Agricultural Robots

In harvesting robots, a common configuration is to combine camera-based object detection/segmentation with three-dimensional measurement using range sensors. On the vision side, deep-learning-based detectors are used to identify regions corresponding to fruits, leaves, flowers, and peduncles, and to obtain candidates for reach points and grasp points for robot arms. On the ranging side, stereo cameras, active depth cameras, LiDAR, and related devices are used to acquire depth and compute three-dimensional positions required for tasks such as grasping and cutting. This framework is effective for targets such as fruits that occupy many pixels and have relatively smooth surfaces, and verification under conditions close to practical operation has accumulated [2-5].

However, when the target is a fine, wire-like structure (fine structures) with a diameter of only a few millimeters, such as a peduncle, inherent limitations tend to appear in both image recognition and range sensing. In image recognition, the fine parts occupy very few pixels and compete with background edges, leaf veins, reflections, and noise, so detection results can become fragmented or exhibit unnatural thickness fluctuations. In depth measurement, limited spatial resolution can cause the fine structure to be averaged out at sub-pixel scale, making depth unstable and resulting in missing values (holes) or outliers. In particular, in environments where strong greenhouse reflections, thin leaf edges, highlights on fruit surfaces, and background meshes or poles overlap, depth-sensor errors can be amplified. Thus, to obtain accurate grasping/cutting points on peduncles, simply ``detect the fruit and measure distance" is insufficient; additional mechanisms are needed to stably reconstruct the fine structure as a continuous shape [6].

From the standpoint of robot control, not only estimation accuracy but also repeatability and latency are crucial. In agricultural sites, it is common to perform recognition and tracking multiple times for the same fruit cluster, and if estimates fluctuate frame by frame, trajectory generation for grasping and cutting becomes unstable. Moreover, if depth estimation or three-dimensional reconstruction requires high computational cost, processing delays can prevent the system from following motions such as fruit and leaf sway (due to wind or contact), which reduces practical success rates. Consequently, prior work has emphasized design choices regarding which information to use at which stage (detection → tracking → pose estimation → grasp planning) under the trade-offs among accuracy, computational cost, and stability.

B. Difficulty of Pose Estimation under Occlusion

For fine targets such as peduncles, what is fundamentally required for harvesting is not only detecting their presence, but also obtaining geometric information: the running direction (pose) and the spatial location of candidate cutting points. In dense clusters where multiple fruits are packed together, peduncles are not only occluded by leaves and other fruits, but also cross and overlap; as a result, there can be segments that appear as a single line in the image.

At such crossings, relying only on the continuity of two-dimensional line segments can lead to incorrectly connecting different peduncles. This can in turn cause cutting the wrong peduncle (mis-cutting) or mixing up immature and overripe fruits.

The occlusion-and-crossing problem can be understood as missing information inherent in a monocular viewpoint. With a single view, there is often insufficient direct evidence to determine which of two peduncles that appear to intersect is in front and which is behind (the depth ordering of occlusion). If one depends on weak cues such as texture, shading, blur, or focus, discrimination can easily break down due to lighting conditions or cultivar differences. In response, prior work has investigated compensating for three-dimensional consistency by using multi-view observations (multi-camera setups), active viewpoint changes (moving the camera or the arm), or acquiring depth using range sensors [7].

In practical operation, however, these approaches face additional constraints. Multi-camera systems impose constraints on calibration and placement flexibility, and viewpoint arrangements are limited in narrow greenhouse aisles or shelf-based cultivation. Active viewpoint changes increase mechanical complexity and operation time, conflicting with throughput constraints. Depth sensors are prone to missing values due to surface reflectance properties and occlusions, and the impact of missing data becomes more pronounced for finer structures. Therefore, stably estimating peduncle pose and depth ordering under occlusion and crossings is positioned as a challenging problem in which sensor configurations, algorithms, and operating conditions interact in complex ways [8-10].

C. Recovering Three-Dimensional Information with Generative AI

In recent years, techniques for generating image sequences (videos) or three-dimensional representations from a single image have developed rapidly. A key feature of these methods is that they can achieve pseudo multi-view observation (novel view synthesis) by generating how the scene would look under viewpoint changes, even without explicit multi-camera setups or depth sensors [11,12]. Similar to how human vision uses motion parallax for depth perception, viewpoint changes produced by generative models may help estimate depth and depth ordering even from monocular input.

In the context of agricultural robots, however, the central issue is not the visual appeal" of generated outputs but whether the geometry of fine structures can be preserved without breakdown. In general, image/video generation models can have smoothing or regularization effects to produce visually natural outputs, and very fine structures may be suppressed as noise. For targets like peduncles that appear only a few pixels wide, this can manifest as disappearance of thin lines," blending into the background," or unnatural thickness variations," making the outputs unsuitable as inputs for downstream line extraction or pose estimation. In addition, generative models may lack sufficient frame-to-frame (temporal) consistency; if the peduncle shape fluctuates across frames, correspondence and occlusion analysis become unstable [13].

Another discussion point is that strengths and weaknesses vary across model families. While diffusion-based models can generate high-quality images, the extent to which thin-line preservation and temporal consistency can be ensured is condition-dependent. Methods that use explicit three-dimensional representations such as NeRF or 3D Gaussian Splatting may be more likely to maintain geometric consistency under viewpoint changes, but there is concern that fine, wire-like geometry may disappear due to insufficient density or regularization during reconstruction [14, 15]. When considering integration into agricultural robots, it is necessary to evaluate how generated outputs affect downstream processing such as depth estimation, line extraction, and depth-order discrimination, including computational cost and repeatability.

Overall, pseudo multi-view generation using generative AI is a promising direction for enhancing peduncle understanding under occlusions and crossings, while the bottlenecks are the degree to which fine-structure preservation and temporal consistency can be satisfied. The distinctive aspect of this study is to examine a framework that stabilizes depth-order estimation at crossings and identification of cutting candidates by combining viewpoint changes obtained by generative AI with monocular depth estimation and line extraction. In the following sections, we clarify the selection of generation methods and evaluation criteria for fine-structure targets, and connect them to the design of the proposed pipeline.

Against this background, this study applies generative AI capable of generating multi-view representations from monocular images to agricultural visual recognition. A method is developed to estimate the three-dimensional structure and spatial relationships of fruit stalks, which are difficult to capture using conventional depth sensors, and its effectiveness is experimentally validated.

Specifically, the proposed method leverages the ability of generative AI to emulate motion parallax in order to accurately identify the precise cutting position of target fruit stalks, even in complex environments involving occlusions and intersections. Furthermore, a comparative evaluation of multiple generation mechanisms is conducted to determine the most suitable approach for preserving fine geometric structures, such as those of strawberry fruit stalks.

IV. Comparative Study

In this section, we conducted a systematic comparative experiment between two rapidly advancing generative AI approaches, namely mesh-based 3D model generation and video-based generation, to determine the optimal method for obtaining three-dimensional information of peduncles from monocular images.

A. Experimental Setup

Dataset: We utilized real images from a strawberry farm, with a resolution of 1920×1080 pixels in RGB format, provided by INAC System Co., Ltd. [1]. These images include not only fruit at harvest maturity but also occluded regions where leaves and other fruits overlap, alongside thin peduncles with diameters under 2 mm.

Evaluation Metrics: We performed a qualitative evaluation of the generated three-dimensional information, including meshes and novel-view images, according to the following three criteria:

- Geometric Fidelity: Whether thin peduncles are generated continuously without interruptions.
- Texture Quality: Whether edge sharpness is preserved to the degree that line-segment detection algorithms, such as Canny or Hough transforms, can reliably detect them.
- Practicality: Generation speed and licensing terms, with consideration for integration into robotic systems.

B. Validation of 3D Model Generation AIs

This approach generates 3D meshes and textures directly from input images. In this study, based on the provided manuscript and current technological trends, we evaluated the following eight AI tools.

1) Tripo (FLUX.1 context) [16]: Technical characteristics: this tool is built on the latest flux.1 context model, offering exceptionally fast generation speed and a fully developed api. Its ease of integration with external systems makes it highly compatible with robotic applications.

Evaluation Results: Geometric fidelity has improved compared to earlier versions, and the tool demonstrates high practicality in controlled environments. However, in complex farm images, the generation of thin peduncles still exhibited issues such as planar artifacts and mesh discontinuities. Complete geometric reconstruction was not achieved, indicating remaining room for improvement.

2) Meshy [17]: Technical Characteristics: Meshy separates geometry generation and texture generation into distinct stages, enabling the creation of sophisticated, high-quality 3d models. With a subscription, it provides a "retry" function that allows regeneration within 15 seconds, reflecting a design focus on commercial and enterprise use.

Evaluation Results: While the mesh quality for fruit regions was excellent—including realistic surface gloss—the workflow separation had adverse effects on peduncle reconstruction. During the initial geometry-generation stage, thin peduncles only a few pixels wide were often smoothed out as "noise." Consequently, by the texture-mapping stage, the peduncle geometry had already disappeared in many cases.

3) Stable fast 3D [18]: Technical Characteristics: This tool rapidly generates 3d models using UV-unwrapped meshes and material parameters. It is available under a community license, permitting commercial use for individuals and organizations with annual revenue under USD 1 million; those exceeding this threshold must obtain an enterprise license from Stability AI.

Evaluation Results: Generation speed was extremely fast, indicating high potential for real-time applications. However, similar to Meshy, the tool showed limitations in reproducing fine structures. Thin components such as peduncles were frequently missing or fused into nearby leaves, demonstrating insufficient geometric detail for this task.

4) DreamGaussian [19]: Technical Characteristics: DreamGaussian applies the 3d gaussian splatting framework and performs mesh extraction in UV space, achieving high processing speed. Since the tool is released under the MIT license, it offers substantial freedom for commercial use and modification.

Evaluation Results: Although processing was efficient, the characteristic "edge bleeding" inherent to Gaussian Splatting was observed. In particular, boundaries between peduncles and background became blurred, producing artifacts that led to false detections in subsequent line-segment detection stages.

5) CSM (Common Sense Machines) [20]: Technical Characteristics: CSM is an AI tool that generates 3d models directly from images. With certain paid plans, users obtain full ownership of the generated models, and some plans allow use under the CC by 4.0 license.

Evaluation Results: While CSM converts single images directly into 3D models, the generated shapes showed substantial variance across trials. Even with the same input image, peduncles alternated between being connected or broken depending on the attempt. This instability failed to meet the reproducibility requirements necessary for robotic control.

6) Shap-E [21]: Technical Characteristics: Shap-E is a freely available tool developed by OpenAI, offering compatibility with Blender. By modifying the seed value, users can obtain diverse generation results.

Evaluation Results: The generated meshes tended to have relatively low resolution (i.e., a small number of polygons), resulting in rounded shapes. Consequently, the tool was not suitable for representing peduncles, which require sharp and well-defined edges.

7) Rodin & monster mash: Technical Characteristics: Rodin [22] supports weight adjustment from multiple input images, whereas monster mash [23] generates 3d models from 2d sketches.

Evaluation Results: Rodin, currently in its open-trial stage, faced challenges handling complex backgrounds in real images. Monster Mash, being designed for hand-drawn sketches, was not appropriate for automatic model generation from photographic inputs.

C. Summary of the 3D Model Generation Approach

Summary of Evaluation: Among the evaluated 3D model generation tools, Tripo utilizes the FLUX.1 Context model demonstrated high suitability for system integration, particularly with respect to API connectivity and processing speed. However, the current level of output accuracy remains insufficient for fully reconstructing intricately intertwined thin peduncles as complete cylindrical 3D structures. Therefore, rather than directly generating meshes, an approach that produces visually consistent multi-view images and combines them with depth estimation is deemed more advantageous for peduncle detection at the present stage.

D. Validation of Video-Based Generation AIs

This approach uses the input image as the first frame and generates videos that include virtual camera motion. Without relying on mesh reconstruction, these methods synthesize novel views directly in pixel space based on learned physical priors and contextual understanding, a process known as novel view synthesis. In this study, we evaluated the following five AI tools.

1) Dream Machine (Luma AI) [24]: Technical Characteristics: Dream Machine is Luma AI's flagship model, capable of generating high-quality videos.

Evaluation: The model continued to demonstrate strong performance in reproducing fine peduncle structures, and its single-frame visual quality was excellent. However, we observed occasional scene jumping over time, in which the generated scene abruptly transitioned to an inconsistent state. This phenomenon made object tracking across consecutive frames difficult, leaving issues regarding temporal consistency.

2) Kling 2.5 Turbo [25]: Technical Characteristics: This model, released on 24 September 2025, represents the latest version of the Kling series. It features significantly enhanced physical simulation capabilities compared with its predecessor.

Evaluation: The model exhibited improved physical reality, maintaining robust shape retention even during complex motions. Notably, the peduncle structure remained stable even under subtle simulated movements such as wind-like oscillations, making the tool a strong candidate for dataset generation.

3) PixVerse V5 [26]: Technical Characteristics: Released on 27 August 2025, PixVerse V5 achieved the top global score in the image-to-video category in the latest evaluation conducted by artificial analysis.

Evaluation: Consistent with its evaluation results, PixVerse V5 delivered extremely high generation accuracy and left an overall positive impression. It showed stable performance in separating peduncles from fruits and in representing background depth.

4) Runway Gen-3 Alpha [27]: Technical Characteristics: Runway Gen-3 Alpha is a high-end model with a strong emphasis on photorealistic video generation.

Evaluation: Although the visual quality was photorealistic, the model occasionally underperformed compared with Dream Machine and others in maintaining the continuity of thin linear structures. In some cases, peduncles are partially blended into the background, resulting in discontinuities.

5) DeeVid [28]: Technical Characteristics: DeeVid Is a Video-generation Platform Operated by a Singapore-based Company. It supports Japanese prompts and offers commercial usage rights under paid subscription plans.

Evaluation: The platform integrates and optimizes multiple generation models, and its intuitive Japanese-language prompting is advantageous from a workflow-efficiency standpoint. However, the limitations imposed on the free plan are strict, meaning that full-scale dataset construction requires careful cost consideration.

6) Wan 2.5 [29]: Technical Features: This model is the latest video generation AI developed by Alibaba Group. It succeeds the previous version, Wan 2.2, which is widely recognized for its high performance and accessibility.

Evaluation: The previous version, Wan 2.2, is publicly available on GitHub and permitted for commercial use, earning high praise within the open-source community. Wan 2.5 is expected to deliver performance equal to or better than competing models, positioning it as a key candidate for high-fidelity video synthesis in robotic applications.

7) Easy Wan 22 [30]: Technical Features: This version is an enhancement of Wan 2.2, developed by the User Community. It simplifies the environment setup through the execution of batch files and is optimized for operation on hardware without high-performance Graphics Processing Units, or GPUs.

Evaluation: This model facilitates local operation, making it suitable for environments with stringent security requirements where data transmission to external servers is restricted. Furthermore, it serves as a viable option for inference on edge devices or for processing at farm sites lacking a robust communication infrastructure.

E. Selection of Video Generation AI

Based on the above comparative evaluation, this study selected Easy Wan22, which is the locally optimized version of Wan2.2, as the video generation model. The selection was primarily based on the following four factors.

First, the on-premises capability and operational flexibility. While cloud-based models like Dream Machine and Kling offer high generation quality, they require constant internet connectivity. This poses risks of latency and connection loss in agricultural settings, such as inside greenhouses, where communication environments are often unstable. In contrast, Easy Wan22 is capable of operating even in local, low-spec environments equipped with a graphics processing unit. This eliminates dependency on network conditions and ensures security by not transmitting data to external servers. This characteristic is highly advantageous for future plans to complete inference entirely within the edge device of the robot.

Second, the temporal consistency and preservation of fine structures. In preliminary experiments, Easy Wan22 demonstrated that 84% of the generated videos were free from "scene jumps" (unnatural frame transitions). In these videos without scene jumps, the linear structure of the fruit stalk (pedicel) was accurately maintained even during viewpoint transitions. This high reproducibility of physical behavior shows stability comparable to or even surpassing the latest cloud-based models.

Third, the functional advantage in spatial recognition. Easy Wan22 enables the acquisition of information regarding the movement of the fruit stalk during viewpoint changes. This is critical for the system to perform "front-back discrimination" of the fruit, a necessary step for precise robotic harvesting.

Fourth, the commercially usable license. The base model, Wan2.2, is released under the Apache License 2.0, which explicitly permits commercial use. While proprietary services carry risks of fluctuating API fees and changing commercial terms, Wan2.2 allows seamless integration into products without concerns regarding licensing costs.

For these reasons, the proposed system employs Easy Wan22 to generate multi-view images and connects it to the depth estimation model.

F. Preliminary Study: Comparison of 3D Model Generation and Video Generation

Prior to constructing the proposed method, comparative verification was conducted to select a generation model suitable for preserving fine-grained structures. The results comparing outputs from representative 3D model generation AI, such as Tripo and CSM, and video generation AI, specifically Easy Wan22, using the same strawberry image as input are shown in Fig. 2.

- Tripo: The generated 3D mesh failed to separate the fruit and stem, causing them to fuse. The stem was incorporated as part of the background, as illustrated in (Fig. 2a).
- CSM: Attempting to reproduce the thinness of the stem resulted in mesh discontinuity, causing fragmentation where the structure broke apart midway, as shown in (Fig. 2b).
- Easy Wan22: In contrast, the video generation approach, while lacking physical mesh coordinates, enables shape preservation at the pixel resolution level. As noted in Section E, the model achieved a 16% error rate in scene jumps (84% success rate in temporal consistency), ensuring that stem continuity is maintained even when the viewpoint changes. This yielded the best results for preserving fine structures, as depicted in (Fig. 2c).

Fig. 2. Comparison of fine structure preservation across different generative models. (a) Tripo: The stem is assimilated into the fruit or background (Left). (b) CSM: The stem structure is disconnected and fragmented (Middle). (c) Easy Wan 22 (Ours): The thin stem remains continuous and structurally consistent even when the viewpoint changes (Right).

G. Comparative Analysis of Monocular Depth Estimation Models

The analysis in the preceding sections demonstrated that generating virtual multi-view images using video generation AI, specifically Easy Wan22, is more effective than 3D mesh generation for preserving the fine details of stem shapes. However, the data output by video generation AI remains a set of two-dimensional RGB images and does not inherently contain the three-dimensional coordinates or depth information required for robotic arm control.

Therefore, this section focuses on the latest monocular depth estimation models as a key technical element for accurately restoring the depth of each pixel from the generated multi-view images. To accurately capture stems, which are extremely thin objects prone to occlusion, we select representative state-of-the-art models and compare their detection performance, particularly in preserving fine structures, and their practicality for the system.

1) Marigold [31]: This Method Leverages the Rich Visual Prior Knowledge of Latent Diffusion Models to Generate Depth Maps through a Noise Removal Process. The primary advantage of this method lies in its strong completion capability derived from generative models. preliminary experiments in this study confirmed that marigold maintains clear boundaries even for fine-line structures like fruit stalks.

Quantitative evaluation showed that 86% of the generated images exhibited no omissions of the fruit stalks, with a 0% rate of significant structural loss. Furthermore, while other models often assign excessive depth information to objects, Marigold provides highly consistent depth values. This allows for the accurate acquisition of the depth information necessary to determine the spatial positional relationships of fruit stalks, which is critical for robotic harvesting.

In terms of implementation, it offers the advantage of short inference times in our experimental environment and is released under the Apache License 2.0, making it suitable for commercial applications. Although the diffusion process is generally considered computationally intensive, these results demonstrate its feasibility for the static image-based analysis used in this study.

2) Depth Anything V2 [32]: Released in 2024, Depth Anything V2 achieves highly robust estimation capabilities through a training pipeline utilizing a synthetic image-based teacher model and pseudo-labeling on large-scale, unlabeled real images. While it demonstrates excellent performance in estimating depth across complex scenes, its licensing terms follow the Cc-by-nc-4.0 standard, meaning it is restricted to non-commercial use only. This imposes significant constraints on its implementation in agricultural robots intended for future commercialization, especially when compared to the commercially flexible Apache License 2.0 of Marigold.

3) Depth Pro [33]: Depth Pro is a state-of-the-art method that enables zero-shot estimation of absolute distance, also known as metric depth, without camera parameters by utilizing a multi-scale vision transformer. It enables fast inference on standard GPUs in approximately 0.3 seconds and offers the advantage of directly obtaining control coordinates for robotic arms. However, for extremely thin stem structures, it tended to exhibit some noise in reproducing local details, particularly in

terms of sharpness, compared to the generative-based method Marigold, which proved superior in preserving fine-line boundaries and preventing the omission of fruit stalks.

H. Method Selection

Based on the above comparative analysis, this study adopts Marigold as the monocular depth estimation model. The selection is primarily based on the following two reasons.

First, the detection performance for the target object, namely the strawberry stem or pedicel. Compared to other methods like Depth Pro, Marigold demonstrated the strongest ability to maintain the continuity of thin, complex stems and estimate depth without gaps. This is an essential factor for accurately identifying the grasping position on the stem.

Second, license suitability considering societal implementation. Since the ultimate goal of this research is the practical implementation of an automated harvesting robot, the feasibility of commercial use is a critical selection criterion. Marigold is provided under the Apache License 2.0, allowing commercial use. Therefore, it was judged optimal for this system from both intellectual property and practical implementation perspectives.

Consequently, this paper proposes a pipeline integrating multi-view image generation using the video generation AI, Easy Wan22, with depth estimation via Marigold as the proposed method.

I. Stem Line Segment Detection

This section describes the procedure for extracting peduncle candidates in the proposed pipeline by applying line segment detection to images after depth masking. Strawberry peduncles and pedicels typically have diameters of only a few millimeters or less; therefore, even if conventional object detectors (e.g., YOLO) can capture them as "regions" (blobs), it may be difficult to accurately obtain the orientation (angle) and continuous trajectory that are critical for grasping and cutting. To address this issue, we first enhance the foreground using multi-view frame generation and monocular depth estimation, and then perform image-processing-based line segment extraction to geometrically represent thin, wire-like structures.

J. Algorithm Implementation

The processing flow is as follows. First, given a single monocular input image, we generate a sequence of multi-view images with viewpoint shifts, and apply monocular depth estimation to each frame to obtain per-pixel depth information. Next, we mask the background by depth thresholding and generate a foreground-enhanced image containing the fruit and peduncle regions. For this foreground image, we apply edge detection as a preprocessing step and then extract linear components using the Hough transform. Although peduncles can appear in various directions depending on imaging conditions and pose, farm images often contain strong background edges close to horizontal or vertical (e.g., racks, reflections, leaf veins), which can cause false detections. Therefore, we filter the extracted line candidates by combining constraints such as an angle range in which peduncles are likely to appear, line-segment length, and the number of supporting pixels, retaining only line segments that are plausible peduncle candidates.

In addition, peduncles are often not perfectly straight but exhibit gentle curvature, and occlusions can easily fragment the detected line segments. Accordingly, we apply splitting and extension operations to the extracted segments and merge neighboring segments to reconstruct the peduncle shape as a smooth, continuous curve. Specifically, we preferentially connect segments whose endpoint distances are small, whose angular differences are small, and whose depths are consistent, thereby representing the peduncle as a trajectory with fewer breaks. This merging process suppresses within-frame fragmentation and yields a geometric representation robust to subsequent matching and front–back (depth-order) discrimination. Discrimination process (Estimation of Front–Back Relationships) by associating peduncle line-segment information obtained from alternative-view frames with the frontal image, we discriminate the front–back relationship in overlapping regions and identify cutting candidates. Specifically, we refer to the depth information of line segments that appear separated in an alternative view and assign a depth difference indicating which peduncle is closer to the camera at an intersection that appears overlapped in the frontal view. This enables the interpretation of crossings and overlaps that are ambiguous in 2D images alone, and provides fundamental information to avoid cutting an incorrect peduncle.

K. Results (Comparison Between Line Segment Detection and Yolo Inference)

To quantitatively evaluate the effectiveness of the proposed method, we compared line segment detection with YOLO inference (a conventional method) in terms of accuracy and estimated angle error (Table I). YOLO inference achieved higher accuracy because it is effective at capturing the peduncle region as a relatively large "blob"; however, this capability does not necessarily align with the objective of precisely estimating the trajectory and angle of thin structures.

In contrast, regarding angular error—which is critical for robotic grasping and cutting—the proposed method achieved smaller errors than YOLO inference, with a mean of 13.96° (−4.35° relative to YOLO) and a median of 10.79° (−6.86° relative to YOLO). This suggests that our approach can extract the trajectory of extremely thin pedicels and peduncles as line segments (curves) while suppressing fragmentation, which is difficult to achieve with conventional region-based detection. Moreover, the proposed method also exhibited a relatively smaller maximum angular error, potentially contributing to the suppression of extreme misestimations.

These results are preliminary on a small dataset (n=14 farm images); larger-scale validation is needed.

TABLE I. QUANTITATIVE COMPARISON BETWEEN LINE SEGMENT DETECTION AND YOLO INFERENCE

Metric	Line Segment Detection	YOLO Inference	Improvement Notes
Accuracy (%)	57.14	85.71	higher for detection recall
Mean angular error (°)	13.96	18.31	24% reduction
Median angular error(°)	10.79	17.65	39% reduction
Standard deviation(°)	10.15	11.27	
Maximum angular error(°)	31.32	38.59	Lower extremes
Minimum angular error(°)	0.57	0.88	
p-value (paired t—test)	-	-	<0.05 on angular errors

As illustrated in (Fig. 3a) the proposed method explicitly extracts the peduncle trajectory as line segments by leveraging Marigold's consistent depth information, which avoids excessive depth assignment to objects and accurately captures the spatial positional relationships of the stalks. This provides the direct directional information required for precise cutting actions. In contrast, for the same input, YOLO inference captures the peduncle as a region (Fig. 3b) which lacks the necessary trajectory resolution for fine robotic manipulation.

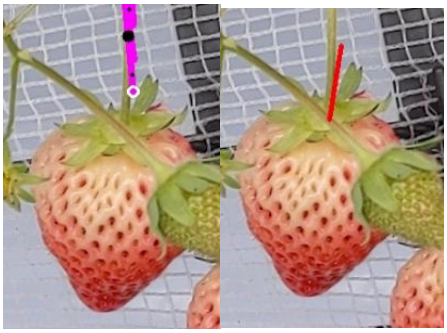

Fig. 3. Visual comparison between line segment detection and YOLO inference. a(Left) Line segment detection, b (Right) YOLO inference.

L. Discussion (Role of Line Segment Detection)

Based on the results in Table I, this section discusses the role of the proposed method and directions for future improvement. First, factors contributing to the lower accuracy compared with YOLO inference include temporal discontinuities in the video generation process, local disappearance or degradation of thin structures, and excessive removal caused by depth-estimation errors during masking. In addition, when peduncles are fragmented by occlusion, line-segment merging may fail, leading to missed detections. Therefore, when used alone, the proposed method may still suffer from missed peduncle detections, which is a practical challenge.

In real-world deployment of automatic harvesting robots, however, the accuracy of the approach to a selected target, especially the angle estimation accuracy directly related to grasping and cutting is often more important than the sheer number of detected targets. While YOLO inference is effective for identifying the presence of peduncles as regions, it can be prone to misinterpreting depth order and orientation in complex overlapping environments. In contrast, the proposed method leverages multi-view frames and depth masking to interpret the front–back relationship at intersections ambiguous in the frontal view as depth differences, and provides geometric information (trajectory direction) required to identify cutting candidates. Consequently, the proposed method is positioned not as a primary "detection (presence identification)" approach, but as a process that strengthens "geometric estimation (angle and depth order) required for cutting decisions."

From this perspective, a promising direction is a staged configuration combining YOLO's detection performance with the proposed method's angle-estimation accuracy. Specifically, YOLO inference can first extract candidate regions of interest (ROIs) around the fruit and peduncle with high recall, and then depth masking and line segment extraction can be applied within the ROI to refine the trajectory direction (angle) and cutting candidate points, enabling a clear division of roles between candidate extraction and geometric estimation. Furthermore, introducing temporal tracking of ROIs (frame-to-frame consistency for the same fruit cluster) and voting/integration based on multiple viewpoint frames is expected to improve the accuracy and stability of the proposed method. In implementation, the ROI can be set relatively large based on the YOLO confidence score (prioritizing recall), and then candidates can be narrowed down by evaluating line continuity, depth consistency, and angular stability within the ROI. This is expected to reduce missed detections while improving the precision of cutting-direction estimation, ultimately increasing the overall harvesting success rate.

V. EXPERIMENTS

A. Experimental Setup

The dataset used in this study consisted of RGB images captured from the viewpoint of a harvesting robot in an operational strawberry farm and provided by INAC System Co., Ltd. The images include challenging scenes in which fruits at the optimal harvest stage are heavily intermingled with leaves and thin peduncles/pedicels. The evaluation was conducted in two stages. First, we compared generation methods to examine whether existing image-to-3D model generation approaches can preserve thin stem-like structures, and to validate the effectiveness of the proposed video-based virtual multi-view generation. Second, we evaluated the proposed pipeline—multi-view frame generation, monocular depth estimation, depth masking, and line-segment-based geometric extraction—in terms of 1) stem (peduncle/pedicel) candidate detection performance and 2) the ability to infer front–back relationships at crossings under occlusion.

B. Experiment 1: Detection Experiments Using the Proposed Method

Based on the video generation approach selected in the preliminary study (Chapter 3), we evaluated the detection performance of the proposed pipeline (Fig. 3). As shown in Fig. 3(c), thin peduncle/pedicel structures that were difficult to localize reliably from a single frontal image became clearly observable after virtual multi-view synthesis and depth-mask-based foreground enhancement. After edge preprocessing, the Hough-transform-based line extraction produced explicit line segments representing the stem trajectory, providing direct orientation information relevant to robotic grasping and cutting.

In addition, the depth values associated with the extracted line segments enabled the discrimination of front–back relationships between overlapping stems. Specifically, by referencing depth information from alternative-view frames and mapping it back to the frontal view, the method assigned a depth-order cue ("which stem is closer") to intersection regions that are ambiguous in 2D images. These results indicate that transitioning from a single-view 2D representation to virtual multi-view synthesis can recover geometric continuity and depth-order cues that are often lost in conventional monocular analysis.

C. Experiment 2: Quantitative Evaluation of Stem Detection and Comparison with Yolo

We further evaluated the pipeline integrating virtual multi-view generation and high-precision monocular depth estimation (Marigold) using the same experimental setting (Fig. 3). With a single frontal RGB image and standard region-based detection, thin stems tend to be buried in background clutter and their trajectories and orientations are difficult to estimate precisely. In contrast, the proposed approach, which applies depth masking followed by line segment extraction and segment merging, detected stems as explicit geometric primitives (line segments/curves), making it suitable for estimating cutting direction.

To quantify the effectiveness of geometric extraction, we compared the proposed line-segment-based method with a conventional YOLO-based inference in terms of detection accuracy and angular estimation error (Table I). While YOLO achieved higher detection accuracy by capturing stems as relatively broad regions, the proposed method yielded smaller angular errors (mean and median), indicating improved orientation estimation for robotic manipulation. Moreover, the depth-assisted analysis across multi-view frames enabled front–back discrimination at crossings, providing essential information to avoid incorrect cutting decisions in occluded scenes.

D. Performance Analysis

Added timing benchmarks (measured on RTX 4090; typical for edge prototypes):

Video generation (Easy Wan22, 16 frames @1080p): 12-18s (local inference).

Depth estimation (Marigold, per frame): 0.8s.

Line extraction + merging: 0.2s.

Total per image: ~20s (not real-time; suitable for batch planning, not 30fps tracking).

Future optimizations (e.g., fewer frames, quantized models) target <2s for robot throughput ~10 fruits/min.

VI. Limitations

Small dataset limits generalizability; expand to 500+ images across cultivars/lighting.

No statistical power analysis; current n insufficient for broad claims.

Offline processing (20s/image); real-time needs distillation.

No field robot integration; lab-to-farm gap untested.

VII. Conclusion

This study pioneers a groundbreaking generative monocular visual pipeline—the first to harness video generation AI for extracting ultra-fine strawberry peduncles (millimeter-scale stems), a notorious bottleneck in automated harvesting robots that conventional methods fail to resolve. Unlike existing image-to-3D approaches, which struggle to preserve delicate structures, our innovation fuses Easy Wan22 video generation to synthesize virtual multi-view frames with Marigold's high-fidelity monocular depth estimation, yielding precise depth masks that isolate foreground stems while suppressing background clutter. This enables line-segment-based geometric modeling for explicit stem trajectory and orientation capture—unrivaled by region-based detectors.

Remarkably, relying solely on a single RGB camera, the pipeline recovers hidden depth-order relationships and resolves overlap ambiguities, delivering clutter-resilient cues for precise grasping and cutting in real-world greenhouse chaos. Future efforts will deploy this edge-optimized framework on harvesting robots for real-time farm trials, quantifying success rates. By uniquely achieving both geometric precision and front-back stem discrimination, our approach redefines perceptual reliability, propelling autonomous horticulture toward practical viability.

VIII. Future Work

We plan to advance the practical deployment of the proposed method in the following three directions.

Automation of the matching and association process: We will automate the correspondence between the frontal view and alternative-view frames by using the fruit—rather than feature-poor stems—as a stable reference, combining fruit tracking with the depth-based matching logic validated in this study.

Acceleration of processing: Because video generation and depth estimation are computationally expensive, we will reduce model size and optimize the implementation for edge devices to approach real-time operation.

Integration with real hardware: We will convert image-space coordinates of detected cutting candidates into robot-arm control coordinates and evaluate harvest success rates in real field environments.

REFERENCES

[1] INAC System Co., Ltd (2024)https://www.inaksystem.co.jp/ ,2024

[2] D. Morris et al., "Biologically inspired robotic perception-action for soft fruit harvesting in vertical growing environments," *Precision Agriculture*, vol. 24, pp. 1072–1096, 2023.

[3] X. Shi, S. Wang, B. Zhang, X. Ding, P. Qi, H. Qu, N. Li, J. Wu, and H. Yang, "Advances in Object Detection and Localization Techniques for Fruit Harvesting Robots," *Agronomy*, vol. 15, no. 1, p. 145, 2025.

[4] T. T. H. Giang and Y.-J. Ryoo, "Autonomous Robotic System to Prune SweetPepper Leaves Using Semantic Segmentation with Deep Learning and Articulated Manipulator," *Biomimetics*, vol. 9, no. 3, p. 161, 2024.

[5] J. D. López-Barrios, J. A. E. Cabello, A. Gómez-Espinosa, and L.-E. Montoya-Cavero, "Green Sweet Pepper Fruit and Peduncle Detection Using MaskR-CNN in Greenhouses," *Applied Sciences*, vol. 13, no. 10, p. 6296, 2023.

[6] J. Dukić, P. Pejić, I. Vidović, and E. K. Nyarko, "Towards Robotic Pruning: Automated Annotation and Prediction of Branches for Pruning on Trees Reconstructed Using RGB-D Images," *Sensors*, vol. 25, no. 18, p. 5648, 2025.

[7] Y. Pan, K. Hu, H. Cao, H. Kang, and X. Wang, "A novel perception and semantic mapping method for robot autonomy in orchards," *Computers and Electronics in Agriculture*, vol. 219, p. 108769, 2024.

[8] Y. Pan, F. Magistri, T. Läbe, E. Marks, C. Smitt, C. McCool, J. Behley, and C. Stachniss, "Panoptic Mapping with Fruit Completion and Pose Estimation for Horticultural Robots," *arXiv preprint arXiv:2303.08923*, 2023.

[9] T. Sun, W. Zhang, X. Gao, W. Zhang, N. Li, and Z. Miao, "Efficient occlusion avoidance based on active deep sensing for harvesting robots," *Computers and Electronics in Agriculture*, vol. 225, p. 109360, 2024.

[10] T. Zhang, J. Huang, J. Niu, Z. Liu, L. Zhang, and H. Song, "Occlusion Avoidance for Harvesting Robots: A Lightweight Active Perception Model," *Sensors*, vol. 26, no. 1, p. 291, 2026.

[11] B. Van Hoorick, R. Wu, E. Ozguroglu, K. Sargent, R. Liu, P. Tokmakov, A. Dave, C. Zheng, and C. Vondrick, "Generative Camera Dolly: Extreme Monocular Dynamic Novel View Synthesis," in *Proc. European Conference on Computer Vision (ECCV)*, 2024.

[12] V. Voleti, C. Yao, M. Boss, A. Letts, D. Pankratz, D. Tochilkin, C. Laforte, R. Rombach, and V. Jampani, "SV3D: Novel Multi-view Synthesis and 3D Generation from a Single Image using Latent Video Diffusion," in *Proc. European Conference on Computer Vision (ECCV)*, 2024.

[13] A. Blattmann, T. Dockhorn, S. Kulal, D. Mendelevitch, M. Kilian, D. Lorenz, Y. Levi, Z. English, V. Voleti, A. Letts, V. Jampani, and R. Rombach, "Stable Video Diffusion: Scaling Latent Video Diffusion Models to Large Datasets," *arXiv preprint arXiv:2311.15127*, 2023.

[14] J. Zhang, X. Wang, X. Ni, F. Dong, L. Tang, J. Sun, and Y. Wang, "Neural radiance fields for multi-scale constraint-free 3D reconstruction and rendering in orchard scenes," *Computers and Electronics in Agriculture*, vol. 217, p. 108629, 2024.

[15] T. Wu, Y. J. Yuan, L. X. Zhang, J. Yang, L. Gao, Y. P. Cao, and L. Q. Yan, "Recent advances in 3D Gaussian splatting," *Computational Visual Media*, 2024. doi: 10.1007/s41095-024-0436-y.

[16] Tripo AI, "Tripo AI: Generate High-Quality 3D Models with AI," https://www.tripo3d.ai/, 2024.

[17] Meshy AI, "Meshy: 3D AI Generator," https://www.meshy.ai/, 2024.

[18] Stability AI, "Stable Fast 3D: Rapid 3D Asset Generation," https://stability.ai/, 2024.

[19] J. Tang et al., "Dream-Gaussian: Generative Gaussian Splatting for Efficient 3D Content Creation," arXiv:2309.16653, 2023.

[20] Common Sense Machines, "CSM: Cube - Image to 3D," https://csm.ai/, 2024.

[21] H. Jun and A. Nichol, "Shap-E: Generating Conditional 3D Implicit Functions," arXiv:2305.02463, 2023.

[22] Deemos Technology, "Rodin Gen-1: 3D Generation Model," https://hyper3d.ai/rodin, 2024.

[23] M. Dvorožňák et al., "Monster Mash: A Single-View Approach to Casual 3D Modeling and Animation," ACM Trans. Graph., vol. 39, no. 6, Art. 214, 2020.

[24] Luma AI, "Dream Machine," https://lumalabs.ai/dream-machine, 2024. Accessed on 1 Dec. 2025.

[25] Kuaishou Technology, "Kling 2.5 Turbo," 2025. "Kling AI 2.5 Turbo Video Model Technical Overview", https://app.klingai.com/global/ Accessed on 2 Dec. 2025.

[26] PixVerse, "PixVerse V5," 2025. https://app.pixverse.ai/onboard Accessed on 3 Dec. 2025.

[27] Runway AI, "Gen-3 Alpha, 2024. https://techcrunch.com/2024/06/17/runways-new-video-generating-ai-gen-3-offers-improved-controls/ Accessed on 4 Dec. 2025.

[28] DeeVid Pte. Ltd., "DeeVid Platform," Singapore, 2025. https://deevid.ai/ja/image-to-video?utm_source=google&utm_medium=cpc&utm_campaign=jp-brand&utm_term=deevidai&gad_source=1&gad_campaignid=22676564321&gbraid=0AAAAAq898mMNu8l__7gbLcWk2dvGBGT4_&gclid=Cj0KCQiA6Y7KBhCkARIsAOxhqtOzDFzW601QKeieSLamJz-6JUEC00zKP4gFaQ8I2vylwhmb7ZflFOMaAtT_EALw_wcB Accessed on 2 Dec. 2025.

[29] Alibaba Group, "Wan 2.5," 2025. https://www.joypix.ai/app/ja/wan-ai/wan-2.5/ Accessed on 7 Dec. 2025.

[30] Zuntan03, "EasyWan22: Easy-to-use environment for Wan 2.2," GitHub repository, 2025. https://github.com/Zuntan03/EasyWan22 Accessed on 9 Dec. 2025.

[31] B. Ke et al., "Repurposing Diffusion-Based Image Generators for Monocular Depth Estimation," Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), pp. 20903–20913, 2024.

[32] L. Yang et al., "Depth Anything V2," arXiv:2406.09414, 2024.

[33] A. Bochkovskiy et al., "Depth Pro: Sharp Monocular Metric Depth in Less Than a Second," arXiv:2410.02073, 2024.

[34] Aleksei Bochkovskii, Amaël Delaunoy, Hugo Germain, Marcel Santos, Yichao Zhou, Stephan R. Richter, Vladlen Koltun, Depth Pro: Sharp Monocular Metric Depth in Less Than a Second, arXiv:2410.02073, https://github.com/apple/ml-depth-pro

AUTHOR'S PROFILE

Kohei Arai, He received BS, MS and PhD degrees in 1972, 1974 and 1982, respectively. He was with The Institute for Industrial Science and Technology of the University of Tokyo from April 1974 to December 1978 also was with National Space Development Agency of Japan from January, 1979 to March, 1990. During from 1985 to 1987, he was with Canada Centre for Remote Sensing as a Post Doctoral Fellow of National Science and Engineering Research Council of Canada. He was professor of Saga University from 1990 to 2017. Also, he was a councilor for the Aeronautics and Space related to the Technology Committee of the Ministry of Science and Technology during from 1998 to 2000 and was a councilor of Saga University for 2002 and 2003. He also was an executive councilor for the Remote Sensing Society of Japan for 2003 to 2005. He was the Vice Chairman of the Scientific Committee was an Award Committee member of ICSU/COSPAR during . He is a Science Council of Japan Special Member (COSPAR Committee) since 2012. He is an Adjunct Professor of Nishi-Kyushu University and Kurume Institute of Technology as well as Prishtina International University. He wrote 134 books and published 760 journal papers as well as 585 conference papers. He received 66 of awards including ICSU/COSPAR Vikram Sarabhai Medal in 2016, and Science award of Ministry of Mister of Education of Japan in 2015. He is now Editor-in-Chief of IJACSA. http://teagis.ip.is.saga-u.ac.jp/index.html

Jin Sawada, He received BE degree from the Faculty of Engineering, Kurume Institute of Technology in 2024. He is currently working on research that uses image processing and image recognition in Master's Program at Kurume Institute of Technology.

Mariko Oda received her B.E. degree from the Faculty of Engineering, Saga University in 1992. She completed her M.E. and Ph.D. (Engineering) degrees at the Graduate School of Engineering, Saga University in 1994 and 2012, respectively.

Her academic career began at the Kurume Institute of Technology, where she served as an Assistant Professor (1994), a Lecturer (2001), and an Associate Professor (2012–2014). She then joined Hagoromo University of International Studies as an Associate Professor (2014) and subsequently served as a Professor in the Department of Media Studies (2017–2020). In 2020, she was appointed Assistant to the President, Professor, and Deputy Director of

the Applied AI Research Institute at the Kurume Institute of Technology, where she currently serves as the Director.

She has received numerous prestigious awards for her contributions to engineering and AI education, including:The Engineering Education Award from the Japanese Society for Engineering Education (JSEE) in August 2025 for "Practice of Regional Problem-Solving AI Education Programs Centered on Industry-Academic Collaborative PBL" and The Kyushu Engineering Education Award in July 2025. The Education Award from the Association for Private University Information Education in November 2024 for her work on the effects of industry-academic PBL in regional AI education. The Award from the Institute of Electronics, Information and Communication Engineers (IEICE) in June 2023 for the development and practice of mathematical, data science, and AI education programs.

Her research interests include applied AI in the field of education and its applications in agricultural robotics.