IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

Visual Scene Understanding-based Task Planning for an Efficient Multipurpose Agricultural Robot System

Yonghyun Park[1,3] and Hyoung Il Son[1,2,3,*], *Senior Member, IEEE*

***Abstract*—This study introduces a visual scene understanding (VSU) pipeline that fuses scene graph generation (SGG) with task planning for agricultural robots. Mask R-CNN detects fruits, leaves, and stems; Object features feed heads for predicates and attributes such as rigidity and ripeness. The resulting graph triggers a rule-based planner that chooses among harvesting, pruning, or thinning and decides on single- or dual-arm execution. Evaluated on a re-annotated custom dataset, the full pipeline reaches 38.9% relationship R@50, 70.1% attribute R@50, 72.3% task-decision accuracy, and 53.7% cooperative-control accuracy. Results show dual-arm selection is twice as sensitive to perception errors as task type assignment. The work provides an agriculture-specific task planning that distinguishes flexible from rigid obstacles, demonstrating that relational and attribute improve perception in agricultural scenes.**

***Index Terms*—agricultural automation, robotics and automation in agriculture and forestry, semantic scene understanding, task planning, visual learning**

I. Introduction

MODERN agriculture is moving from labor-intensive work to data-driven systems called intelligent farming [1], [2]. This shift is driven by climate change, labor shortages, and higher demand for crop diversity. Digital and smart agriculture aim to solve these issues with automation and data [3], [4].

However, prevailing paradigms concentrate on localized optimizations, for instance, site-specific input applications or sensor-based monitoring. Consequently, they often grapple with the inherent variability and unpredictability characteristic of real-world farming environments [5], [6]. Crops naturally exhibit diverse morphologies and can present unforeseen obstacles, including intertwined leaves, densely clustered fruits, or seasonal foliage changes. As a result, existing approaches reliant on rudimentary object detection or partial mechanization may fall short of the demands for continuous adaptation and sophisticated decision-making essential in dynamically evolving farm conditions [7], [8].

Manuscript received: March, 31, 2025; Revised May, 26, 2025; Accepted October, 18, 2025.

This paper was recommended for publication by Editor Hyungpil Moon upon evaluation of the Associate Editor and Reviewers' comments. This work was supported by (organizations/grants which supported the work.)

[1] Y. Park and H. I. Son are with the Department of Convergence Biosystems Engineering, Chonnam National University, Yongbong-ro 77, Gwangju 61186, Republic of Korea dk03378@jnu.ac.kr; hison@jnu.ac.kr

[2]H. I. Son is with the Interdisciplinary Program in IT-Bio Convergence System, Chonnam National University, Yongbong-ro 77, Gwangju 61186, Republic of Korea

[3]Y. Park and H. I. Son are with the Research Center for Biological Cybernetics, Chonnam National University, Yongbong-ro 77, Gwangju 61186, Republic of Korea

*Corresponding author

Digital Object Identifier (DOI): see top of this page.

To address these limitations, agricultural robots must transcend their conventional role as mere automation tools, evolving into intelligent systems endowed with greater adaptability [9]. Such an evolution necessitates capabilities akin to scene understanding, where a robot not only detects objects (e.g., fruits, stems, leaves) but also interprets intricate environmental cues, inter-object relationships, and their semantic attributes. This form of contextual awareness is paramount for the successful execution of multi-step or multi-task operations prevalent in agriculture, such as leaf pruning, flower thinning, fruit thinning, and harvesting [10], [11]. These tasks inherently demand the capacity to distinguish between different operational requirements and to dynamically manage occlusions, overlapping objects, and other environmental complexities.

Despite notable advances in computer vision for agriculture—for example, through CNN- or Transformer-based detection and segmentation—many contemporary pipelines persist in treating objects as isolated entities [12]–[14]. A robot confined to recognizing bounding boxes and class labels typically lacks the cognitive layer required to determine, for instance, whether a specific leaf must be preemptively removed, whether multiple fruits warrant simultaneous handling, or if a dual-arm maneuver would better mitigate stem oscillation. This constrained perceptual capability often engenders inefficiencies in task execution, particularly when the robot must differentiate between pliable obstacles that can be displaced or removed (e.g., leaves) and rigid obstacles that necessitate careful avoidance (e.g., main stems or trellising structures) [15], [16].

Scene graph generation (SGG) has surfaced as an effective approach to bridge this cognitive gap by explicitly modeling objects, their inherent attributes, and the pairwise relationships connecting them [17]. Initially conceived for applications such as image captioning and visual question answering, SGG can be effectively adapted to agricultural contexts. Here, nodes within the graph can represent entities like fruits, leaves, and stems, while edges encode predicates such as *occluded* or *attached-to*. This structured, relational representation offers high-level task planning, empowering the robot to make decisions concerning object interactions and task prioritization [18]–[20].

This paper introduces a visual scene understanding (VSU)-based task planning for agricultural robots operating within complex, unstructured environments (Fig. 1). In contrast to prior research that has predominantly focused on isolated object detection or rudimentary task sequencing based on proximity, our proposed method explicitly encodes both inter-object relationships and critical physical attributes (e.g., rigidity, ripeness). By analyzing the constructed scene graph, the robot can: 1) identify and differentiate among multiple agricul-

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

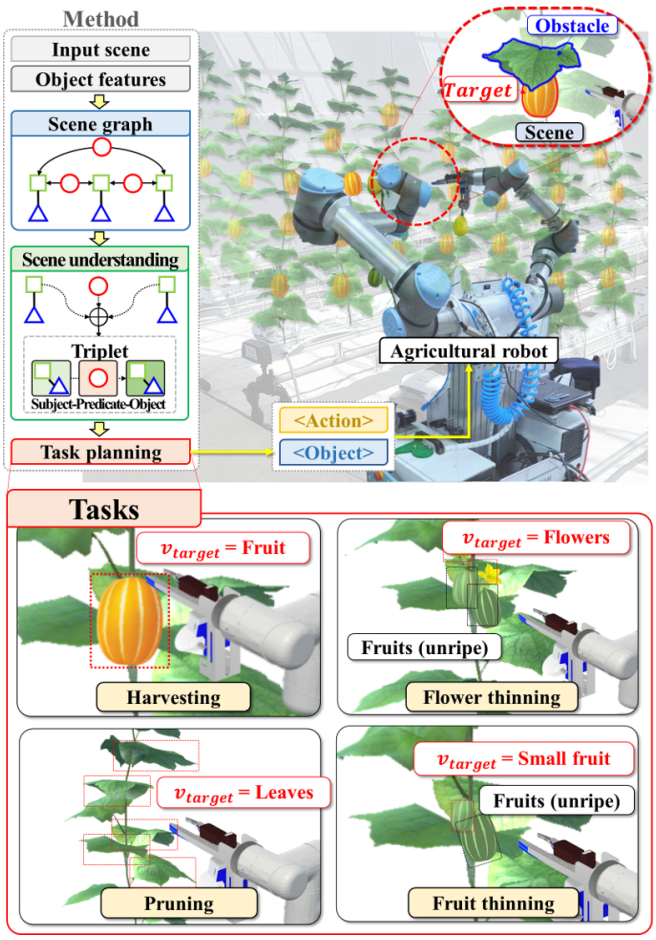

Fig. 1. Flowchart of the proposed system approach.

tural task types (such as harvesting, pruning, and thinning); 2) determine necessary prerequisite actions (e.g., removing flexible occluding leaves while navigating around rigid obstacles); and 3) dynamically re-sequence operational steps as the scene evolves. This approach addresses a significant perception in agricultural robot, where the conventional strategy of universally avoiding all obstacles proves suboptimal compared to selectively managing pliable impediments.

The main contributions of this paper are:

- An SGG framework tailored for agricultural environments, which integrates the representation of objects, their relationships, and pertinent attribute information (e.g., rigidity, ripeness) crucial for sophisticated task planning.
- A context-aware task planning system that exploits the rich relational and attribute structure afforded by SGG, enabling robots to distinguish between diverse agricultural tasks and to make informed decisions regarding obstacle management based on their physical properties.
- Attribute-aware planning scheme is proposed, enabling to guide operations in complex agricultural environments

This research contributes to the development of next-generation agricultural robotic systems possessing enhanced contextual intelligence and greater operational autonomy by embedding a deeper level of relational awareness into their perception and planning modules.

II. Related Work

A. Object Detection in Agricultural Robotics

Object detection is the basis of vision tasks in agricultural robots. It identifies fruits, stems, and leaves. Early methods using color thresholds or handcrafted features worked in controlled labs but failed in real fields with changing light, occlusion, and complex backgrounds [10], [11]. Deep learning methods such as Faster R-CNN, SSD, and YOLO have improved detection speed and accuracy [21]. YOLOv8 in particular shows strong real-time fruit detection and yield estimation [8], [14].

Still, these models focus on bounding boxes and miss context. Tasks like pruning or thinning need more: e.g., checking if a leaf blocks a fruit or if fruits share the same stem. Thus, detection alone is not enough without considering relations among objects in the canopy [13], [22].

B. Scene Understanding in Robotics

Scene understanding means interpreting environments for decision making. Approaches include data fusion, 3D reconstruction, and SLAM [23]. But these often stress geometry and free space, not semantic relations. This works in structured places like homes, where objects follow patterns.

Fields are different: plant layouts and occlusions are highly variable [5], [6]. Some works use attention to find key regions in dense leaves or domain adaptation for crop types, but challenges remain. Leaves tangle, fruits cluster, and stems are half-hidden. The organic and irregular nature of crops demands special scene understanding methods.

C. Scene Graph Generation and Visual Relationship Detection

Scene graph generation (SGG) represents objects, attributes, and relations [17]. It has been used in captioning, VQA, and fine-grained relation tasks. Chalvatzaki et al. [24] linked SGG with language models for robot planning. Agia et al. [25] made Taskography, a framework for task planning with large 3D scene graphs. Ekpo et al. [26] proposed VeriGraph for verifiable planning. Fu et al. [27] combined relation perception with knowledge graphs for dynamic planning. These works focus on indoor tasks. Agricultural scenes are less studied.

D. Task Planning with Contextual Reasoning

Research within the broader field of robotics has indicated that task planning approaches enriched by scene graph representations can lead to more efficient execution, for example, by reducing unnecessary action-switching [18]–[20], [28]. Jiao et al. [18] specifically demonstrated sequential manipulation planning leveraging scene graphs. In a similar vein, Rana et al. [29] explored the use of large language models in conjunction with 3D scene graphs to enable scalable robot task planning.

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

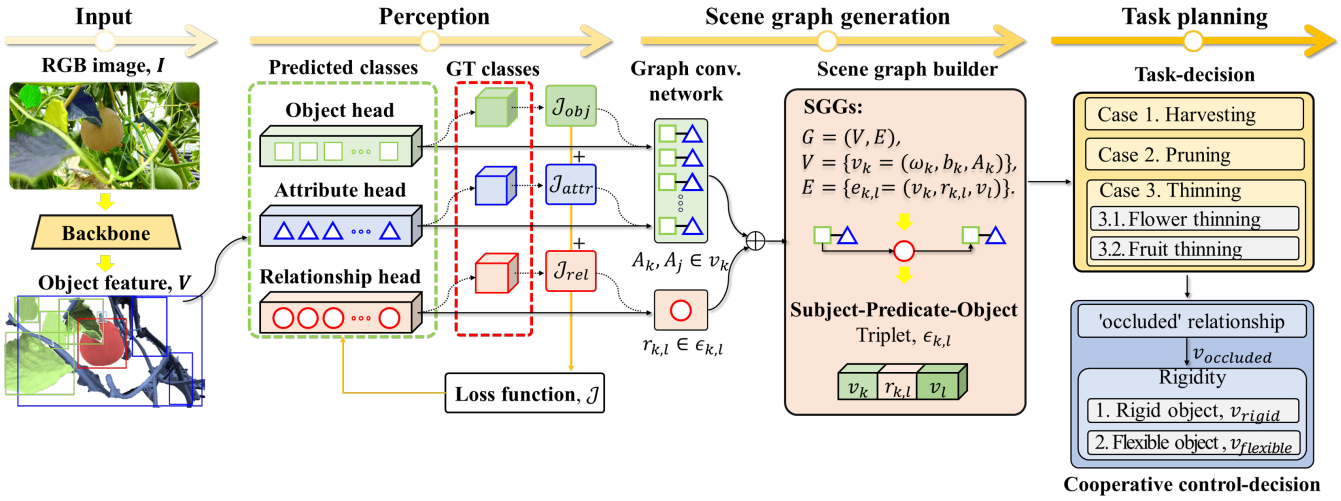

Fig. 2. End-to-end pipeline of the proposed VSU-based agricultural-robot system. RGB input is processed to yield region proposals whose features feed three dedicated heads (class, predicate, attribute). Graph convolution aggregates these cues into a scene graph $G = (V, E)$. This system selects the agricultural task (harvesting, pruning, thinning) and decides whether cooperative dual-arm manipulation is required.

However, a gap persists within agricultural robotics research concerning systems that not only recognize inter-object relationships but also integrate reasoning about physical attributes for more intelligent and adaptive task planning. Many existing agricultural robots tend to treat all obstacles uniformly, typically resorting to avoidance maneuvers irrespective of the obstacle's physical properties. This strategy can be inefficient, particularly when encountering flexible obstacles (e.g., young leaves or pliable stems) that could potentially be managed (e.g., pushed aside or selectively removed) rather than simply avoided [3]. Consequently, a proposed system that captures both object relationships and physical attributes may enable robots to perform multistep processes, such as distinguishing between various agricultural tasks and making intelligent decisions about obstacle management, more effectively than conventional methods.

III. Visual Scene Understanding-based Task Planning

Conventional detection finds objects and locations ("what" and "where") but gives only a shallow view of the scene. Recent work shows that learning object relations (*predicates*) provides deeper understanding [17], [18]. In farms, fruits, leaves, and stems have complex relations. For example, a leaf may *occlude* a fruit, or fruits may share one pedicel. These cues are important for efficient task execution. We therefore integrate SGG to give a relation-aware view of the scene. Fig. 2 shows the pipeline: objects → scene graph → planning (harvest, prune, thin).

A. Scene Graph Generation

1) System overview: A scene graph $G = (V, E)$ offers a structured representation of an image, where V denotes a node set (objects) and E represents an edge set (relationships). SGG extends beyond the capabilities of standard bounding-box detection by explicitly capturing detailed object interactions and their context, which is particularly critical for robust decision-making in multifaceted agricultural tasks [19], [20]. Fig. 2 depicts the conceptual workflow of SGG, while Fig. 2 provides a more detailed schematic of our pipeline, from initial object detection through to relationship inference and graph construction.

2) Object detection and feature extraction: Given an input image I, the CNN detects N objects:

$$p(O \mid I) = \prod_{k=1}^{N} p(c_k, b_k \mid I), \tag{1}$$

where $O = \{(c_k, b_k)\}_{k=1}^{N}$ represents a set of classes c_k and bounding boxes b_k for all detected objects. Each object ω_k has a feature vector $f_k = g(I, b_k)$ and an optional attribute set as follows:

$$A_k \subseteq \{\text{attributes (e.g., ripeness, rigidity, color)}\}. \tag{2}$$

The attribute set A_k is of considerable importance for agricultural operations, as it encodes physical properties (such as the rigidity of stems or pedicels versus the flexibility of leaves, or the ripeness stage of a fruit) that directly influence task planning decisions. For example, these attributes help determine whether an obstructing element can be safely removed or must be carefully navigated around [15], [16].

3) Relationship prediction: For every pair of detected objects (ω_k, ω_l), the relationship prediction model takes their respective feature vectors f_k and f_l as input to infer a predicate $r_{k,l}$:

$$r_{k,l} = R(f_k, f_l). \tag{3}$$

Common examples of such relationships pertinent to agricultural scenes include predicates like `occluded`, `attached to`, `in front of`, and `near`. Understanding these relationships is fundamental for deciphering the complex spatial arrangement of agricultural elements, especially within dense

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

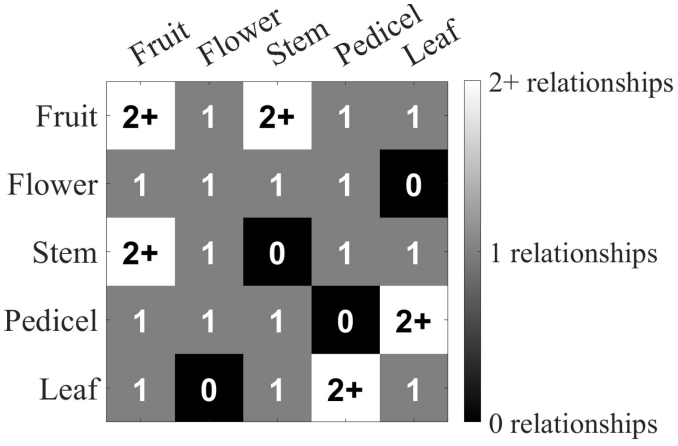

Fig. 3. Ground truth relationship matrix (0: No relationship, 1: one relationship, 2+: Two or more relationships).

TABLE I
EXAMPLE OBJECT, ATTRIBUTE, AND RELATIONSHIP CLASSES FROM THE CUSTOM DATASET.

Objects	Attributes	Relationships
fruit, flower, leaf, pedicel, stem	rigidity (rigid/flexible), ripeness (ripe/unripe), color (red, green, yellow)	attached-to, near, occluded, in-front-of

canopies characterized by multiple occlusions and intertwined plant structures.

Morphological knowledge imposes hard constraints on which object–object relations can physically occur. This method encodes these constraints in a 5×5 binary matrix (Fig. 3).

4) Graph construction: Subsequent to object detection and relationship inference, the system constructs a scene graph G, defined as:

$$\begin{aligned} G &= (V, E), \\ V &= \{\, v_k = (\omega_k, b_k, A_k) \,\}, \\ E &= \{\, \epsilon_{k,l} = (v_k, r_{k,l}, v_l) \,\}, \end{aligned} \tag{4}$$

where each node v_k encodes object class ω_k, bounding box b_k, and associated attributes A_k. Each directed edge $\epsilon_{k,l}$ signifies the predicates $r_{k,l}$ existing between v_k and v_l. This formulation results in triplets of the form $\langle$**subject** $-$ **predicate** $-$ **object**$\rangle$, which serve to explicitly articulate the spatial or semantic connection between elements in the agricultural scene.

5) Training and loss function: The SGG model is trained end-to-end using datasets annotated with object labels, relationship labels, and attribute labels. The composite loss function $\mathcal{J}$ is formulated as:

$$\mathcal{J} = \mathcal{J}_{\text{obj}} + \mathcal{J}_{\text{rel}} + \mathcal{J}_{\text{attr}}, \tag{5}$$

where $\mathcal{J}_{\text{obj}}$ denotes the loss for object detection (both classification and bounding box regression),$\mathcal{J}_{\text{rel}}$ represents the loss for relationship prediction, and $\mathcal{J}_{\text{attr}}$ corresponds to the loss for attribute classification. Table I provides illustrative examples of the object, attribute, and relationship classes utilized in our custom dataset.

B. Visual Scene Understanding

Beyond detecting isolated objects, agricultural robots must reason over *how* items interact. We extend SGG by attaching attributes (rigidity, color, ripeness) to every node and keeping only three task–relevant predicates: `occluded`, `attached-to`, and `near`. This set covers the cues needed for harvesting, pruning, and thinning.

1) Relation–attribute graph: Each image is converted to a graph $G = (V, E)$ where $V = \{v_i = (o_i, b_i, A_i)\}$ stores class o_i, box b_i, and attribute vector A_i, while $E = \{\epsilon_{ij} = (v_i, r_{ij}, v_j)\}$ stores predicate r_{ij}. A typical triplet is

$\langle$**leaf (rigid, green)** $-$ **occluding** $-$ **fruit (flexible, red, ripe)**$\rangle$

which sends the system that the leaf must be removed before grasping the fruit.

2) Occlusion reasoning: Attributes modulate action choice: a `rigid` object obstacles a cutting motion, whereas a `flexible` one is pushed aside.

3) Multi-attribute task sequencing: Combining predicates with ripeness and proximity lets the robot (i) select ripe fruits for harvest, (ii) thin clustered unripe fruits, or (iii) prune excess foliage when leaf density exceeds a threshold.

C. Task Planning

Once the scene graph is constructed and populated, the robot undertakes a two-stage planning process. First, it determines the most appropriate agricultural task to perform from a predefined set and identifies the corresponding target objects, denoted v_{target}. Following this task selection, the system meticulously analyzes the current obstacle conditions as represented in the scene graph to ascertain whether cooperative dual-arm control is necessary or advantageous for the selected task. Based on this comprehensive scene graph analysis, the system selects from four primary agricultural operations: harvesting, pruning, flower thinning, and fruit thinning.

Case 1. Harvesting: If the scene graph contains at least one node v_k classified as a `fruit` and possessing the attribute `ripe`, that fruit is designated as a v_{target} for harvesting (Fig. 1):

$$v_{target} = \{\, v_k \mid A_k = \text{'ripe'}, \omega_k = \text{'fruit'} \,\}. \tag{6}$$

In scenarios presenting multiple ripe fruits, the robot may employ further prioritization logic, potentially based on factors like estimated depth or bounding-box size, to optimize the overall harvesting efficiency [16], [30].

Case 2. Pruning: Excessive or obstructive leaves can significantly impair visibility for perception systems and reduce essential air circulation around targets (Fig. 1). Pruning tasks are initiated if the count of nodes identified as `leaf` within the scene graph surpasses a predefined threshold δ:

$$v_{target} = \begin{cases} v_k, & \text{if } |\{v_k \in V \mid \omega_k = \text{'leaf'}\}| \geq \delta, \\ 0, & \text{otherwise.} \end{cases} \tag{7}$$

The pruning process particularly benefits from attribute information, as this allows the system to differentiate, for example, between rigid stems or pedicels that require cutting and flexible leaves that can be displaced or removed with simpler actions [10], [11].

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

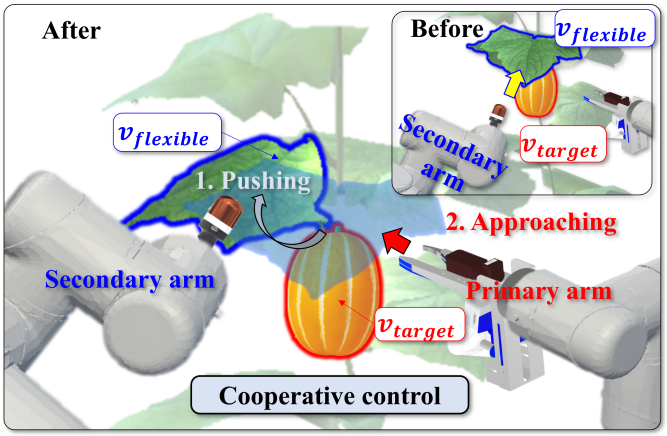

Fig. 4. Cooperative control strategies: 1) Secondary arm pushing a flexible obstacle, 2) Primary-arm operation approaching around a target.

Case 3. Thinning: The practice of thinning, whether targeting flowers or unripe fruits, aims to improve the quality and size of the remaining yield (Fig. 1). This operation is subdivided into flower thinning and fruit thinning based on the specific relationships and attributes identified in E.

Case 3.1. Flower thinning: If a node v_k identified as a `flower` is detected in proximity to multiple nodes v_l classified as `unripe fruit`, specifically where If a `flower` node is detected near multiple `unripe fruit` nodes,

$$\begin{aligned} v_{flower} &= \{\, v_k \mid \omega_k = \text{'flower'}\}, \\ v_{unripe} &= \{\, v_l \mid \omega_l = \text{'fruit'},\ A_l = \text{'unripe'}\}, \end{aligned} \tag{8}$$

and the `near` relationship is identified between $v_k \in v_{flower}$ and $v_l \in v_{unripe}$ for at least two such pairs, these flowers are designated as targets for thinning:

$$\left|\{\, \epsilon_{k,l} \in E \mid v_k \in v_{flower},\, v_k \in v_{unripe},\, r_{k,l} = \text{'near'}\}\right| \geq 2. \tag{9}$$

Case 3.2. Fruit thinning: For the task of thinning unripe fruits,

$$v_{unripe} = \{v_k \in V \mid \omega_k = \text{'fruit'}, A_k = \text{'unripe'}\}. \tag{10}$$

The model prioritizes thinning for a fruit $v_k \in v_{unripe}$ if it is identified as being "near" at least three other unripe fruits, a condition indicative of a dense cluster. Such fruits are considered prime candidates for thinning, as their removal can optimize resource allocation to the remaining fruits within the cluster, potentially leading to improved growth and quality [11], [12].

Cooperative Control Decision: Subsequent to the determination of the primary agricultural task and the identification of the target object(s) v_{target}, the system proceeds to decide whether employing cooperative dual-arm control would be beneficial. This decision is predicated on an analysis of prevailing obstacle conditions derived from the scene graph, with a focus on evaluating occlusion relationships and obstacle attributes to enhance overall execution efficiency (Fig. 1).

The system first identifies all objects occluding the target:

$$v_{occluded} = \{v_j \mid \exists\, \epsilon_{ij} \in E : v_i = v_{target} \wedge r_{ij} = \text{'occluded'}\}. \tag{11}$$

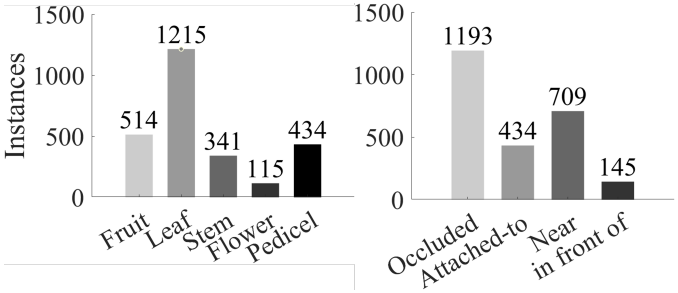

Fig. 5. Dataset statistics for the *Cucumis_melo_300*.

These occluding objects are then categorized based on their 'rigidity' attribute:

$$\begin{aligned} v_{rigid} &= \{v_j \in v_{occluded} \mid A_j = \text{'rigid'}\}, \\ v_{flexible} &= \{v_j \in v_{occluded} \mid A_j = \text{'flexible'}\}. \end{aligned} \tag{12}$$

In the presence of rigid obstacles (such as main stems or thick branches), the system defaults to a single-arm operation, incorporating path planning algorithms to navigate around these structural elements, which typically should not be disturbed. Conversely, if flexible obstacles (like leaves or thin, pliable stems) are identified, cooperative dual-arm control is activated. In this mode, the secondary manipulator is tasked with displacing the flexible obstacle, thereby clearing a path for the primary arm to perform its main designated task. Fig. 4, if included, would provide a visual depiction of these distinct control strategies.

The formal decision logic for engaging cooperative control is:

$$\text{cooperative} = \begin{cases} \text{true}, & \text{if } |v_{flexible}| \geq 1 \\ \text{false}, & \text{otherwise} \end{cases} \tag{13}$$

By integrating such attribute-aware scene understanding into its planning logic, the system can emulate more sophisticated, human-like strategies—actively managing obstacles that can be safely moved or reconfigured while carefully avoiding those that are fixed. This capability is anticipated to enable more efficient and robust task execution compared to conventional robotic approaches that often treat all obstacles in a uniform, undifferentiated manner.

The current planner is rule-based and focuses on three fundamental agricultural tasks: harvesting, pruning, and thinning. These operations are common across a wide range of crops, so adapting the planner to new settings mainly involves adjusting the mapping between crop-specific attributes/relations and these core tasks. While action sequences or thresholds may differ by crop, the underlying decision framework remains applicable.

IV. Experiments and Results

A. Experimental Setup

Dataset: The *Cucumis_melo_300* dataset is crop-specific: all fruits are *Cucumis melo*. Across 300 greenhouse scenes, we annotated 514 fruits, 1,215 leaves, 341 stems, 115 flowers, and 434 pedicels. We also labeled 1,193 `occluded`, 434

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

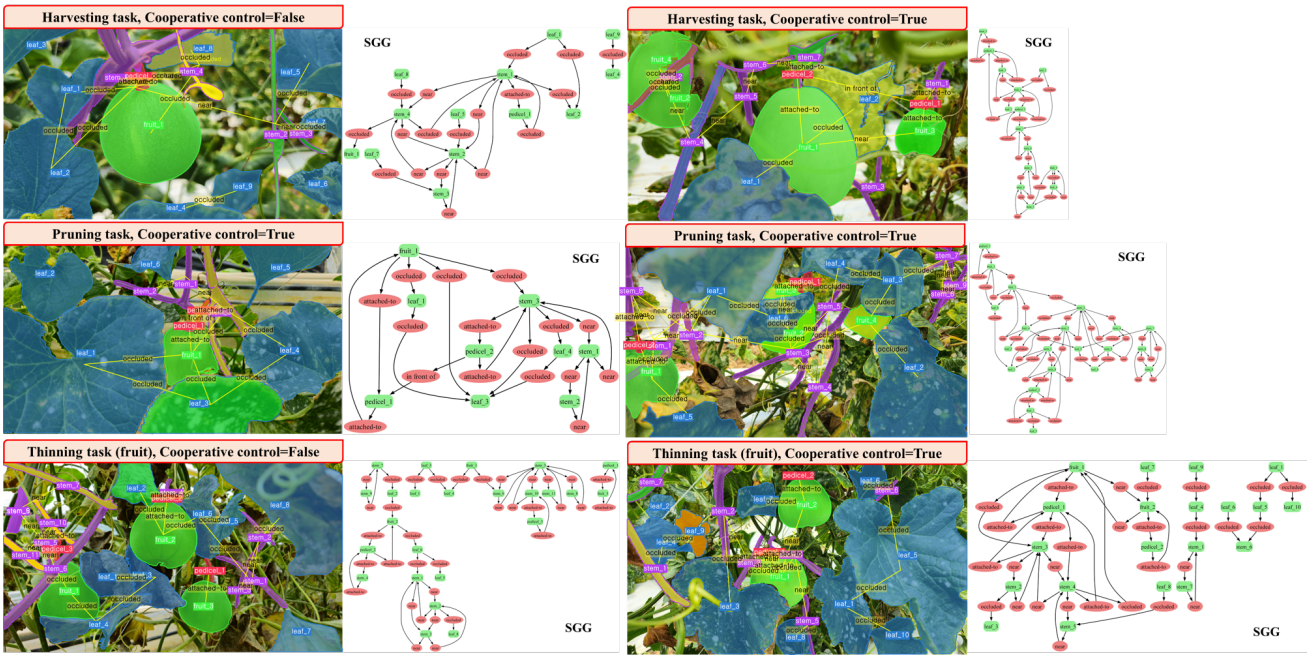

Fig. 6. Qualitative results of the proposed system. For harvesting, pruning, and thinning tasks, the left sub-panel shows perceptions with predicate, while the right sub-panel depicts the corresponding scene graph generated by SGG. The red header states the task inferred by the planner and whether cooperative dual-arm control is required.

`attached-to`, 709 `near`, and 145 `in-front-of` relations (Fig. 5, Table I).

For attributes, we distinguished between flexibility and rigidity based on functional roles. Leaves were labeled as *flexible* objects that can be displaced or removed. Fruits, pedicels, flowers, and stems were labeled as *rigid* objects, since they must be grasped or cut for harvesting or carefully avoided to minimize plant damage.

The dataset was collected in June under natural greenhouse conditions, with images captured in the morning (08:00–12:00) and afternoon (13:00–17:00) to reflect different lighting. Data collection was performed about one week before harvest, so both ripe and unripe fruits were present (237 ripe, 277 unripe).

Network configuration: Mask R-CNN (ResNet-50 + FPN, ImageNet pre-train) generates RoIs; RoI features feed a MotifNet [17] head for predicate classification and a three-layer MLP for attribute prediction. Training employed SGD for 30 epochs with early stopping on validation mR@50.

Metrics and protocol: Two stages are evaluated:

- Perception: Relationship R@50 / mR@50 and Attribute R@50 / mR@50.
- Planning: *Task–decision accuracy* T_{td}-correct selection of harvesting, pruning, thinning; *Cooperative control accuracy* T_{ccd}-correct single- vs. dual-arm choice.

Significance of pairwise differences was assessed with McNemar's χ^2 test ($\alpha = 0.05$). Three operating modes were considered:

1) Predicate classification (PredCls): Ground-truth object bounding boxes and classes are provided; only the relationships are predicted.
2) Scene graph classification (SGCls): Ground-truth bounding boxes are provided. However, the model must predict object classes and their relationships.
3) Scene graph detection (SGDet): Object classes and bounding boxes are not provided; the model must simultaneously detect and classify objects and their relationships.

Ground-truth for task planning: Because no public benchmark couples *scene–graph labels* with *high-level agricultural plans*, ground-truth (GT) required for evaluating the planner was synthesised in two steps:

- Each test image was exhaustively annotated with the object, attribute, and predicate classes of Table I; the result is a *human–verified* scene graph S_i^{GT}.
- The proposed task–planning $\pi(\cdot)$ was executed *once* on every S_i^{GT}; the outputs—task type $\tau_i^{GT} \in \{Harvesting, Pruning, Thinning\}$ and cooperative-control flag—were frozen and treated as reference labels.

Hence, the GT reflects *how the planner ought to behave* under perfect perception.

Task planning accuracy: Let S_i^P denote the *predicted* scene graph produced by the perception stack for image i. Feeding S_i^P into the same policy yields $\tau_i^P = \pi_\tau(S_i^P)$ and $c_i^P = \pi_c(S_i^P)$. Task decision and cooperative–control accuracies are then,

$$T_{td} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}\big[\tau_i^{\mathrm{P}} = \tau_i^{\mathrm{GT}}\big], \qquad T_{ccd} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}\big[c_i^{\mathrm{P}} = c_i^{\mathrm{GT}}\big], \tag{14}$$

where $N = 90$ is the number of test images and $\mathbf{1}[\cdot]$ is the indicator function. In words, accuracy is the proportion of scenes for which the planner, when driven by *noisy* visual

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

TABLE II
SGG PERFORMANCE AND RESULTING TASK PLANNING ACCURACY

Mode	OD	OCA	SGG Performance (%)				Task Planning (%)	
			Rel R@50	Rel mR@50	Attr R@50	Attr mR@50	T_{td}	T_{ccd}
PredCls	GT	GT	68.3	39.7	GT	GT	85.0	77.3
SGCls	GT	90.4	48.8	30.2	GT	GT	78.7	66.0
SGDet	71.4	80.9	38.9	24.5	70.1	52.2	72.3	53.7

OD: object detection, OCA: object class accuracy, Rel: relationship, Attr: attribute

input, reaches the same decision it would have made given the perfect (annotated) graph.

This protocol disentangles *planning logic* from *perception noise*: the policy π and its thresholds are fixed—evaluation changes only the quality of its input; any discrepancy between (τ_i^P, c_i^P) and (τ_i^{GT}, c_i^{GT}) is therefore attributable to errors in detection, attributes, or predicates, enabling a clean perception–planning error attribution. Both metrics of (14) are reported in Table II (Fig. 6).

B. Results and Discussion

Table II reports SGG performance and task-planning accuracy under three standard settings: PredCls, SGCls, and SGDet. These represent increasing difficulty, from using ground-truth boxes and classes (PredCls) to fully predicted detection and classification (SGDet).

End-to-end performance: With SGDet (all components predicted), the pipeline achieved $T_{td} = 72.3\%$ task decision accuracy and $T_{ccd} = 53.7\%$ cooperative control accuracy on 90 test images. Compared to a proximity-only baseline, SGDet improved task decisions by +13.6%p (McNemar $\chi^2 = 12.4$, $p < 0.01$). This confirms that adding relational and attribute cues provides measurable planning benefits.

Sensitivity to perception quality: Moving from PredCls to SGDet caused a 12.7%p drop in T_{td} and a larger 23.6%p drop in T_{ccd}. This shows that task-type selection is relatively robust, while cooperative dual-arm choice is almost twice as sensitive to errors in relation and attribute prediction. In practice, errors in occlusion detection or rigidity labeling directly determine whether the robot engages both arms, leading to larger downstream impact.

Role of detection quality: Attribute R@50 (70.1%) closely follows detection accuracy (71.4%). This indicates that attribute recognition is bounded by object detection quality. Thus, improving detectors tuned for horticultural scenes (e.g., for thin stems or clustered fruits) is the most effective way to boost planning accuracy.

Error analysis: Task errors (T_{td}) mainly arose from ripeness misclassification and false `near` relations in dense foliage. Cooperative-control errors (T_{ccd}) were dominated by occlusion mistakes, rigidity misclassification in reflective leaves, and missed thin stems or pedicels. These led the planner to predict single-arm use when dual-arm action was actually needed. This highlights that cooperative control requires more reliable relational cues than task decision.

Attribute analysis: Recall showed a long-tail distribution. Frequent attributes (*flexible*, *ripe*) achieved higher recall than rare ones (*rigid*, *unripe*). This explains the drop from Attr R@50 to Attr mR@50. Since attributes are tied to object type, mislabels in rigidity or ripeness directly affect both task type and arm coordination. Applying class-balancing (loss re-weighting, oversampling) improved mR@50 with minor precision trade-offs, suggesting that even simple rebalancing can yield tangible benefits.

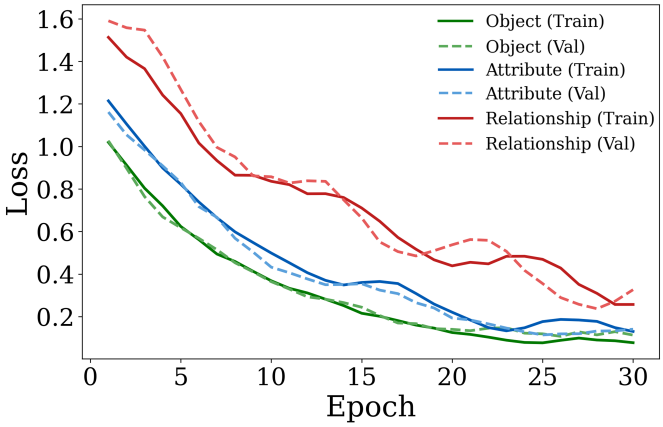

Fig. 7. Training/validation losses for object detection, attribute classification, and relationship prediction.

Training/validation losses: Fig. 7 shows all training losses. Detection and attribute losses converged smoothly with small generalization gaps. Relationship loss was slower and noisier, with a slight rise in later epochs, consistent with the higher difficulty of relation prediction and mild overfitting on the modest dataset. Attribute loss was stabilized using valid-pair masking to suppress label noise.

V. Conclusion and Future Work

This paper introduces a system integrating SGG with object-level and relational information for VSU–based task planning in agricultural robots. By modeling spatial and semantic relationships between crops, leaves, and stems, the system demonstrated enhanced recognition of attributes and relationships in the *Cucumis_melo_300* dataset, improving critical tasks, such as harvesting, pruning, and thinning.

The primary contribution of this work is a VSU-based task planning for agricultural environments that models both spatial relationships and physical properties, enabling robots to make more intelligent decisions about task selection and execution. The experimental results demonstrate that even with moderate SGG performance, the system achieves 72.3% task type accuracy and 53.7% cooperative control accuracy,

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

confirming the viability of this approach for agricultural robots perception and planning. Particularly significant is the system's ability to distinguish between rigid obstacles that must be avoided and flexible obstacles that can be actively managed, representing an advance over conventional approaches that treat all obstacles uniformly.

Future work should focus on several key areas for improvement. Object detection serves as the foundation for both relationship and attribute recognition, making it a priority target for enhancement. Relationship recognition, with the lowest performance metrics, could benefit from transformer-based architectures. Incorporating temporal information and multimodal sensing (e.g., depth or thermal) could also improve robustness to occlusions and attribute inference.

In addition, task planning can be extended. While this study considered harvesting, pruning, and thinning regardless of season, a phenology gate could constrain stage-specific tasks when growth-stage information is available. In practice, one versatile robot may need to handle multiple operations in resource-limited or extreme environments. Future extensions will also include tasks such as sampling and pest control.

ACKNOWLEDGMENTS

This work was supported by Korea Institute of Planning and Evaluation for Technology in Food, Agriculture and Forestry(IPET) through Open-field Smart Agriculture Utilization Model Development Program, funded by Ministry of Agriculture, Food and Rural Affairs(MAFRA)(RS-2025-02307274)

REFERENCES

[1] J. Kim and H. I. Son, "A voronoi diagram-based workspace partition for weak cooperation of multi-robot system in orchard," *IEEE Access*, vol. 8, pp. 20 676–20 686, 2020.

[2] C. Ju, J. Kim, J. Seol, and H. I. Son, "A review on multirobot systems in agriculture," *Computers and Electronics in Agriculture*, vol. 202, p. 107336, 2022.

[3] Y. Park, J. Seol, J. Pak, Y. Jo, C. Kim, and H. I. Son, "Human-centered approach for an efficient cucumber harvesting robot system: Harvest ordering, visual servoing, and end-effector," *Computers and Electronics in Agriculture*, vol. 212, p. 108116, 2023.

[4] E. S. Mohamed, A. Belal, S. K. Abd-Elmabod, M. A. El-Shirbeny, A. Gad, and M. B. Zahran, "Smart farming for improving agricultural management," *The Egyptian Journal of Remote Sensing and Space Science*, vol. 24, no. 3, pp. 971–981, 2021.

[5] V. Rajendran, B. Debnath, S. Mghames, W. Mandil, S. Parsa, S. Parsons, and A. Ghalamzan-E, "Towards autonomous selective harvesting: A review of robot perception, robot design, motion planning and control," *Journal of Field Robotics*, vol. 41, no. 7, pp. 2247–2279, 2024.

[6] H. Zhou, X. Wang, W. Au, H. Kang, and C. Chen, "Intelligent robots for fruit harvesting: Recent developments and future challenges," *Precision Agriculture*, vol. 23, no. 5, pp. 1856–1907, 2022.

[7] Y. Park, J. Seol, J. Pak, Y. Jo, J. Jun, and H. I. Son, "A novel end-effector for a fruit and vegetable harvesting robot: mechanism and field experiment," *Precision Agriculture*, vol. 24, no. 3, pp. 948–970, 2023.

[8] Y. Li, S. Wu, L. He, J. Tong, R. Zhao, J. Jia, J. Chen, and C. Wu, "Development and field evaluation of a robotic harvesting system for plucking high-quality tea," *Computers and Electronics in Agriculture*, vol. 206, p. 107659, 2023.

[9] J. Jun, J. Kim, J. Seol, J. Kim, and H. I. Son, "Towards an efficient tomato harvesting robot: 3d perception, manipulation, and end-effector," *IEEE access*, vol. 9, pp. 17 631–17 640, 2021.

[10] H. Williams, D. Smith, J. Shahabi, and et al., "Modelling wine grapevines for autonomous robotic cane pruning," *Biosystems Engineering*, vol. 235, pp. 31–49, 2023.

[11] M. Hussain, L. He, J. Schupp, D. Lyons, and P. Heinemann, "Green fruit segmentation and orientation estimation for robotic green fruit thinning of apples," *Computers and Electronics in Agriculture*, vol. 207, p. 107734, 2023.

[12] S. Sui, M. Li, Z. Li, Y. Zhao, C. Wang, W. Du, X. Li, and P. Liu, "A comb-type end-effector for inflorescence thinning of table grapes," *Computers and Electronics in Agriculture*, vol. 217, p. 108607, 2024.

[13] L.-E. Montoya-Cavero, R. D. de León Torres, A. Gómez-Espinosa, and J. A. E. Cabello, "Vision systems for harvesting robots: Produce detection and localization," *Computers and electronics in agriculture*, vol. 192, p. 106562, 2022.

[14] F. Gao, W. Fang, X. Sun, Z. Wu, G. Zhao, G. Li, R. Li, L. Fu, and Q. Zhang, "A novel apple fruit detection and counting methodology based on deep learning and trunk tracking in modern orchard," *Computers and Electronics in Agriculture*, vol. 197, p. 107000, 2022.

[15] L. Gong, W. Wang, T. Wang, and C. Liu, "Robotic harvesting of the occluded fruits with a precise shape and position reconstruction approach," *Journal of Field Robotics*, vol. 39, no. 1, pp. 69–84, 2022.

[16] J. Kim, H. Pyo, I. Jang, J. Kang, B. Ju, and K. Ko, "Tomato harvesting robotic system based on deep-tomatos: Deep learning network using transformation loss for 6d pose estimation of maturity classified tomatoes with side-stem," *Computers and Electronics in Agriculture*, vol. 201, p. 107300, 2022.

[17] Y. Cong, W. Liao, H. Ackermann, B. Rosenhahn, and M. Y. Yang, "Spatial-temporal transformer for dynamic scene graph generation," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 16 372–16 382.

[18] Z. Jiao, Y. Niu, Z. Zhang, S.-C. Zhu, Y. Zhu, and H. Liu, "Sequential manipulation planning on scene graph," in *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2022, pp. 8203–8210.

[19] Z. Ni, X. Deng, C. Tai, X. Zhu, Q. Xie, W. Huang, X. Wu, and L. Zeng, "Grid: Scene-graph-based instruction-driven robotic task planning," in *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2024, pp. 13 765–13 772.

[20] H. Jiang, B. Huang, R. Wu, Z. Li, S. Garg, H. Nayyeri, S. Wang, and Y. Li, "Roboexp: Action-conditioned scene graph via interactive exploration for robotic manipulation," *arXiv preprint arXiv:2402.15487*, 2024.

[21] O. M. Lawal, "Yolomuskmelon: quest for fruit detection speed and accuracy using deep learning," *IEEE Access*, vol. 9, pp. 15 221–15 227, 2021.

[22] T. Kim, D.-H. Lee, K.-C. Kim, and Y.-J. Kim, "2d pose estimation of multiple tomato fruit-bearing systems for robotic harvesting," *Computers and Electronics in Agriculture*, vol. 211, p. 108004, 2023.

[23] M. Han, Z. Zhang, Z. Jiao, X. Xie, Y. Zhu, S.-C. Zhu, and H. Liu, "Reconstructing interactive 3d scenes by panoptic mapping and cad model alignments," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2021, pp. 12 199–12 206.

[24] G. Chalvatzaki, A. Younes, D. Nandha, A. T. Le, L. F. Ribeiro, and I. Gurevych, "Learning to reason over scene graphs: a case study of finetuning gpt-2 into a robot language model for grounded task planning," *Frontiers in Robotics and AI*, vol. 10, p. 1221739, 2023.

[25] C. Agia, K. M. Jatavallabhula, M. Khodeir, O. Miksik, V. Vineet, M. Mukadam, L. Paull, and F. Shkurti, "Taskography: Evaluating robot task planning over large 3d scene graphs," in *Conference on Robot Learning*. PMLR, 2022, pp. 46–58.

[26] D. Ekpo, M. Levy, S. Suri, C. Huynh, and A. Shrivastava, "Verigraph: Scene graphs for execution verifiable robot planning," *arXiv preprint arXiv:2411.10446*, 2024.

[27] H. Fu, C. Liu, and X. Li, "Dynamic task planning: An integrated approach with scene relation perception and knowledge graphs," in *2024 36th Chinese Control and Decision Conference (CCDC)*. IEEE, 2024, pp. 1539–1543.

[28] T. Li, F. Xie, Q. Qiu, and Q. Feng, "Multi-arm robot task planning for fruit harvesting using multi-agent reinforcement learning," in *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2023, pp. 4176–4183.

[29] K. Rana, J. Haviland, S. Garg, J. Abou-Chakra, I. Reid, and N. Suenderhauf, "Sayplan: Grounding large language models using 3d scene graphs for scalable robot task planning," *arXiv preprint arXiv:2307.06135*, 2023, 10.48550/arXiv.2307.06135.

[30] Y. Park, C. Kim, and H. I. Son, "Fast and stable pedicel detection for robust visual servoing to harvest shaking fruits," *Computers and Electronics in Agriculture*, vol. 220, p. 108863, 2024.