



Article

A Multi-Task Model for Quality Recognition of Seedling Transplantation

Xiao Han ^{1,2,3}, Huarui Wu ^{1,2,3} and Wang Guo ^{1,2,3,*}

¹ National Engineering Research Center for Information Technology in Agriculture, Beijing 100097, China; hxyu33@sina.com (X.H.); wuhr007@nercita.org.cn (H.W.)

² Beijing Research Center for Information Technology in Agriculture, Beijing Academy of Agriculture and Forestry Sciences, Beijing 100097, China

³ Laboratory of Digital Village Technology, Ministry of Agriculture and Rural Affairs, Beijing 100097, China

* Correspondence: guow007@nercita.org.cn

Abstract

The mechanical characteristics of duckbill seeders, combined with complex environmental disturbances in the field, often lead to quality problems such as exposed seedlings and stem-buried seedlings when mechanically transplanted. These issues directly affect the subsequent yield and efficiency of automated harvesting. This paper is based on the YOLOv11 architecture and studies an automatic recognition model (Yolov11-ARM) for the transplanting quality of cabbage seedlings. The model applies a custom CPNMViTBv3 backbone that integrates Transformer-based global context modeling with efficient CNN feature reuse, replacing the standard C3f module to more effectively capture structural details in seedlings. The detection head is augmented with multi-scale feature aggregation and a lightweight SENetV2 attention mechanism, enhancing discriminative capability for small seedling instances. Additionally, a dynamically adaptive loss function for keypoint estimation is introduced, which adjusts according to object scale and error distribution, thereby ensuring training stability and robust localization performance. Experimental results show a high mean Average Precision (mAP@0.5) of 98.7% for seedling status (object) detection and 99.5% for keypoint localization with the studied model. This research provides an effective technical solution for the real-time, multi-parameter quality assessment of automated transplanting.

Keywords: Yolov11-ARM; cabbage seedling; keypoint detection; multi-task model; quality recognition



Academic Editor: Zhengjun Qiu

Received: 17 December 2025

Revised: 1 February 2026

Accepted: 4 February 2026

Published: 1 August 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

Plug tray seedling transplanting is widely used in commercial vegetable production [1]. In this system, seedlings are cultivated in standardized plug trays until an appropriate growth stage, after which they are transplanted into the field. Manual transplanting, however, suffers from low efficiency, high labor intensity, and rising labor costs, making the transition to intelligent transplanting technologies an inevitable trend [2,3]. Mechanized equipment alleviates labor demands while improving planting efficiency. Nevertheless, mechanical seedling pickup and planting could lead to operational failures such as clamping, clogging, and dropping, resulting in quality defects, including seedling injury, exposure, improper burial, or lodging. Since transplanting quality directly affects crop yield, integrating real-time quality monitoring modules into automated transplanters provides essential data for advancing intelligent machinery, enabling adaptive adjustments and

process optimization. During mechanical transplanting, the motion of the equipment and its forward speed frequently cause seedling tilting due to traction from the duckbill planter, compromising verticality [4,5].

Advances in machine vision have spurred the application of intelligent recognition algorithms in real-time detection, enabling the identification of seedling stems, leaves, and other key agricultural features to improve transplanting quality. Current research focuses on two phases: pre-transplant dynamic assessment of seedling viability [6–15] and post-transplant monitoring of seedling status [16–19]. Pre-transplant efforts emphasize empty plug detection and lodging identification. Li et al. developed an adaptive transplanting system incorporating an asymmetric light-field-coupled visual sensing module that effectively suppresses ambient light interference, enabling accurate removal of weak, missing, or damaged seedlings before transplanting [9]. Chen et al. analyzed vibration signals from transplanting manipulators and image features of plug seedlings to interpret real-time vibration spectral characteristics, mitigating the impact of mechanical resonance on transplanting trajectories. Their vision-based health screening achieved over 96.10% accuracy in identifying healthy seedlings [10]. Li et al. integrated a deep learning-driven selective transplanting robot that combines seedling selection, sorting, and transplanting mechanisms [11]. Zhao et al. measured collision contact areas between seedlings and soil using sensors and conducted single-factor experiments to analyze the effects of seedling age and operational speed on collision stress distribution and damage mechanisms [12]. Jin et al. optimized transplanting paths through seedling recognition, reducing stem and leaf damage by 11.11%, albeit with a slight increase in per-cycle time (0.029 s) [13]. Zhao et al. applied YOLOv5 to classify empty plugs and weak, damaged, and healthy seedlings in pepper and tomato plug trays [14]. Xue et al. developed a computational fluidic model for rice [15].

These studies primarily utilize machine vision for recognizing seedling status and enabling selective transplanting to improve quality, or they refine transplanting mechanisms to minimize seedling damage. In addition, for seedling phenotype recognition, most existing algorithms use 3D reconstruction to obtain seedling phenotype parameters [16–18]. While providing depth information, these 3D point cloud methods typically involve higher computational complexity, greater hardware cost (e.g., multi-camera or LiDAR setups), and longer processing latency. For the task of real-time posture monitoring on moving transplanters—where low latency, modest computational footprint, and cost-effectiveness are critical—a well-designed 2D vision-based approach presents a more practical and deployable solution. This is particularly relevant for embedded systems integrated into agricultural machinery.

In post-transplant seedling status detection, Wen [19], Na [20], and Wu [21] improved verticality through mechanical structural optimization. Wu et al. developed a system to classify buried, exposed, and healthy seedlings. Jin et al. designed an online monitoring system integrating machine vision and deep learning [22]. By combining multi-sensor modules with the YOLO-RDS adaptive rotation detection algorithm, the system could identify 3D spatial orientation of pepper seedling verticality. Experimental results demonstrated 92.56% detection accuracy with absolute errors below 3°, providing reliable technical support for intelligent transplanting quality assessment.

Current research on transplanting quality definition and parameter extraction faces two major limitations:

- (1) Limited detection scope: Most computer vision studies focus on isolated metrics such as verticality, neglecting multifactorial quality assessments—including root-naked, missed, and lodged transplants—in crops such as cabbage.

- (2) **Incomplete feature utilization:** Existing object detection methods do not support comprehensive seedling posture analysis. Further investigation is needed to enhance detection outputs for robust feature extraction.

Correspondingly, the work presented in this paper is organized around the following sequence:

1. **Task Formulation:** We define a unified real-time vision task that combines multiclass seedling status detection (normal, exposed, stem-buried) with precise stem keypoint localization for posture quantification.
2. **Model Design:** To meet the demands of this task, we introduce YOLOv11-ARM, an enhanced multi-task detection model. Its design incorporates a hybrid CNN–Transformer backbone for more comprehensive feature representation, a multi-scale detection head augmented with attention mechanisms, and a dedicated loss function for improved keypoint regression.
3. **Implementation & Evaluation:** The studied model is trained and evaluated on a dedicated cabbage seedling dataset. We assess its performance in terms of accuracy and efficiency, with the aim of providing a practical and extensible framework for automated transplant quality assessment.

Therefore, this paper presents a multi-task real-time detection model (YOLOv11-ARM) based on an enhanced YOLOv11 framework (You Only Look Once v11, YOLOv11) [23]. The model identifies three transplanting postures (exposed, stem-buried, and normal seedlings) and precisely locates stem keypoints. Through the integration of a hybrid backbone for enriched feature representation, a multi-scale aggregation detection head with attention mechanisms, and a dynamic adaptive loss function, the developed model achieves high-precision and efficient detection of seedling status and structure in complex field environments. This high detection accuracy directly supports the automation loop by providing reliable input for real-time decision-making. In practical deployment, such a system can significantly reduce the need for manual inspection and corrective labor, lower the rate of transplanting failures that necessitate costly re-planting, and ultimately contribute to higher operational efficiency and consistency in mechanized transplanting lines. This work provides a practical technical solution for real-time quality monitoring during mechanical transplanting, contributing to the intelligent advancement of transplanting equipment. Furthermore, by enabling accurate keypoint detection, the model facilitates the quantitative extraction of crop phenotypic traits, establishing a foundation for subsequent growth monitoring and agronomic analysis.

2. Materials and Methods

2.1. Data Acquisition

Cabbage seedling images were collected at Taizhouwu Agricultural Base in Haidian District, Beijing (116°46' E, 40°18' N). The transplanted cabbage cultivar was ZhongGan 21, with 5–7 true leaves at transplanting. The images were acquired under varying lighting conditions to improve model generalization, including both sunny and overcast days. As shown in Figure 1, to accurately identify transplanting postures, datasets were captured perpendicular to the planting direction, with a 30° angle between the camera sightline and ground horizon, at 30–40 cm from seedlings. Images were captured using an iPhone 13 Pro Max smartphone (Apple Inc., Cupertino, CA, USA), and its camera provides sufficient resolution and clarity to distinguish seedling stem features, meeting the core requirement of this study. However, in real-world applications, the deployment is more flexible. The algorithm is designed to be robust to variations in shooting height and angle, provided that the cabbage stem remains clearly visible in the image. This ensures the method can

be adapted to different field setups and equipment configurations without significant performance degradation.

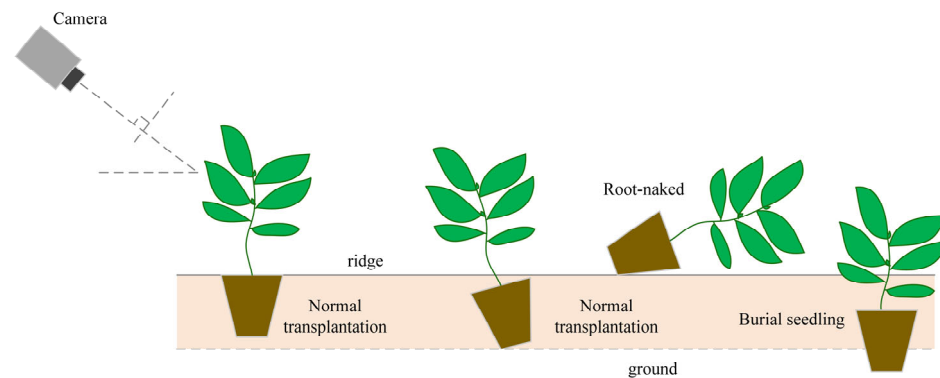


Figure 1. Schematic diagram of data collection method for transplanted seedlings.

Image categories included exposed seedlings, stem-buried seedlings, and normally transplanted seedlings. The class distribution was set at a ratio of approximately 3:1:1. Additional images of normally transplanted seedlings were collected to support more detailed feature learning, as keypoint detection in such seedlings is more challenging. A total of 466 RGB images were collected during transplanting. Python 3.8 scripts cropped raw images to minimize interference from complex backgrounds.

2.2. Postural Keypoint Analysis of Transplanted Cabbages

As shown in Figure 2, this paper uses Labelme (<https://github.com/wkentaro/labelme>, accessed on 15 January 2026) to annotate the bounding boxes and keypoints of cabbage seedlings. The model is designed to classify seedlings into three postural categories: exposed seedlings, stem-buried seedlings, and normally transplanted seedlings. The dataset comprises a total of 466 RGB images, which were expanded through augmentation. The original images across these three categories were collected in an approximate ratio of 4:1:1, corresponding to roughly 321, 58, and 87 images for normally transplanted, root-naked, and burial seedlings, respectively. To support this classification and enable precise posture analysis, four keypoints were manually annotated along the visible stem of each normally transplanted seedling. This specific design is based on a balance between morphological fidelity and practical annotation efficiency:

- (1) Number of Points: Fewer than four points proved insufficient to reliably capture the natural curvature of the stem for robust skeleton fitting, while more than four would significantly increase manual annotation burden without a commensurate improvement in posture discrimination for our task. Four points optimally support fitting a smooth, representative skeletal curve.
- (2) Location Rationale: The points are strategically placed at morphologically significant positions: the first point at the stem-soil interface (base), the last point at the apical meristem (growth tip), and the two intermediate points at approximately one-third and two-thirds of the visible stem length. This distribution ensures the annotated skeleton effectively represents the overall stem orientation and any critical bending, which are direct indicators for distinguishing among the three transplanting postures.

In addition, a bounding box was drawn to localize each seedling instance.

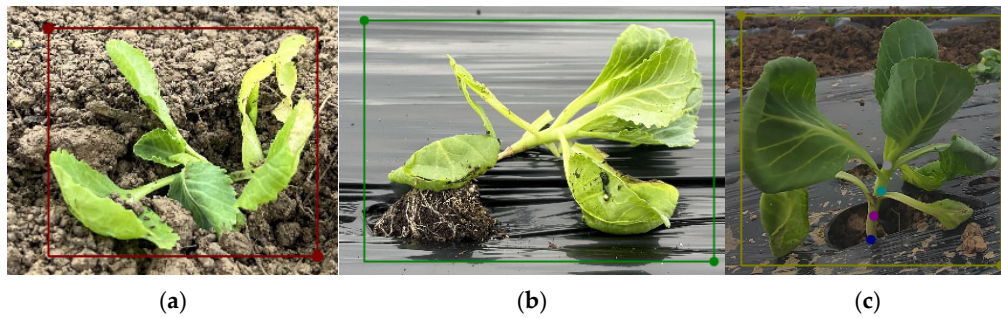


Figure 2. Annotated data illustration: (a) Burial seeding; (b) Root-naked; (c) Keypoints of stem (Four colored dots (from bottom to top) represent the key points used for fitting the seedling stem).

2.3. Data Preprocessing

All seedlings were annotated using labelme. Bounding boxes were drawn for each seedling, and four keypoints were marked along the stem—a design chosen to balance between accurate skeleton fitting and real-time inference performance. The original dataset of 466 images provides a sufficient basis to learn the target postures, as the seedling phenotypes of different cabbage varieties are similar at the transplanting stage. Data augmentation techniques, including horizontal flipping, color transformation, random blurring, and contrast adjustment, were applied to the original images. This strategy artificially expands the data distribution, introducing variability in lighting, texture, and focus that mimics real-world conditions, thereby directly mitigating the risk of overfitting to the limited original samples. This process expanded the total dataset to 2562 images. This complete set of augmented images was then randomly divided into training, validation, and test sets according to an 8:1:1 ratio, resulting in 2050, 256, and 256 images, respectively. To standardize the input size, all images were resized to 384×288 .

The images were collected under varying lighting conditions to improve generalization. However, the dataset primarily represents the specific field conditions of this study. Performance in environments with significantly different soil types or extreme weather remains to be validated in future work.

3. A Posture Detection Model for Transplanted Seedlings

In order to accurately identify the status of transplanted seedlings, this paper constructs a multi-task recognition model. This model can distinguish between exposed seedlings, stem-buried seedlings, and normally transplanted seedlings, while locating key points for transplanted seedlings.

YOLOv11-Pose, a single-stage keypoint detection algorithm originally developed by Glenn-Jocher for human pose estimation, comprises input, backbone, neck, and detection head modules. When adapted to seedling posture recognition, the model faces challenges including highly clustered stem keypoint distributions, significant phenotypic discrepancies between human and seedling structures, and complex lighting variations in agricultural environments. As shown in Figure 3, this study introduces targeted modifications (Yolov11-ARM) to the YOLOv11n-Pose baseline framework and studies an automatic recognition model.

To better capture both local details and global context of seedling stems, we construct the CPNMViTBv3 module (Figure 4). It integrates Transformer-based global modeling within a Cross-Stage Partial (CSP) framework, improving feature flow and representation for delicate plant structures.

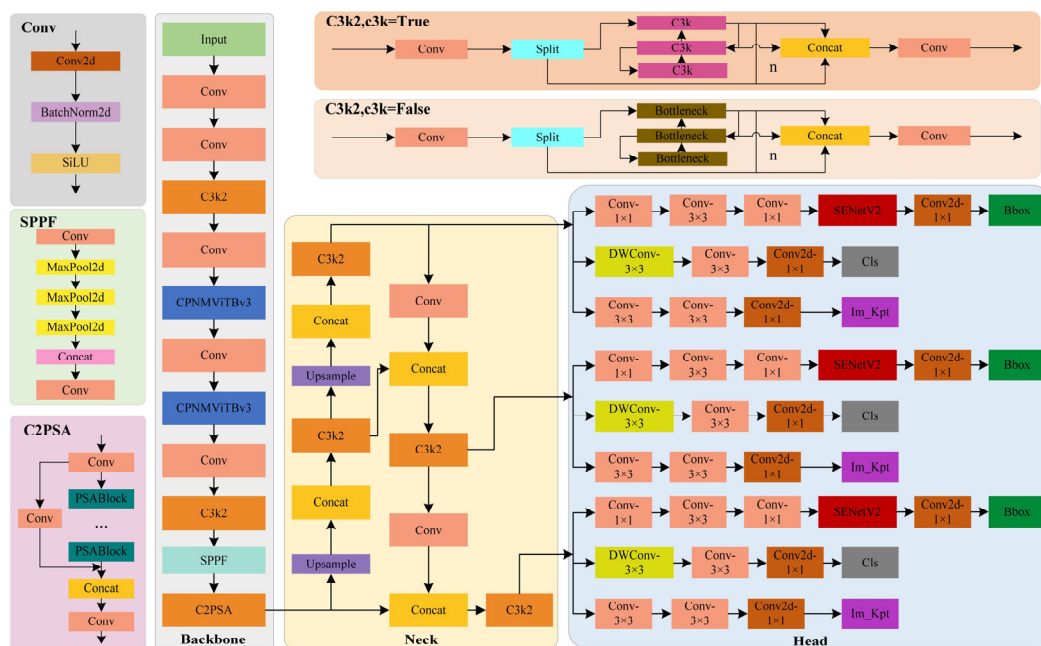


Figure 3. Yolov11-ARM architecture.

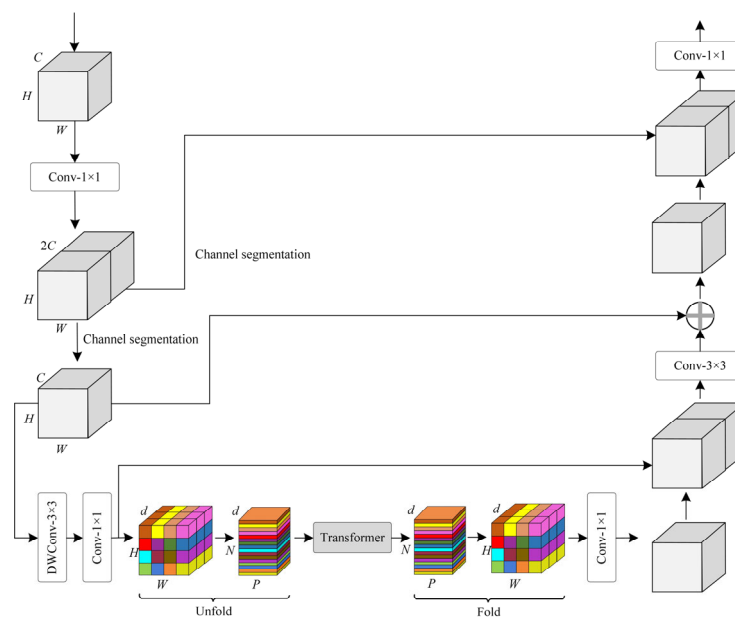


Figure 4. CPNMViTBv3 module.

In addition, we substantially modify the original detection head (Figure 5). The improved head introduces a SENetV2 attention module (whose structure is detailed in Figure 6) for dynamic feature recalibration and employs a multi-scale feature fusion strategy. Furthermore, it incorporates a novel keypoint loss function (detailed in Section 2.3) tailored for seedling stem fitting.

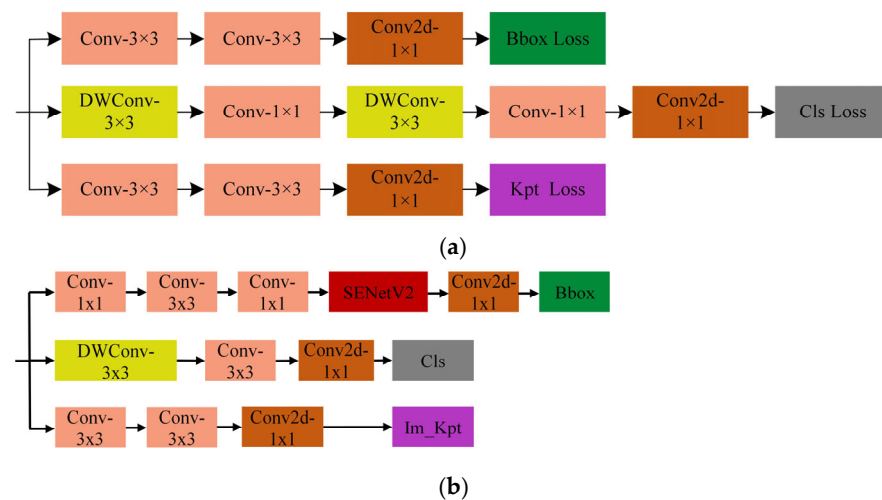


Figure 5. YOLOv11-pose detection head. (a) YOLOv11-pose detection head; (b) Scale aggregation detection head.

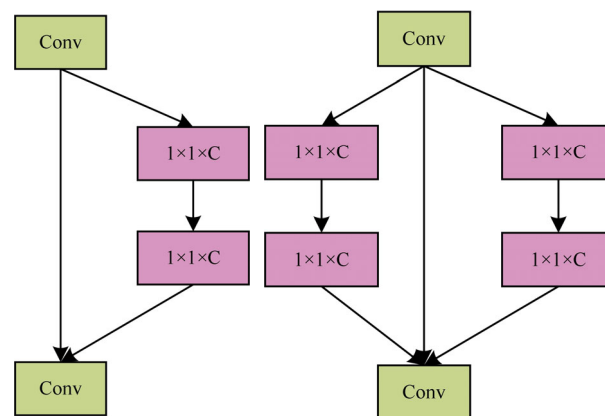


Figure 6. SENetV2 module.

3.1. Construction of CPNMViTBv3 Module

The C3K2 (where C3 inherits from the traditional Cross-Stage-Partial connections structure, K denotes its flexible Kernel strategy, and 2 indicates a specific configuration variant) module, a key enhanced feature fusion architecture in YOLOv11, optimizes the traditional C3 (Cross-Stage-Partial connections (CSP) with 3 convolutions) structure [24] by employing a flexible convolutional kernel strategy. It utilizes smaller kernels in shallow layers to capture fine-grained details and larger kernels in deeper layers to integrate broader semantic information. While this design is effective for hierarchical local feature fusion, its purely convolutional nature imposes a fundamental constraint on capturing long-range contextual dependencies, which are crucial for understanding complex scenes with occlusions or diverse postures.

To complement local feature extraction with structured global context modeling, this study proposes a novel module, the CSP-Enhanced Mobile Vision Transformer V3 (MobileViTv3) Block (CPNMViTBv3), illustrated in Figure 4. The module integrates a Transformer-based branch (inspired by MobileViTv3) for global relation modeling alongside a concurrent local feature stream via a CSP structure, enabling a balance between local precision and global awareness.

The YOLOv11-Pose backbone network employs CPNMViTBv3 modules to extract hierarchical features from the 4th (mid-level) and 6th (deep-level) layers, enhancing feature representation and information propagation efficiency. CPNMViTBv3 adopts a modular design with cross-layer connections to strengthen feature flow, enabling direct interaction

between low-level and high-level features while minimizing redundant computation. As shown in Equation (1), A 1×1 convolution first expands input feature channels, followed by partitioning into two streams:

$$X' = \text{Conv}_{1 \times 1}(X) \text{ where } X' \in R^{H \times W \times (2C)} \quad (1)$$

Here, X is the input feature map, and X' is the output with doubled channels. H , W , C denote height, width, and channel number, respectively. One stream processes deep features through n MobileViTv3 blocks [25], while the other retains original features via residual connections. MobileViTv3 integrates multi-head self-attention mechanisms and feed-forward neural networks to capture global contextual dependencies and refine feature representations. For input features Y , the MobileViTv3 transformation is defined as Equation (2):

$$Y' = \text{MobileViTv3}(Y) = \text{Attention}(Y) + \text{FeedForward}(Y) \quad (2)$$

where $\text{FeedForward}(\cdot)$ denotes a position-wise feed-forward network applied to each token independently. $\text{Attention}(Y)$ computes multi-head self-attention as shown in Equation (3):

$$\text{Attention}(Y) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (3)$$

Q , K , and V are the Query, Key, and Value matrices, respectively, $Q, K, V \in R^{N \times d_k}$, where $N = H \times W$ is the number of spatial tokens and d_k is the key dimension. After being processed by MobileViTv3, the feature map is reshaped into patch form and fused with multi-scale features through Transformer:

During the inference process, CPNMViTBv3 enables the model to capture relevant features in complex visual tasks better while maintaining efficiency, promoting semantic information integration and channel interaction. This module optimizes the global local collaborative expression ability of deep features in the backbone network without significantly increasing computational costs and is suitable for detecting scenarios with a set of key points.

3.2. Scale Aggregation Detection Head

The YOLOv11-Pose model conducts two innovations in detection head design: multi-task collaborative optimization and lightweight dynamic perception. An adaptive feature decoupling head separates object detection and keypoint prediction into parallel branches, employing dynamic channel reweighting to allocate task-specific feature weights.

The object detection and keypoint prediction are separated into parallel branches by adaptive feature decoupling head. It consists of three decoupling heads, namely prediction classification score and regression coordinates. Combined with the task alignment allocation mechanism, the classification and positioning performance are improved through the joint optimization of classification score and IoU, while the low-quality prediction frame is suppressed. At the same time, Yolov11 adopts depth-separable convolution to reduce the number of model parameters (as shown in Figure 5). However, the detection head only relies on single-scale high-level features, and it is easy to lose small object information. The fixed anchor frame design makes it difficult to match the diversity of target scales, resulting in high missing rate of small targets and inaccurate boundary regression.

As shown in Figure 5, this study optimizes the detection head through modular feature enhancement and attention mechanisms. Multi-level input features are upsampled and laterally connected to construct multi-scale feature maps integrating rich semantic and detailed information. Scale-specific anchor boxes dynamically adapt parameters across feature hierarchies to match object scale variations. Hierarchical prediction jointly balances

global localization for large targets and fine-grained recognition for small ones, elevating multi-scale detection accuracy.

The detection head incorporates a bottleneck architecture to streamline module parameters while integrating SENetV2 (Squeeze-and-Excitation Network V2 block) for dynamic adaptive weight generation [26]. This dual mechanism suppresses redundant information interference and strengthens model robustness in complex scenarios.

As shown in Figure 6, SENetV2 improves upon the SENet framework by integrating Squeeze-and-Aggregation Excitation (SaE) modules that combine multi-branch fully connected topologies with channel attention mechanisms. This fusion elevates global information integration and feature representation precision through dual strategies: multi-path feature interaction and dynamic channel calibration, with only minor parameter increases. Figure 6 contrasts SENet and SENetV2 architectures, emphasizing SENetV2's aggregated structure within residual blocks to enhance feature discriminability and global context modeling.

The SaE module operates through three stages: Global average pooling compresses input feature tensors into vectors, processed by four parallel dimension-reducing fully connected layers. Adaptive weight fusion generates channel attention maps, while branch features pass through nonlinear transformations. Dimensional restructuring aligns processed features spatially with original inputs. By establishing cross-channel long-range dependencies, this architecture dynamically recalibrates semantic feature weights, improving multi-scale object discrimination.

3.3. Loss Function

The basic loss function of YOLOv11 consists of four parts: classification loss, bounding box loss, keypoint loss, and keypoint confidence loss. The total training loss (L_{total}) is formulated as:

$$L_{total} = \sum_{s,i,j,k} (\lambda_{cls} L_{cls} + \lambda_{box} L_{box} + \lambda_{kpts} L_{kpts} + \lambda_{kp_conf} L_{kp_conf}) \quad (4)$$

where λ_{cls} , λ_{box} , λ_{kpts} , λ_{kp_conf} represents task-specific weighting factors. They dynamically balance the optimization focus between classification (L_{cls}), bounding-box regression (L_{box}), keypoint localization (L_{kpts}), and keypoint confidence tasks (L_{kp_conf}).

For classification, the loss function (L_{cls}) adopts Binary Cross-Entropy (BCE) to measure the difference between predicted probabilities p_i and ground truth labels y_i as shown in Equation (5):

$$L_{cls} = \frac{1}{N} \sum_i L_i = \frac{1}{N} \sum_i \left[\underbrace{-y_i \cdot \ln(p_i)}_{\text{When the truth label is 1}} + \underbrace{(1 - y_i) \cdot \ln(1 - p_i)}_{\text{When the truth label is 0}} \right] \quad (5)$$

Here, N represents the total sample count. The BCE loss quantifies prediction errors and drives parameter optimization through gradient descent and improves classification accuracy.

To more effectively identify the keypoint locations and their correlations in seedlings, this paper has improved the keypoint loss function L_{kpts} . Due to the gradient vanishing of traditional OKS loss when there is a large error in keypoint localization Our enhanced formulation integrates dynamic threshold adjustment and smooth gradient transition mechanisms for improved robustness. Let d_n represent the squared Euclidean distance between the n -th predicted keypoint (x_n^p, y_n^p) and ground truth (x_n^g, y_n^g) as follows:

$$d_n = (x_n^p - x_n^g)^2 + (y_n^p - y_n^g)^2 \quad (6)$$

This distance is then normalized by the object scale to ensure consistent error measurement across seedlings of different sizes. Considering the standard deviation of keypoints and the area of the bounding box, the larger the object area s , the larger the allowable absolute error should be. Thus, normalizing the error of the distance as follows:

$$e_{norm_n} = \frac{d_n}{\left(\frac{2}{N_{kpts}}\right)^2 \cdot s \cdot 2} \quad (7)$$

where N_{kpts} represents the total number of keypoints. Keypoint detection difficulty varies across samples due to occlusion or blurring in some seedlings. A dynamic threshold (Th) is computed based on the mean (μ_e) and standard deviation (σ_e) of the normalized errors within a training batch. This allows the loss to adaptively focus on samples of moderate difficulty. Therefore, based on the mean standard deviation calculation, this paper obtains Equations (8) and (9):

$$Th = \mu_e + \alpha \cdot \sigma_e \quad (8)$$

$$\mu_e = \frac{\sum_{n=1}^{N_{kpts}} e_{norm_n} \delta(v_n > 0)}{\sum_{n=1}^{N_{kpts}} \delta(v_n > 0)}, \sigma_e = \sqrt{\frac{\sum_{n=1}^{N_{kpts}} (e_{norm_n} \delta(v_n > 0) - \mu_e)^2}{\sum_{n=1}^{N_{kpts}} \delta(v_n > 0)}} \quad (9)$$

where α represents the hyperparameter controlling threshold strictness. $\delta(v_n > 0)$ represents the visibility flag of the key point. When $v_n > 0$, its value is 1; otherwise, it is 0. μ_e represents the mean of the normalized errors e_{norm_n} over the batch. σ_e represents standard deviation of the normalized errors e_{norm_n} over the batch. In the above equation, as the threshold approaches the mean loss, more losses will be classified as high error regions, accelerating convergence. When the threshold is further away from the loss mean, more losses are classified as low error areas, and the prediction is finely adjusted. Due to the difference in absolute position error magnitude between large and small objects. For example, the 10-pixel error of a large object may have a similar actual impact as the 5-pixel error of a small object.

Minor errors: Exponential decay loss enhances sensitivity to subtle deviations (steep gradients).

Major errors: Linear loss prevents gradient diminishment.

To ensure smooth gradient flow around this threshold, a sigmoid function is used for transition, leading to the final weighted loss. Meanwhile, in order to avoid gradient jumps at the threshold, sigmoid (t_n) is used for smooth transition as shown in Equation (10):

$$\begin{cases} t_n = \text{sigmoid}\left(\frac{e_{norm_n} - Th}{\sigma_e}\right) \\ loss_n = (1 - t_n) \cdot \underbrace{(1 - \exp(-e_{norm_n}))}_{e_{norm_n} < Th} + t_n \cdot \underbrace{(e_{norm_n} - Th + 1 - \exp(-Th))}_{e_{norm_n} \geq Th} \end{cases} \quad (10)$$

Finally, the weighted average loss is used to obtain Equation (11):

$$L_{kpts} = \frac{1}{\sum_{n=1}^{N_{kpts}} \delta(v_n > 0)} \sum_{n=1}^{N_{kpts}} \delta(v_n > 0) \cdot \left[\begin{array}{l} (1 - t_n) \cdot (1 - \exp(-e_{norm_n})) + \\ t_n \cdot (e_{norm_n} - Th + 1 - \exp(-Th)) \end{array} \right] \quad (11)$$

To evaluate the loss function of keypoint confidence, this paper uses BCE to measure the difference between predicted confidence and actual existence. As shown in

Equation (12), minimizing this loss can effectively improve the model's confidence (L_{kpt_conf}) in unobstructed points.

$$L_{kp_conf} = \sum_{n=1}^{N_{kpts}} BCE(p_{kpts_n}, \delta(v_n > 0)) \quad (12)$$

p_{kpts_n} is the confidence level of the n th keypoint.

The improved loss function utilizes a dynamic threshold mechanism to adaptively adjust the loss function based on sample difficulty, enhancing the robustness of the model to objects of different scales. In addition, using a smooth Sigmoid transition function avoids the gradient explosion or disappearance, making the training process more stable.

4. Experimental Setup

The model training learning rate is set to 0.01, the weight decay of the learning rate is 0.005, the border loss gain is 7.5, the classification loss gain is 0.5, and the attitude loss gain is 12.0. After 300 epochs of training, the model is evaluated for each round of training, and the optimal cycle is selected based on the mAP@0.95. The parameters of the experimental training and testing platform are shown in the Table 1 to ensure that the hardware meets the basic requirements for model training.

Table 1. Training environment configuration.

Configuration	Parameter
CPU	20 vCPU Intel (R) Xeon (R) Platinum 8457C
Memory	100 GB
GPU	L20
Video memory	48 GB
Training environment	Cuda 11.3
Operating system	ubuntu20.04
Development environment	PyTorch 1.10.0; Python 3.8

Select precision (P), recall (R), and mean Average Precision (mAP) as evaluation indicators:

$$Precision(P) = \frac{T_P}{T_P + F_P} \times 100\% \quad (13)$$

$$Recall(R) = \frac{T_P}{T_P + F_N} \times 100\% \quad (14)$$

$$mAP@0.95 = \frac{1}{10} \sum_{t=0.5}^{0.95} AP_t, \quad AP = \int_0^1 \rho(R) dR \quad (15)$$

Note: T_P refers to the number of positive samples that are accurately recognized as true positives; F_P refers to the number of positive samples identified as true positives but mistakenly classified as false positives; F_N refers to the number of negative samples not correctly recognized as false negatives. c is the number of categories. mAP@0.5 represents the accuracy value when the threshold of object similarity is 0.50. mAP@0.95 evaluates a variety of object similarity threshold values ranging from 0.5 to 0.95, with increments of 0.05 (for instance, 0.5, 0.55, 0.6, and so on, up to 0.95) [27].

Furthermore, this research uses GFLOPs and the total number of parameters as metrics to measure the model's size.

5. Experiment and Result Analysis

5.1. Comparative Experiments

To objectively show the effectiveness of the studied model, comparative experiments were conducted. These experiments included YOLOv7-Pose, YOLOv8-Pose, YOLOv11-Pose, MD-YOLO (i.e., Multi-scale Dense YOLO) [28] and DSW-YOLO (i.e., A detection method for ground-planted strawberry fruits) [29]. Among these, YOLOv7-Pose and YOLOv8-Pose represent traditional approaches to keypoint detection. The experimental results are presented in Table 2.

Table 2. Comparison of different models.

Models	Object Detection				Keypoint Detection				Model Size (MB)	GFLOPS
	Precision/%	Recall/%	mAP@0.5/%	mAP@0.95/%	Precision/%	Recall/%	mAP@0.5/%	mAP@0.95/%		
YOLOv7-Pose	80.5	85.4	87.1	71.4	--	-	-	-	76.25	101.6
YOLOv8-Pose	97.8	82.4	91.0	81.1	98.3	88.5	93.7	92.4	3.08	8.3
YOLOv11-Pose	99.5	93.9	96.7	85.1	98.5	98.2	98.3	98.0	2.65	6.6
MD-YOLO	99.2	92.2	95.7	86.6	95.7	91.7	94.2	92.3	3.10	8.3
DSW-YOLO	94.4	60.8	78.0	65.7	97.7	98.9	99.3	99.3	3.54	8.3
Yolov11-ARM	99.5	97.8	98.7	86.7	98.5	99.0	99.5	99.4	3.54	9.6

As shown in Tabel 2, Yolov11-ARM shows competitive performance in both object and keypoint detection tasks. For object detection, it records the highest recall (97.8%) and mAP@0.5 (98.7%), along with a strong mAP@0.95 (86.7%). In keypoint detection, the model also performs robustly, reaching 98.5% precision, 98.9% recall, 99.5% mAP@0.5, and 99.4% mAP@0.95—matching or exceeding all other compared methods. It surpasses YOLOv8-Pose by 7.7 percentage points in recall and 7.7 points in mAP@0.5 for object detection, while also performing better across all keypoint detection metrics. Compared to YOLOv11-Pose, our model gives higher Recall and mAP@0.5 in object detection, together with improved Recall and mAP@0.95 in keypoint detection. Although DSW -YOLO shows similar keypoint detection results, our model has significantly stronger object detection capability, with a 20.7% higher recall and 20.7% higher mAP@0.5. In terms of efficiency, the proposed model has a size of 3.54 MB and computational cost of 9.6 GFLOPS—slightly more than YOLOv11-Pose and MD-YOLO, but still maintains a compact and efficient structure. It keeps a clear real-time advantage over the much larger YOLOv7-Pose. These findings indicate that the studied model provides a good balance between detection performance and computational cost, showing potential for real-time seedling structure analysis and keypoint estimation in agricultural automation.

Figure 7 shows some visual detection results. In the keypoint detection task, the keypoint positions detected by YOLOv8 Pose, YOLOv11 Pose, and MD-YOLO all showed significant deviations. The confidence level obtained by YOLOv7 Pose is significantly lower than other algorithms, and there are still significant deviations in the detection positions of the four key points. The model size is too large, posing a challenge for deployment in agricultural scenes where high processing speed is needed.



(a)



(b)



(c)



(d)

Figure 7. Cont.

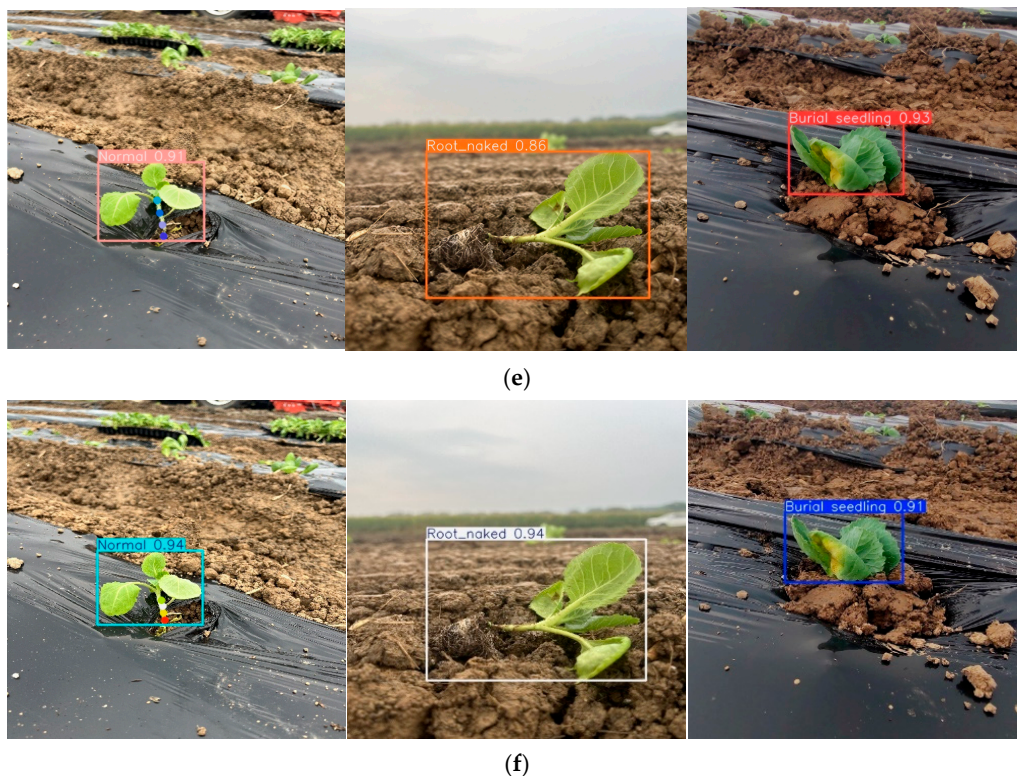


Figure 7. The detection performance of different algorithms under different seedling postures (The four colored dots (from bottom to top) in the leftmost image represent the key points used for fitting the seedling stem): (a) YOLOv7-Pose; (b) YOLOv8-Pose; (c) YOLOv11-Pose; (d) MD-YOLO; (e) DSW-YOLO; (f) YOLOv11-ARM.

5.2. Model Training Analysis and Ablation Experiment

Figure 8 presents the training dynamics, showing the evolution of loss and validation accuracy (mAP@0.95) over epochs. The loss curves for YOLOv11-ARM decrease smoothly and converge stably after approximately 150 epochs, indicating effective optimization without severe overfitting or underfitting. More importantly, the validation mAP@0.95 curves closely follow the training trends, reaching a plateau near the final performance (86.7% for object detection, 99.4% for keypoint detection). This close alignment between training and validation metrics demonstrates that the model is learning generalizable features.

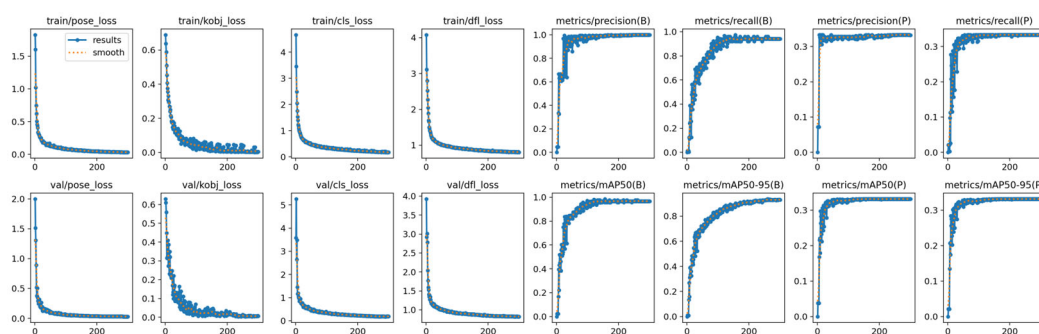


Figure 8. The training dynamics.

A fine-grained analysis of classification errors further confirms the robustness of YOLOv11-ARM. As shown in Figure 9, the normalized confusion matrix shows good discrimination (values of 1.00) for burial and normal seedlings, and for background. For root-naked seedlings, the accuracy is 0.94, with the only minor misclassification (0.06) being as background, indicating no harmful confusion between the three target posture classes.

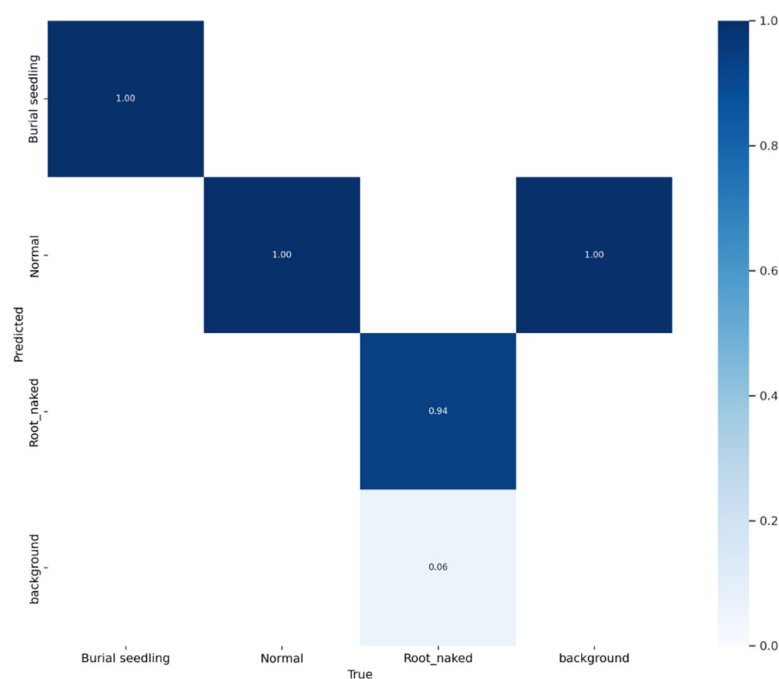


Figure 9. The image of confusion matrix normalized.

To verify the impact of adding CPNMViTBv3, scale aggregation detection head, and the studied loss function on the overall performance of the YOLOv11-ARM model, ablation experiments were designed. The experimental results of different improvement strategies for ablation are shown in Table 3.

Table 3. Ablation experiment.

Module		Object Detection					Keypoint Detection			
CPNMViTBv3	Scale Aggregation	Loss	Precision/%	Recall/%	mAP@0.5/%	mAP@0.95/%	Precision/%	Recall/%	mAP@0.5/%	mAP@0.95/%
✓	✓	✓	99.5	93.9	96.7	85.1	98.5	98.2	98.3	98.0
			99.5	97.4	98.2	86.9	98.5	99.0	99.3	99.0
			99.5	96.1	97.5	85.3	98.5	99.0	99.0	98.7
✓	✓	✓	98.6	97.5	98.4	87.2	98.5	98.5	98.7	98.4
✓	✓		99.5	97.8	98.7	86.7	98.5	99.0	99.5	99.4

As shown in Table 3, the incorporation of the CPNMViTBv3 backbone brings clear improvements to object detection performance. With this module added, the recall rises from 93.9% to 97.4%, and mAP@0.5 increases from 96.7% to 98.2%, indicating its effectiveness in capturing more discriminative features. When the scale aggregation mechanism is introduced alone, object detection recall improves to 96.1% and mAP@0.5 to 97.5%, demonstrating its role in enhancing multi-scale representation.

However, combining both CPNMViTBv3 and scale aggregation results in a slight drop in precision (98.6%) compared to using either module alone (99.5%). This can be attributed to an initial optimization challenge when integrating two complementary but distinct feature enhancement pathways. The CPNMViTBv3 module enriches features with global contextual information, while the scale aggregation head focuses on fusing multi-scale local details. Their simultaneous introduction increases the complexity and diversity of the feature representation, which may temporarily introduce some ambiguity during training, leading to a marginal increase in false positives and thus a dip in precision. Notably, this combination yields the highest mAP@0.95 (87.2%) among all configurations, indicating improved localization accuracy for difficult cases. This suggests a temporary balance where

the model gains in precise localization (reflected in mAP@0.95) while slightly adjusting its classification selectivity (precision) in the short term.

These improvements are established from a high-performance baseline (YOLOv11-Pose mAP@0.5: 96.7%). In this context, the absolute gains translate to substantial relative reductions in error—for instance, the increase to 98.7% mAP@0.5 corresponds to a reduction in the error rate by approximately 61%. The gains are also consistent across multiple related metrics (recall, mAP@0.5, mAP@0.95) for both detection tasks. To ensure the robustness of these comparisons, all ablation experiments were conducted under controlled conditions with fixed random seeds. This controlled setup supports the conclusion that the observed improvements are attributable to the architectural modifications.

The full model, integrating CPNMViTBv3, scale aggregation, and the dynamic adaptive loss, yields the best overall performance: 99.5% precision, 97.8% recall, 98.7% mAP@0.5 in object detection, and 99.5% mAP@0.5 together with 99.4% mAP@0.95 in keypoint detection. This confirms that the joint use of all three components enables more robust detection and keypoint estimation across varying seedling structures and scales.

This paper compares the loss functions of different models on the same validation set, and the map function curve is shown in Figure 10. By comparing the mAP@0.95 curves of target classification and keypoint detection separately, it can be clearly seen that as the training epochs increase, the loss value gradually decreases and eventually stabilizes. The improved model has a more stable convergence process. This indicates that the improved model can optimize parameters more effectively during the training process, thereby improving the overall performance of the model.

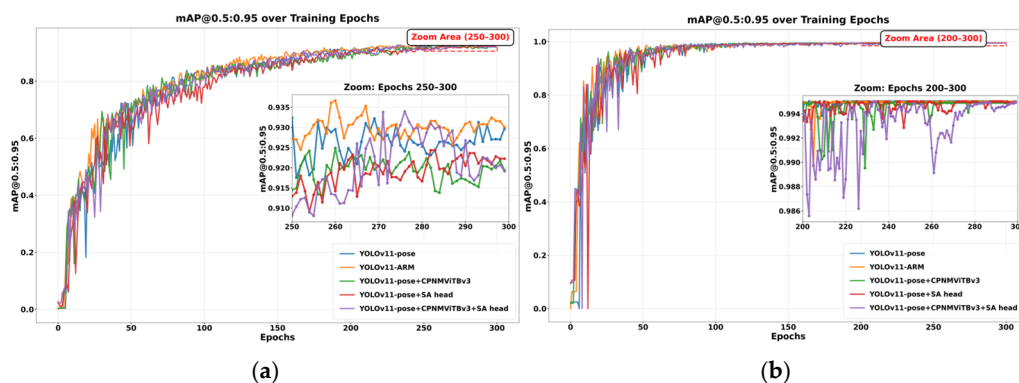


Figure 10. The detection performance of different algorithms under different seedling postures. (a) target classification loss; (b) keypoint detection loss.

To further verify that the model improvements enhance attention to cabbage seedlings and keypoints, this paper extracts weight features from the output layers of the Neck network, utilizing Grad-CAM for feature visualization as shown in Figure 11. YOLOv7-Pose produces the weakest and most scattered activations, failing to concentrate on important regions such as the stem. In comparison, both YOLOv11-ARM and DSW-YOLO generate strong, focused activations around the stem area. They also maintain good attention to edge regions, which helps in identifying bare roots and buried seedlings. Among them, YOLOv11-ARM highlights more precise and compact areas than the other models, showing its ability to extract features more effectively. This stronger visual focus matches its higher recall and mAP scores in both object and keypoint detection, as shown in the earlier performance table, confirming its advantage in accurately recognizing seedling structures.

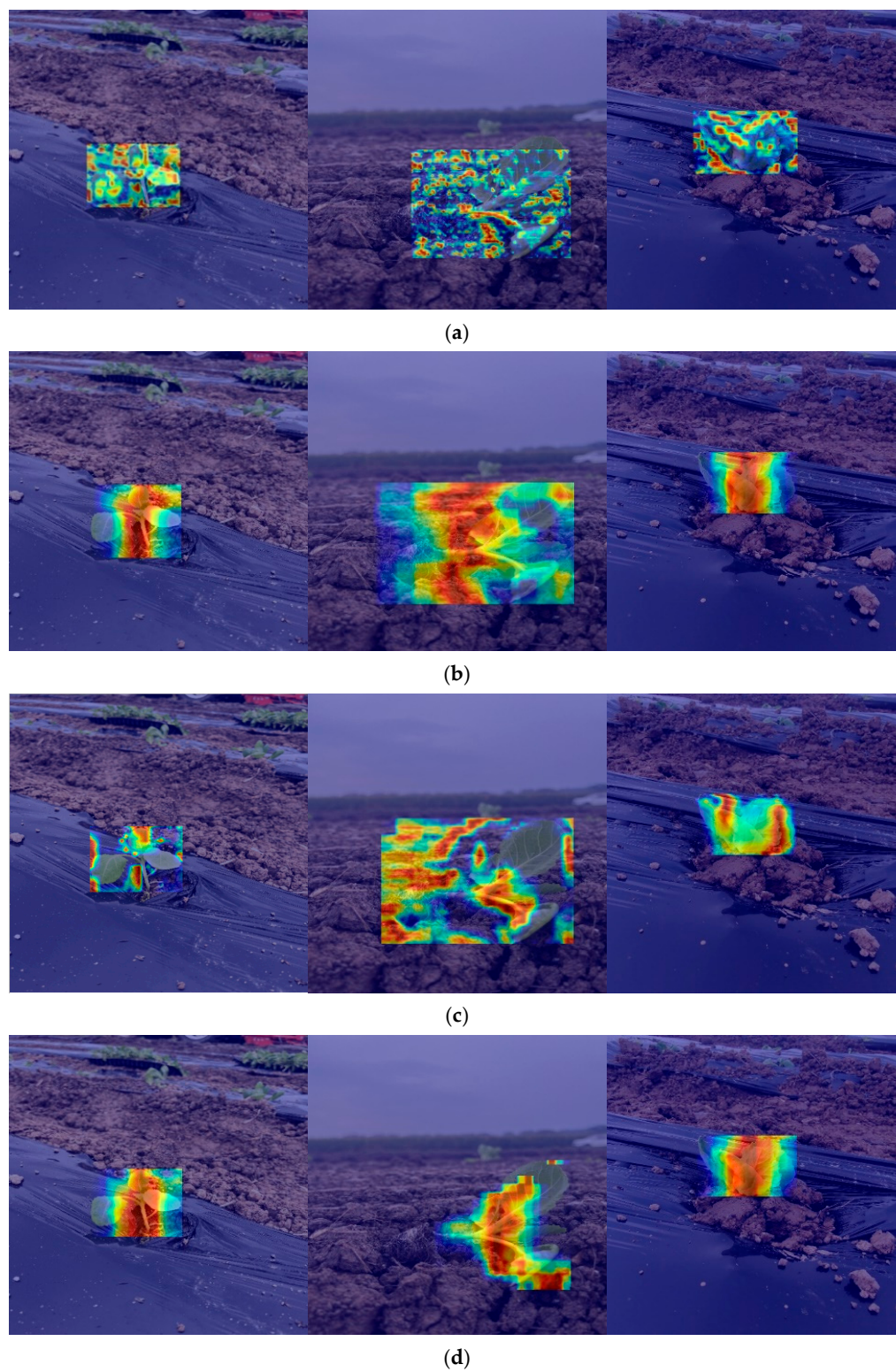


Figure 11. *Cont.*

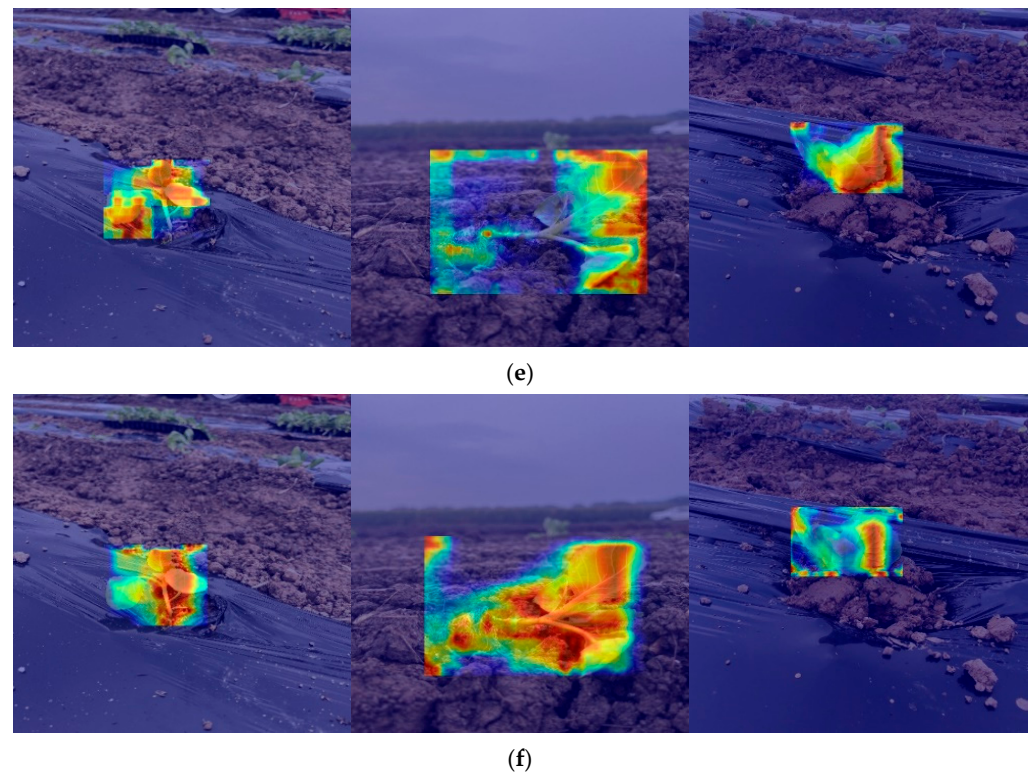


Figure 11. The heatmaps of different algorithms. (a) YOLOv7-Pose; (b) YOLOv8-Pose; (c) YOLOv11-Pose; (d) MD-YOLO; (e) DSW-YOLO; (f) YOLOv11-ARM. The colored regions represent the heatmaps output by the models.

In crop phenotype recognition, there is often a focus on large-scale information extraction methods such as 3D reconstruction or leaf area measurement. However, these methods face challenges related to high model complexity and insufficient information extraction. In contrast, detecting crop keypoints allows for a more accurate extraction of vital information, such as leaf length, leaf span, and the positional relationships of crop organs. This model fosters information complementarity through crop keypoint recognition, enabling a more comprehensive expression of crop phenotype characteristics and enhancing the accuracy of phenotype recognition. This paper conducts a systematic evaluation and comparison of several keypoint detection models on a cabbage seedling dataset. Furthermore, it studies a novel model, YOLOv11-ARM, which incorporates a hybrid backbone for enriched feature representation, a multi-scale detection head, and a dynamic loss function. This integrated approach achieves an improved balance between accuracy and efficiency, leading to more reliable detection of seedling structures.

In terms of key-point detection, Du et al. applied YOLOv8 to identify picking points for tomatoes and strawberries, but their focus was primarily on single-point detection during a specific crop period [30]. In contrast, this paper groups keypoints to detect various critical points at different growth stages of cabbage seedlings. A significant advantage of this method is its ability to balance model size and precision. Compared to other innovative models and traditional keypoint detection methods, it reduces training complexity while simultaneously improving detection accuracy.

Beyond algorithmic comparisons, the practical value of a vision-based detection model lies in its integration with automated machinery. For instance, recent work by Magar et al. developed an automatic seedling extraction device using an embedded system and mechanical toggle mechanism, focusing on the robust physical execution of transplanting [31]. While that study optimized extraction mechanics and field efficiency, our work provides the complementary “perception” capability. Specifically, the YOLOv11-ARM

model can serve as the vision module for such a system, enabling real-time assessment of transplanting quality (e.g., identifying exposed or buried seedlings) post-extraction. The compact size (3.54 MB) and computational efficiency (9.6 GFLOPs) of our model make it suitable for deployment on the resource-constrained embedded systems that control such agricultural devices. Thus, the two studies address different yet complementary aspects of intelligent transplanting: reliable physical execution and intelligent visual inspection. Their integration could lead to a closed-loop system where the extractor handles planting, and the vision module provides immediate quality feedback for process adjustment.

6. Conclusions

This study integrates keypoint detection with object recognition to identify and localize phenotypic keypoints in transplanted cabbage seedlings. The studied YOLOv11-ARM model incorporates a hybrid CPNMViTBv3 backbone for enriched feature representation, a multi-scale aggregation detection head, and a dynamic adaptive loss function. The primary findings, compared to other benchmark models, are as follows:

The ablation experiments indicate that each proposed component contributes positively to the model's overall performance. Introducing the CPNMViTBv3 backbone alone raises the recall for object detection from 93.9% to 97.4% and mAP@0.5 from 96.7% to 98.2%. The scale aggregation mechanism independently improves object detection recall to 96.1%. The complete model, integrating all three components, delivers the best overall results, achieving 98.7% mAP@0.5 in object detection and 99.5% mAP@0.5 in keypoint detection. This confirms that the combined architectural enhancements work synergistically to improve detection robustness across varying seedling structures and scales.

In comparative experiments, YOLOv11-ARM shows strong performance. It surpasses YOLOv8-Pose by 7.7 percentage points in recall and 7.7 points in mAP@0.5 for object detection. Compared to YOLOv11-Pose, it offers higher recall and mAP@0.5 in object detection. Although DSW-YOLO exhibits similar keypoint detection accuracy, our model demonstrates significantly stronger object detection capability, with over 20% higher recall and mAP@0.5. Furthermore, visualization results from Grad-CAM confirm that YOLOv11-ARM generates more precise and compact attention maps focused on critical regions like the stem, leading to more accurate keypoint localization compared to models with scattered activations, such as YOLOv7-Pose.

The compact model size (3.54 MB) and controlled computational cost (9.6 GFLOPs) reported in this study indicate its inherent suitability for real-time applications on resource-constrained devices. While direct measurement of frames-per-second (FPS) on specific edge hardware is a necessary step for full deployment validation, the presented efficiency metrics establish a strong foundation for such future work.

The hybrid CNN–Transformer architecture and keypoint-based strategy establish an extensible paradigm for real-time plant phenotyping. This framework can be readily adapted to quantify other agronomic traits (e.g., leaf angle or stem thickness) by redefining keypoint annotations, broadening its potential impact beyond transplanting quality assessment. The studied architectural improvements target general challenges in plant phenotyping, such as capturing long-range context and multi-scale features. The core designs, including the CPNMViTBv3 module and the enhanced detection head, are portable in concept and could be adapted to other real-time detection architectures, extending the potential impact of this work.

While this study systematically evaluates the studied model on a dedicated cabbage seedling dataset, the validation remains within a controlled experimental setting. The absence of large-scale testing in diverse, real-world production environments may affect the model's generalizability and robustness for practical keypoint identification. Future

work will focus on deploying the model in field conditions and developing an integrated seedling monitoring system. This will help validate its performance under variable agricultural settings and enhance its utility for real-time phenotypic analysis in intelligent transplanting applications.

Author Contributions: Conceptualization, X.H. and H.W.; methodology, X.H.; validation, X.H., H.W. and W.G.; writing—original draft preparation, H.W.; writing—review and editing, X.H.; visualization, W.G. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Reform and Development Project of Beijing Academy of Agricultural and Forestry Sciences—Research on Identification and Precise Positioning Technology for the Entire Growth Cycle of Open Field Cabbage (GGFZ20240104).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available on request from the corresponding author (the data are not publicly available due to privacy restrictions.).

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Xu, T.; Li, X.; He, J.; Han, S.; Wang, G.; Yin, D.; Zhou, M. Research Progress and Future Prospects of Key Technologies for Dryland Transplanters. *Appl. Sci.* **2025**, *15*, 8073. [\[CrossRef\]](#)
2. Reza, M.N.; Karim, M.R.; Ali, M.R.; Lee, K.-H.; Bicamumakuba, E.; Lee, K.Y.; Chung, S.-O. Field Evaluation of a Transplanter and a Collector Under Development for Korean Spring Cabbage Production in Greenhouses. *AgriEngineering* **2025**, *7*, 226. [\[CrossRef\]](#)
3. Rahim, U.; Wajid, A.; Usman, G. Deep Reinforcement Learning for Robust USVs Navigation in Diverse Environmental Scenarios. *Int. J. Data Inform. Intell. Comput.* **2025**, *4*, 1–10. [\[CrossRef\]](#)
4. Liu, W.; Tian, S.; Wang, Q.; Jiang, H. Key Technologies of Plug Tray Seedling Transplanters in Protected Agriculture: A Review. *Agriculture* **2023**, *13*, 1488. [\[CrossRef\]](#)
5. Jadhav, V.C.; Tated, R.G.; Bhangale, J.H.; Sancheti, S.D.; Gore, K.B. A review on vegetable seedlings transplantation systems and their importance in Indian agriculture. *Int. J. Veg. Sci.* **2025**, *31*, 914–947. [\[CrossRef\]](#)
6. Du, X.; Gao, K.; Si, L. Design and Experiments of a Vision-Guided Integrated Cavity Tray Seedling Culling and Replanting System. *IEEE Access* **2023**, *11*, 144545–144559. [\[CrossRef\]](#)
7. Jin, X.; Tang, L.; Li, R.; Ji, J.; Liu, J. Selective transplantation method of leafy vegetable seedlings based on ResNet 18 network. *Front. Plant Sci.* **2022**, *13*, 893357. [\[CrossRef\]](#) [\[PubMed\]](#)
8. Fu, W.; Gao, J.; Zhao, C.; Jiang, K.; Zheng, W.; Tian, Y. Detection Method and Experimental Research of Leafy Vegetable Seedlings Transplanting Based on a Machine Vision. *Agronomy* **2022**, *12*, 2899. [\[CrossRef\]](#)
9. Li, M.; Xiao, L.; Ma, X.; Yang, F.; Jin, X.; Ji, J. Vision-Based a Seedling Selective Planting Control System for Vegetable Transplanter. *Agriculture* **2022**, *12*, 2064. [\[CrossRef\]](#)
10. Chen, K.; Yuan, Y.; Zhao, B.; Zhou, L.; Niu, K.; Jin, X.; Gao, S.; Li, R.; Guo, H.; Zheng, Y. Digital Twins and Data-Driven in Plant Factory: An Online Monitoring Method for Vibration Evaluation and Transplanting Quality Analysis. *Agriculture* **2023**, *13*, 1165. [\[CrossRef\]](#)
11. Li, M.; Zhu, X.; Ji, J.; Jin, X.; Li, B.; Chen, K.; Zhang, W. Visual perception enabled agriculture intelligence: A selective seedling picking transplanting robot. *Comput. Electron. Agric.* **2025**, *229*, 109821. [\[CrossRef\]](#)
12. Jin, X.; Zhao, G.; Li, M.; Li, S.; Zhou, L.; Liang, J.; Xie, X.; Hou, B. Optimization of transplanting mechanism working parameters based on a coupled machine-soil-pot seedling model. *Comput. Electron. Agric.* **2025**, *238*, 110786. [\[CrossRef\]](#)
13. Jin, X.; Tang, L.; Li, R.; Zhao, B.; Ji, J.; Ma, Y. Edge recognition and reduced transplantation loss of leafy vegetable seedlings with Intel Realsense D415 depth camera. *Comput. Electron. Agric.* **2022**, *198*, 107030. [\[CrossRef\]](#)
14. Zhao, S.Y.; Lei, X.J.; Liu, J.Z.; Jin, Y.C.; Bai, Z.C.; Yi, Z.Y.; Liu, J.L. Transient multi-indicator detection for seedling sorting in high-speed transplanting based on a lightweight model. *Comput. Electron. Agric.* **2023**, *211*, 107996. [\[CrossRef\]](#)
15. Xue, K.; Gao, K.; Kuang, F.; Zhang, S.; Liao, J.; Zhu, D. Machinery-plant-paddy soil coupling model based numerical simulation method of mechanical transplanting process of big rice seedling. *Comput. Electron. Agric.* **2022**, *198*, 107053. [\[CrossRef\]](#)
16. de Villiers, H.A.C.; Otten, G.; Chauhan, A.; Meesters, L. Autoencoder-based 3D representation learning for industrial seedling abnormality detection. *Comput. Electron. Agric.* **2023**, *206*, 107619. [\[CrossRef\]](#)

17. Chen, Q.; Huang, S.; Liu, S.; Zhong, M.; Zhang, G.; Song, L.; Zhang, X.; Zhang, J.; Wu, K.; Ye, Z.; et al. Multi-view 3D reconstruction of seedling using 2D image contour. *Biosyst. Eng.* **2024**, *243*, 130–147. [[CrossRef](#)]
18. Okabe, Y.; Hiraguri, T.; Endo, K.; Kimura, T.; Hayashi, D. Classification of Tomato Harvest Timing Using an AI Camera and Analysis Based on Experimental Results. *AgriEngineering* **2025**, *7*, 48. [[CrossRef](#)]
19. Wen, Y.S.; Zhang, J.X.; Tian, J.Y.; Duan, D.S.; Zhang, Y.; Tan, Y.Z.; Yuan, T.; Li, X.J. Design of a traction double-row fully automatic transplanter for vegetable plug seedlings. *Comput. Electron. Agric.* **2021**, *182*, 106017. [[CrossRef](#)]
20. Na, M.; Teng, L.; Zhou, Z.; Zhou, K.; Wang, J.; Zhou, M. Design and experiment of fully automatic slide-track type dryland pot seedling transplanting mechanism. *Trans. Chin. Soc. Agric. Mach.* **2021**, *52*, 54–61. [[CrossRef](#)]
21. Wu, G.; Yu, G.; Zhou, H.; Ge, Y.; Ye, A.; Wang, L. Reverse Design and Tests of Three-arms Rotary Rice Seedling Transplanting Mechanism. *Trans. Chin. Soc. Agric. Mach.* **2022**, *53*, 84–91. [[CrossRef](#)]
22. Jin, X.; Zhu, X.; Xiao, L.; Li, M.; Li, S.; Zhao, B.; Ji, J. YOLO-RDS: An efficient algorithm for monitoring the uprightness of seedling transplantation. *Comput. Electron. Agric.* **2024**, *218*, 108654. [[CrossRef](#)]
23. Cheng, S.; Han, Y.; Wang, Z.; Liu, S.; Yang, B.; Li, J. An Underwater Object Recognition System Based on Improved YOLOv11. *Electronics* **2025**, *14*, 201. [[CrossRef](#)]
24. Wang, C.-Y.; Liao, H.-Y.M.; Wu, Y.-H.; Chen, P.-Y.; Hsieh, J.-W.; Yeh, I.-H. CSPNet: A New Backbone that can Enhance Learning Capability of CNN. In Proceedings of the CVPR Workshops, Seattle, WA, USA, 14–19 June 2020.
25. Xia, Z.; Zhou, H.; Yu, H.; Hu, H.; Zhang, G.; Hu, J.; He, T. YOLO-MTG: A lightweight YOLO model for multi-target garbage detection. *Signal Image Video Process.* **2024**, *8*, 5121–5136. [[CrossRef](#)]
26. Niu, F.; Sheng, Y.; Wang, J.; Zheng, X.; Liu, K.; Lin, Y.; Wang, W.; Li, G. Domain generalization for sea cucumber detection: Tackling background color variability in aquaculture settings. *Aquac. Int.* **2025**, *33*, 339. [[CrossRef](#)]
27. Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft COCO: Common Objects in Context. In *Computer Vision—ECCV 2014*; Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T., Eds.; Lecture Notes in Computer Science; Springer: Cham, Switzerland, 2014; Volume 8693. [[CrossRef](#)]
28. Tian, Y.; Wang, S.; Li, E.; Yang, G.; Liang, Z.; Tan, M. MD-YOLO: Multi-scale Dense YOLO for small target pest detection. *Comput. Electron. Agric.* **2023**, *213*, 108233. [[CrossRef](#)]
29. Du, X.; Cheng, H.; Ma, Z.; Lu, W.; Wang, M.; Meng, Z.; Jiang, C.; Hong, F. DSW-YOLO: A detection method for ground-planted strawberry fruits under different occlusion levels. *Comput. Electron. Agric.* **2023**, *214*, 108304. [[CrossRef](#)]
30. Du, X.; Meng, Z.; Ma, Z.; Lu, W.; Cheng, H. Tomato 3D pose detection algorithm based on keypoint detection and point cloud processing. *Comput. Electron. Agric.* **2023**, *212*, 108056. [[CrossRef](#)]
31. Magar, A.P.; Nalawade, S.M.; Walunj, A.A.; Gaikwad, B.B.; Shelke, G.N.; Khadatkhar, A.; Sawant, C.P.; Modi, R.U. Development and optimization of an automatic transplanting device for plug tray vegetable seedlings using embedded system. *Sci. Rep.* **2025**, *15*, 22483. [[CrossRef](#)] [[PubMed](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.