Smart Agricultural Technology 11 (2025) 100903

Contents lists available at ScienceDirect

Smart Agricultural Technology

journal homepage: www.journals.elsevier.com/smart-agricultural-technology

Automated detection and segmentation of baby kale crowns using grounding DINO and SAM for data-scarce agricultural applications

Gianmarco Goycochea Casas [a,*], Zool Hilmi Ismail [b], Mohd Ibrahim Shapiai [b], Ettikan Kandasamy Karuppiah [c]

[a] Department of Forest Engineering, Federal University of Viçosa, Viçosa, MG 36570-900, Brazil
[b] Center for Artificial Intelligence and Robotics, Universiti Teknologi Malaysia, Jalan Sultan Yahya Petra, Kuala Lumpur 54100, Malaysia
[c] NVIDIA Corporation, Singapore, Singapore

ARTICLE INFO

Keywords:
Multimodal AI
Precision agriculture
Computer vision
Zero-Shot Learning
Automated Labeling
NVIDIA

ABSTRACT

This research addresses the significant challenge of data scarcity in agriculture by introducing an automatic pipeline for plant detection and segmentation. The primary objective was to detect and segment the crown area of baby kale (Brassica oleracea var. sabellica) during its early growth stages without relying on extensive data training or manual annotations, providing an alternative for scenarios with insufficient data. A dataset comprising aerial images of baby kale plants was gathered over a three-week period in a controlled environment. The model was processed using the NVIDIA GeForce RTX 4060 GPU. Grounding DINO was employed for plant detection based on textual prompts, and bounding boxes were generated to locate the central plant in each image. The detected regions were then processed using SAM to extract precise segmentation masks of the plant crown. The segmentation results were validated by comparing the automated method with manually annotated ground truth using statistical metrics, including Spearman's correlation, RMSE%, and the Wilcoxon signed-rank test. The automated approach demonstrated a strong correlation ($\rho = 0.956$) with manual annotations across all weeks, with RMSE% decreasing as plants matured. While Week 1 exhibited lower agreement ($\rho = 0.581$, RMSE% = 56.246 %) due to segmentation challenges at early growth stages, performance improved significantly in Week 2 ($\rho = 0.945$, RMSE% = 24.834 %) and Week 3 ($\rho = 0.996$, RMSE% = 11.733 %). The statistical validation confirmed a significant difference between manual and automated annotations; however, the automated method consistently captured the growth trend of the plants. In conclusion, while the pipeline offers a promising approach for plant detection and segmentation in data-scarce environments, its limitations, especially in early growth stages, should be considered. The study contributes by demonstrating a practical approach to overcoming data scarcity in agriculture using multimodal AI models capable of zero-shot and few-shot learning. This approach paves the way for more adaptive AI-driven agricultural monitoring systems, addressing data scarcity challenges in precision farming.

1. Introduction

Data scarcity poses a significant challenge in training Convolutional Neural Networks (CNNs) for artificial intelligence applications. Researchers continue to explore this arduous topic, implementing strategies to overcome data scarcity. These include *data augmentation*, a widely used technique to artificially increase the size of the training set through transformations such as rotation, scaling, and flipping [1]; *Generative Adversarial Networks* (GANs) that can generate synthetic data mimicking the characteristics of real data, helping to augment the training set [2]; *self-supervised learning*, which utilizes pretext tasks to learn visual representations without the need for labels, although its performance may be inferior to supervised learning [3]; and *feature-based knowledge distillation (FKD)*, which employs expert models to guide the training of new models in data-limited domains [4].

The agricultural sector is not immune to these challenges, facing data scarcity as a major hurdle. A lack of high-quality and diverse datasets limits the development of robust machine learning models [5]. The dynamic and variable nature of agricultural systems demands precise and up-to-date data for informed decision-making. Additionally,

* Corresponding author.
E-mail address: gianmarco.casas@ufv.br (G.G. Casas).

https://doi.org/10.1016/j.atech.2025.100903
Received 7 February 2025; Received in revised form 19 March 2025; Accepted 19 March 2025
Available online 20 March 2025

existing datasets often lack adequate labelling and quality, hindering the training of accurate models. Despite these challenges, deep learning has revolutionized agriculture by improving crop classification, yield estimation, and pest detection [6–8,9]. However, the scarcity of high-quality, robust data poses a significant barrier to fully realizing the potential of deep learning in agriculture. The time-consuming and costly nature of collecting, curating, and labelling agricultural datasets has discouraged many businesses from investing in this area.

To address this data scarcity, an approach involving the utilization of multimodal AI (Artificial Intelligence) models. These models are capable of processing and integrating both visual and textual information, coupled with segmentation models, to effectively leverage the existing data provided by suppliers.

Multimodal AI models excel at processing and integrating information from various sources like text, images, sounds, and videos. This allows for a deeper and more holistic understanding of content. Their ability to handle complex tasks that require combining different types of data has made them increasingly popular. A prime example is the creation of a foundational AI model trained on a massive dataset of diverse information. This foundational model can then be quickly adapted to perform a wide range of tasks. This showcases the immense potential of multimodal AI to make AI systems more adaptable and useful across various fields [10].

Multimodal AI models are increasingly being applied in agriculture to enhance various aspects of farming and crop management. Recent scientific studies have explored these applications in depth. This method effectively addresses challenges in agricultural disease detection and question-answering systems, providing new perspectives and tools for the development of intelligent agriculture [11].

One of the models is Grounding DINO, which incorporates a transformer-based detector called DINO (DETR with Improved DeNoising Anchor Boxes). This AI system operates within the field of computer vision, with primary emphasis on object detection. By integrating language comprehension, it also intersects with natural language processing, thus creating a multimodal AI model that can analyze and merge both visual and textual data. The model is an open-set object detector that integrates the Transformer-based DINO detector with grounded pre-training, enabling the detection of arbitrary objects based on human inputs like category names or referring expressions. The key to open-set object detection is incorporating language into a closed-set detector to generalize open-set concepts. To effectively fuse language and vision modalities, the model is conceptually divided into three phases: a feature enhancer, language-guided query selection, and a cross-modality decoder for cross-modality fusion [12].

Building on the solid groundwork laid by its predecessors, Grounding DINO unveils a series of sophisticated open-set object detection models crafted by the experts at IDEA Research. These cutting-edge models are designed with the ambitious goal of pushing the boundaries of open-set object detection even further, offering new levels of accuracy and versatility in identifying and analyzing objects in diverse and evolving environments [13].

Grounding DINO model could be utilized as a strategy to overcome data scarcity in object detection tasks. Several key factors highlight its usefulness in this context: it offers generalization to unseen objects (open-set detection), allowing for the detection of arbitrary objects based on input text. This means it can recognize objects without the need for explicit training with images of those objects. Furthermore, it excels in zero-shot and few-shot learning, enabling zero-shot detection where it can detect objects it has not been trained on, thereby reducing the need for labeled data. Additionally, by detecting objects based on text descriptions, it facilitates the automated generation of annotations, reducing the need for manual labeling of large volumes of data.

The Segment Anything Model (SAM), developed by Meta AI, represents a significant advancement in vision foundation models for image segmentation tasks. It is distinguished by its capability to perform zero-shot segmentation, which allows it to segment objects into images without necessitating task-specific training [14,15]. The integration of both DINO and SAM models combines object detection and segmentation abilities. This combination utilizes the strengths of both models to handle complex visual tasks, especially in open-world scenarios where zero-shot detection and segmentation are needed [16].

In this study, we focused on the early growth stage of Kale (*Brassica oleracea* var. *sabellica*), known as baby Kale, cultivated in a laboratory under artificial light. Baby Kale is significant due to its economic viability in vertical farming systems, owing to its high harvest index and short growth cycles, which allow for multiple harvests per year [17]. This species is celebrated for its high content of bioactive components, including phenolic compounds, glucosinolates, chlorophylls, carotenoids, and minerals such as calcium, potassium, iron, magnesium, vitamins C and E, as well as unsaturated fatty acids and proteins [18]. This nutritional profile has caught the attention of the nutraceutical industry, which has recently begun to recognize the health benefits of kale [19, 20].

Due to the importance of this vegetable, we implemented the Grounding DINO model along with the SAM with the main objective of detecting and segmenting the crown area during the early growth stage of Kale, without the need for extensive data training and labeling as an alternative for cases of insufficient data.

Early measurement of the plant crown provides critical insights into growth dynamics under varying treatment conditions. This information is essential for understanding how plants respond to experimental nutritional factors. By detecting and segmenting the crown area at an early stage, it is possible to predict plant growth and adjust agricultural practices accordingly (J and [21]). Moreover, early detection enables timely interventions, which can save plants and prevent the spread of diseases to others [22].

2. Materials and methods

2.1. Dataset

Data were collected from aerial photographs taken at random of various baby kale (*Brassica oleracea* var. *sabellica*) plants at a distance of 1 meter. The photographs were taken over a three-week period, from the first week of germination to the third week of growth. In total, 175 photographs were received, with 54 photographs from the first week, 58 photographs from the second week, and 63 photographs from the third week. From these, photographs that met processing requirements were selected, such as no stains, in focus on the plant, clear images, and images with no overexposure or opacity. In total, 85 photographs met the processing requirements, with 35 photographs for the first week, 41 photographs for the second week, and 9 photographs for the third week. The latter decreased considerably because several photographs superimposed plants on other plants. Consequently, photographs featuring a single plant were prioritized.

The plants were grown in a plant factory at The Malaysian Agricultural Research and Development Institute (MARDI). The plant factory is a 2400 square meters, 20-meter-high facility that uses a controlled environment to optimize plant growth. Factors such as lighting, irrigation, fertilization, temperature, humidity, and ventilation were controlled and adjusted with a single treatment for the same plants. The plants are grown under multi-colored light-emitting diodes (LEDs).

2.2. Building model

The system used featured a 13th Gen Intel(R) Core (TM) i7–13700H processor with a clock speed of 2.40 GHz. It had 8.00 GB of installed RAM and operated on a 64-bit operating system on an x64-based processor. It was equipped with an NVIDIA GeForce RTX 4060 GPU, which had 8 GB of GDDR6 memory and a bus width of 128 bits.

This process aims to accurately detect and segment the crown of the baby kale plant, leveraging the advanced capabilities of the Grounding

DINO and Segment Anything Model (SAM) models. The following sections delve into the specific stages of this process, outlining the methodology and techniques employed at each step to achieve precise and reliable crown detection and segmentation (Fig. 1):

2.2.1. Image resizing

Initially, all images were resized to a standard dimension of 640×640 pixels. This initial resizing step is crucial for ensuring consistency in subsequent processing stages and optimizing the performance of both detection and segmentation models. By establishing a uniform input size, we mitigate potential variations in image dimensions that could negatively impact model accuracy and efficiency. This standardization also streamlines data handling and reduces computational overhead during model training and inference

2.2.2. DINO grounding configuration

Grounding DINO is an open-source object detection model that combines the Transformer-based DINO detector with grounded pre-training, enabling the detection of arbitrary objects through human input, such as category names or referential expressions. Installation involves cloning the Grounding DINO repository and configuring the necessary Python environment. Subsequently, the model is loaded using a specific configuration file and pre-trained weights.

a. Plant Detection and Coordinate Normalization

In this stage, the Grounding DINO model is employed to detect the plant within each resized image. The model takes as input an image and a text prompt—in this case, the word "*plant*"—and generates bounding boxes that automatically indicate the plant's location within the image. For each image, the box closest to the image's center is selected. The coordinates of this box are then normalized by dividing the center coordinates (cx, cy) and the dimensions (w, h) by the image's width and height, respectively. This ensures that the coordinates fall within a [0, 1] range, facilitating their use in subsequent stages.

The model introduces three core components:

- Feature Enhancer: Improves the extracted features from images and text.
- Language-Guided Query Selection: Selects relevant queries from image features using text guidance.
- Cross-Modality Decoder: Fuses vision and language features to generate object predictions.

The mathematical foundations of Grounding DINO are based on Transformer architectures, particularly DETR (DEtection TRansformer) and DINO (DETR with Improved Training) [12]. Below, we outline key formulations:

2.2.2.1. Language-guided query selection. Let $X_I \in R^{N_I \times d}$ be the image features, and $X_T \in R^{N_T \times d}$ be the text features. Note here that N_I is the number of image tokens N_T; is the number of text tokens; and d is the feature dimension.

Grounding DINO selects top N_q image queries using the following formula (1):

$$I_{N_q} = Top_{N_q}\left(Max^{(-1)}\left(X_I X_T^T\right)\right) \quad (1)$$

Note here that $Max^{(-1)}$ performs a maximum operation along the last dimension; X_T^T is the transposed text feature matrix; and Top_{N_q} selects the top N_q query indices.

Each selected query consists of Content Part (learnable during training) and Positional Part (modeled as dynamic anchor boxes).

The following pseudocode (Algorithm), resembling PyTorch syntax, details the Language-Guided Query Selection process. This algorithm, described in Liu et al. [12], selects the most relevant image queries based on a given text input, optimizing query selection for downstream object detection.

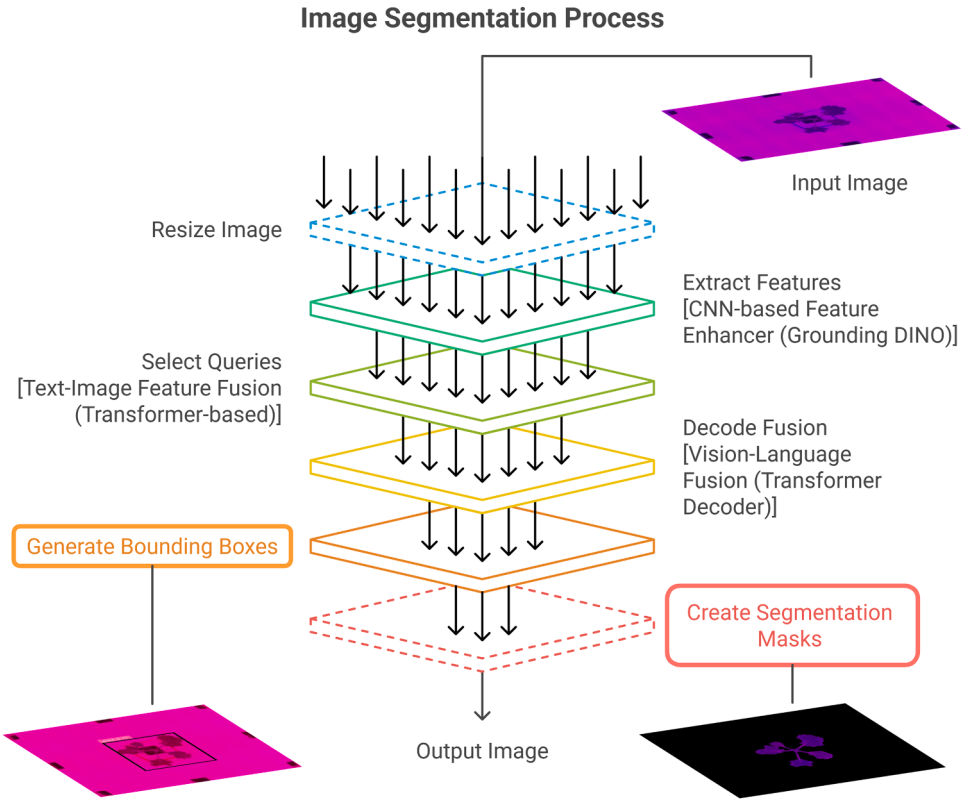

Fig. 1. Schematic representation of the image segmentation process using grounding DINO and SAM.

Algorithm

```
"""
    Input:
        image_feat: (bs, num_img_tokens, ndim)
        text_feat: (bs, num_text_tokens, ndim)
        num_query: int
    Output:
        topk_idx: (bs, num_query)
    """
    # Compute similarity between image and text features
    logits = torch.einsum("bic,btc->bit", image_feat, text_feat) # bs, num_img_tokens, num_text_tokens
    # Select the most relevant image features
    logits_per_img_feat = logits.max(−1)[0] # bs, num_img_tokens
    topk_idx = torch.topk(logits_per_img_feat, num_query, dim=1)[1] # bs, num_query
```

2.2.2.2. Cross-modality decoder. The model introduces a Cross-Modality Decoder, which differs from DINO by incorporating text-based cross-attention. Each decoder layer consists of:

- Self-Attention: Computes attention across all queries.
- Image Cross-Attention: Merges image and query embeddings.
- Text Cross-Attention: Injects text information for semantic alignment.
- Feed-Forward Network (FFN).

Formally, the query update at each layer follows (2):

$$Q' = \text{Self} - \text{Attn}(Q) + \text{Image} - \text{Cross} - \text{Attn}(Q, X_I) + \text{Text} - \text{Cross} - \text{Attn}(Q, X_T) + \text{FFN}(Q) \quad (2)$$

Note here that Q' represents the queries from the previous decoder layer.

2.2.2.3. Loss function. The model uses a Hungarian Matching Loss, which consists of Classification Loss ($\mathscr{L}_{\text{cls}}$)—Cross-entropy loss for object classification, Box L1 Loss ($\mathscr{L}_{L_1}$)—Measures the absolute difference in bounding box coordinates, and Generalized IoU Loss (GIoU) ($\mathscr{L}_{\text{GIoU}}$)—Evaluates the overlap between predicted and ground-truth boxes. The total loss function is (3):

$$\mathscr{L} = \lambda_{\text{cls}}\mathscr{L}_{\text{cls}} + \lambda_{L_1}\mathscr{L}_{L_1} + \lambda_{\text{GIoU}}\mathscr{L}_{\text{GIoU}} \quad (3)$$

Note here that loss weights are set as $\mathscr{L}_{\text{cls}} = 2.0$, $\mathscr{L}_{L_1} = 5.0$, and $\mathscr{L}_{\text{GIoU}} = 2.0$.

b. Coordinates to Pixels Conversion

The normalized coordinates were converted back to pixel values by multiplying them by the image's width and height (640 in this case). From these pixel coordinates, the *xmin* (4), *ymin* (5), *xmax* (6), and *ymax* (7) values were calculated, representing the top-left and bottom-right corners of the bounding box in pixel terms. This conversion was necessary to ensure that the labels generated by the Grounding DINO model aligned with the input requirements of the SAM model. The formulas used for this conversion are as follows:

$$\text{xmin} = \text{cx} - (\text{w} / 2); \quad (4)$$

$$\text{ymin} = \text{cy} - (\text{h} / 2); \quad (5)$$

$$\text{xmax} = \text{cx} + (\text{w} / 2); \text{ and} \quad (6)$$

$$\text{ymax} = \text{cy} + (\text{h} / 2) \quad (7)$$

2.2.3. SAM configuration for mask extraction

The SAM is a cutting-edge image segmentation model. Its adaptable segmentation capabilities provide unprecedented versatility for image analysis tasks.

After installing SAM, the model is loaded, and a predictor is configured. For each image, using previously obtained bounding box coordinates, SAM generates a mask that segments the region of interest—in this case, the kale plant's crown. The resulting mask is applied to the original image, yielding a segmented image where only the region of interest is visible.

SAM's architecture comprises three primary components [14]:

- Image Encoder: Extracts high-dimensional feature representations from an input image. The image encoder of SAM is based on a Vision Transformer (ViT-H/16) trained with Masked Autoencoders (MAE). The encoded image representation is given by (8):

$$E = f_{ViT}(I) \quad (8)$$

Note here that E is the encoded image feature map of size C×H′×W'; I is the input image of size; and f_{ViT} is the vision transformer model.

The encoded image representation is downscaled by a factor of 16, meaning (9):

$$H' = \frac{H}{16};\ W' = \frac{W}{16} \quad (9)$$

- Prompt Encoder: Encodes various prompts such as bounding boxes, points, or free-form text. Prompts are encoded based on their type:

Points: A point (x, y) is embedded using positional encoding (10):

$$p = PE(x, y) + e_{\substack{fg \\ bg}} \quad (10)$$

Note here that $PE(x, y)$ is a sinusoidal positional encoding; and $e_{\substack{fg \\ bg}}$ is a learned embedding for foreground/background differentiation.

Bounding Boxes: Represented as two corner points (x_1, y_1) and (x_2, y_2) (11):

$$b = PE(x_1, y_1) + e_{top-left} + PE(x_2, y_2) + e_{bottom-right} \quad (11)$$

Masks: Downsampled using two convolutional layers before embedding

- Mask Decoder: Predicts segmentation masks from the encoded prompts and image features. The mask decoder combines the encoded image and prompts to produce a segmentation mask M. The process is:

Attention-based Fusion: Uses a Transformer decoder with cross-attention layers (12):

$$T = \text{SelfAttn}(P) + \text{CrossAttn}(P, E) \quad (12)$$

Note here that P represents the encoded prompt features; and E is the encoded image features.

Mask Prediction: The final segmentation mask M is computed as (13):

$$M = \sigma(\text{W}_\text{m}\text{T}) \quad (13)$$

Note here that W_m is a learned weight matrix; and σ is the sigmoid activation function

Finally, the process of automatic mask extraction is explained using the labeling intersections by the Grounding DINO model:

1. Model Loading: The pre-trained SAM model is loaded into memory.
2. Image Preprocessing: The input image is loaded and preprocessed to match the model's expected input format.
3. Prompt Definition: A bounding box is defined to specify the region of interest within the image. This bounding box was generated by the Grounding DINO model automatically.
4. Image Encoding: The preprocessed image is passed through the image encoder to obtain its embedding.
5. Prompt Encoding: The bounding box prompt is processed by the prompt encoder to generate its embedding.
6. Mask Generation: The mask decoder combines the image and prompt embedding to produce the segmentation mask. The area of a mask was calculated by summing nonzero pixels in the binary segmentation mask. This provides a direct measure of the object size in pixels. Given a binary segmentation mask M of size H×W (height and width), the mask area A is computed as (14):

$$A = \sum_{i=1}^{H} \sum_{j=1}^{W} 1\left(M_{i,j} > 0\right) \tag{14}$$

Note here that $M_{i,j}$ is the pixel value at position (i, j); and $1\left(M_{i,j} > 0\right)$ is an indicator function that returns 1 if the pixel is part of the mask (nonzero) and 0 otherwise.

7. Output: The segmented image is saved or displayed as needed.

2.3. Model validation

The crown area of the baby kale plant was computed using two different methods: manually labeled masks and automatically generated masks using the Grounding DINO + SAM model. These two datasets were then compared statistically to evaluate their agreement and reliability. The Shapiro-Wilk test was performed to assess the normality of the data distribution. Since the p-value was below the significance threshold ($p < 0.05$), we reject the null hypothesis, indicating that the data do not follow a normal distribution. Given the non-normal distribution of the data, the Spearman correlation coefficient (ρ) (15), the Root Mean Square Error Percentage (RMSE%) (16), and the Wilcoxon signed-rank test (17) were used for further validation.

$$\rho = 1 - \frac{6\sum d_i^2}{n(n^2-1)} \tag{15}$$

Note here that d_i^2 is the difference between the ranks of corresponding values and n is the number of observations

$$RMSE\% = \frac{RMSE}{\overline{X}} \times 100 \tag{16}$$

Note here that RMSE $= \sqrt{\frac{1}{n}\sum (X_i - Y_i)^2}$, X_i and Y_i individual values from the two datasets and $\overline{X}$ mean of the manually labeled masks dataset.

$$W = \sum R^+ \tag{17}$$

Note here that R^+ is the sum of the ranks of the positive differences between paired values

3. Results

3.1. Visual analysis of detection and segmentation

We have evaluated the performance of the DINO model for object detection and the SAM model for image segmentation in the context of plant image analysis. Fig. 2 illustrates the results of images processing using the Grounding DINO + SAM model for object detection and segmentation of kale plants over three consecutive weeks. Each row represents the same experimental setup at different time points (Week 1, Week 2, and Week 3), showcasing the model's performance in detecting and segmenting the plant crown as it grows.

The raw images serve as the input data, captured under controlled environmental conditions. These images exhibit the plant's structural development over time. The Grounding DINO model identifies the plant within the image and generates bounding boxes with confidence scores, which indicate the model's certainty in detecting the object. The bounding boxes enclose the region of interest, which is later used as input for the segmentation model. Using the bounding boxes from the detection stage, the SAM model extracts segmentation masks, isolating the plant crown from the background. The segmented masks accurately delineate the plant structure, allowing for further quantitative analysis of its area.

3.2. Temporal analysis of automated plant detection and segmentation

To further analyze the performance of the automated plant detection and segmentation process, a temporal evaluation was conducted by monitoring the plants at two-day intervals over a three-week period (Fig. 3). Each column corresponds to a specific week, while each row represents a different day within that week (Day 2, Day 4, and Day 6). The bounding boxes generated by Grounding DINO highlight the detected plants, automatically selecting the central plant in each image. The detection confidence score is displayed within each bounding box, showing a progressive increase in detection reliability as the plant grows. The masks extracted using SAM isolate the plant's crown, highlighting its morphology at different stages of growth. The segmentation results demonstrate an increase in leaf expansion and structural complexity over time.

Week 1, the plants exhibit small leaf structures with lower detection confidence scores, ranging from 0.27 to 0.48. The segmented masks show minimal foliage, corresponding to early plant development. Week

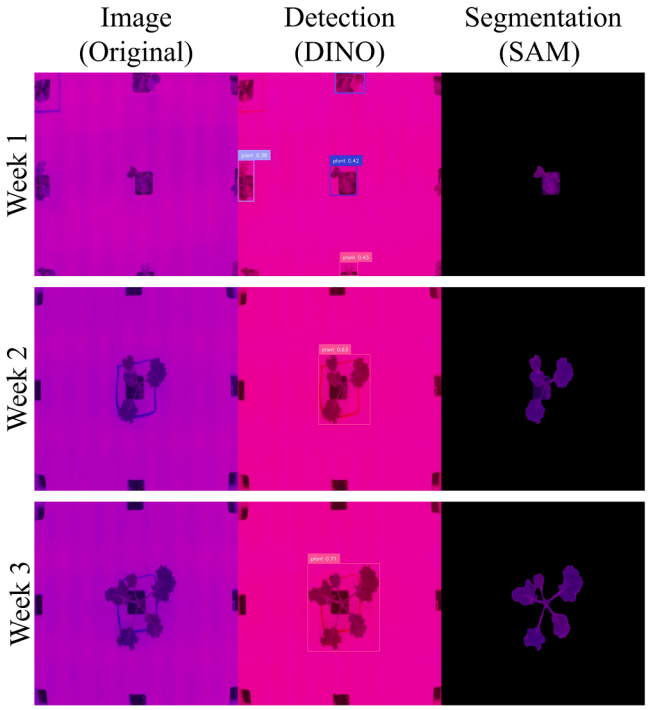

Fig. 2. Automated detection and segmentation of baby Kale plants using Grounding DINO and SAM across different growth stages.

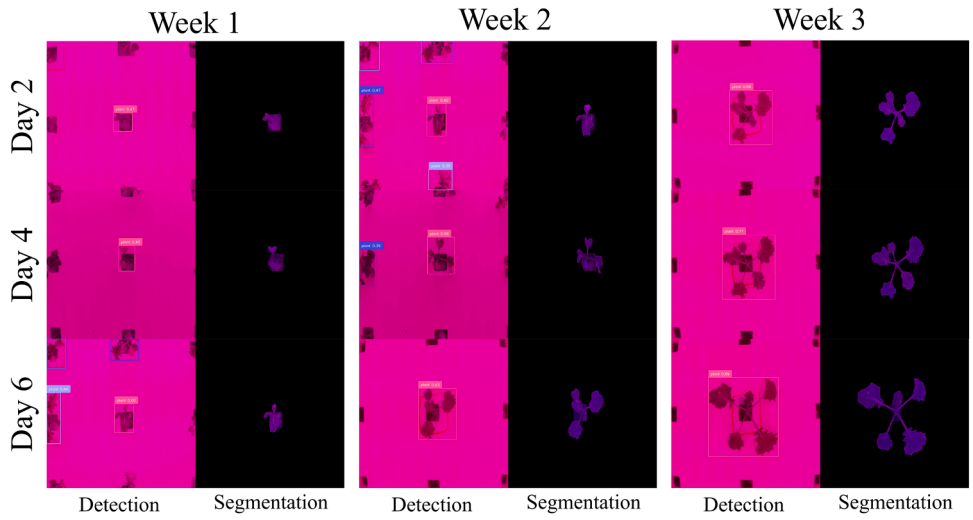

Fig. 3. Automated temporal detection and segmentation of baby Kale plants over a three-week growth period.

2, the plant structures become more defined, leading to higher confidence scores (0.39 to 0.61) in detection. Segmentation results reveal more pronounced leaf expansion, capturing finer details in the plant morphology. Week 3, the plants have significantly expanded, with detection scores reaching their highest values (0.60 to 0.71). Segmentation masks accurately delineate the larger, more developed plant structures, demonstrating the robustness of Grounding DINO + SAM in capturing complex morphological variations.

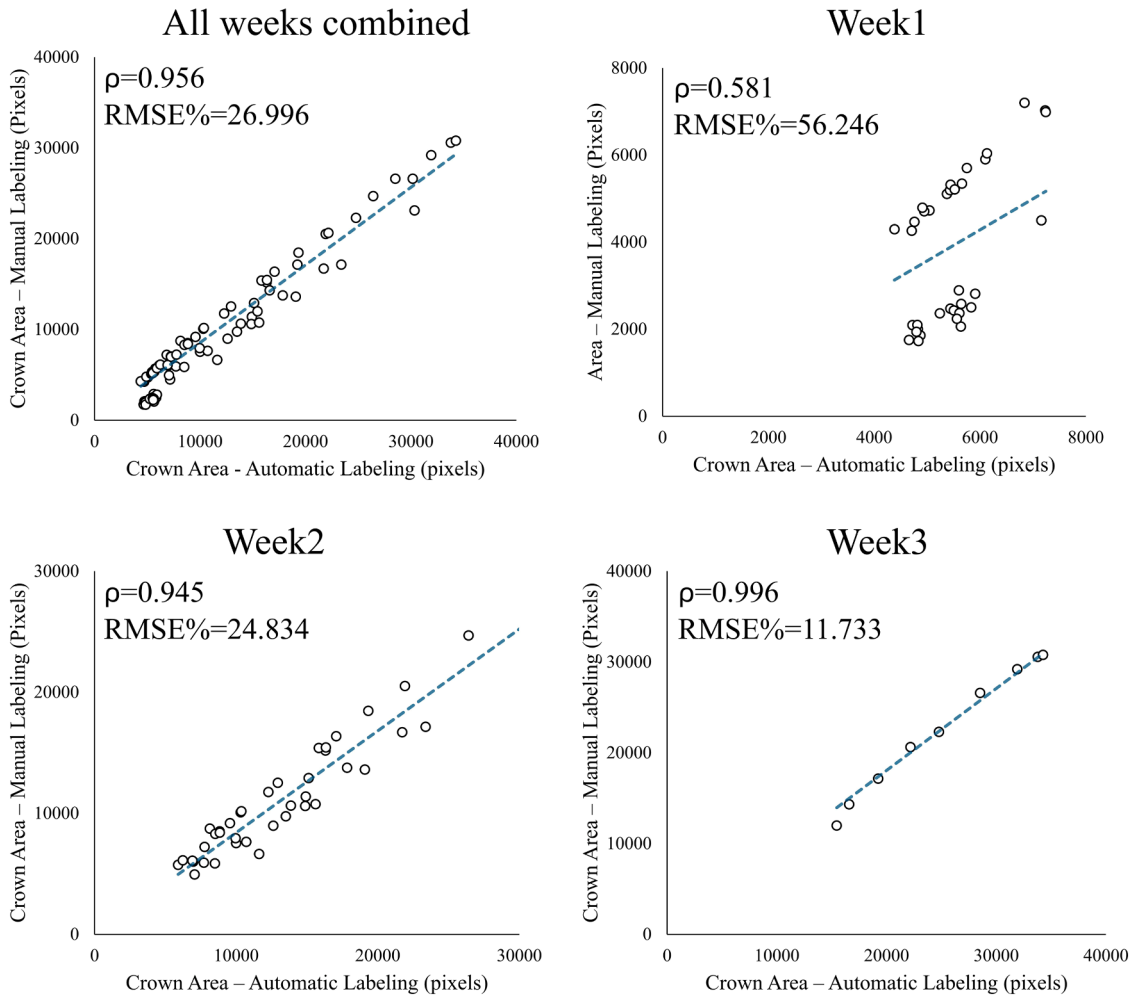

Fig. 4. Spearman correlation (ρ) and Root mean square error percentage (RMSE%) between automatic (Grounding DINO + SAM) and manual labeling of baby Kale crown area.

3.3. Performance analysis of automated labeling for baby kale segmentation

The results were summarized in Fig. 4, where we analyzed the data both cumulatively (all weeks combined) and separately for each week. All weeks combined showed a strong agreement between automatic and manual labeling, with $\rho = 0.956$ and RMSE% = 26.996 %, indicating high consistency across the dataset. Week 1 exhibited the lowest correlation ($\rho = 0.581$) and the highest error (RMSE% = 56.246 %), suggesting significant discrepancies between manual and automatic annotations during this period. Week 2 presented an improved correlation ($\rho = 0.945$) and a lower RMSE% (24.834 %), indicating better agreement between the two methods. Week 3 demonstrated the highest correlation ($\rho = 0.996$) and the lowest RMSE% (11.733 %), confirming a strong alignment between the two approaches in the later stage (Table 1).

We applied the Wilcoxon signed-rank test to assess whether there was a significant difference between manual and automatic crown area. Since the p-value was significantly lower than the 0.05 threshold, the null hypothesis was rejected, indicating a statistically significant difference between the two datasets.

4. Discussion

Developing multimodal AI models in agriculture that perform object detection and segmentation without the need for extensive training or annotations is a challenging endeavor. While current research primarily focuses on models that still require some level of training and annotated data to function effectively, fully eliminating the need for these elements is not yet feasible. Nonetheless, research is progressing toward reducing these requirements, with the goal of creating models that can operate effectively with minimal labeled data, thereby enhancing their applicability in diverse agricultural settings. For instance, a study proposes an innovative approach that integrates image, text, and sensor data using deep learning technologies. This method addresses challenges in agricultural disease detection and question-answering systems, providing new perspectives and tools for the development of intelligent agriculture. While this approach enhances the efficiency of data utilization, it still relies on a certain amount of training data to achieve high accuracy [11]. Another study explores case studies where combining techniques and data sources accelerates progress in personalized nutrition and invasive pest detection. The study highlights challenges such as obtaining high-quality training data and the need for machine learning techniques customized for agricultural use. This indicates ongoing efforts to reduce the dependency on extensive annotations, though some level of labeled data remains necessary [23].

In this study, we employed an entirely automatic pipeline for plant detection and segmentation, eliminating the need for manual annotations and additional model training. The Grounding DINO + SAM framework was applied directly to raw images without any prior fine-tuning, demonstrating its adaptability to real-world agricultural data. Since each image contains multiple plants, an automated selection process was implemented to extract only the central plant. The selection criteria were based on identifying the bounding box closest to the center of the image.

From Week 1 to Week 3, an increase in the plant's size was evident. The segmentation masks generated by SAM effectively capture the morphological changes, highlighting the model's ability to adapt to different plant growth stages (Fig. 2). The temporal analysis confirms that the proposed method successfully tracks plant growth across different times. The detection accuracy improves as the plant develops, and the segmentation process remains consistent, providing precise masks throughout the three-week period (Fig. 3). This automated pipeline offers a scalable and efficient solution for plant monitoring applications in precision agriculture.

Table 1
Statistical validation results for automated and manual labeling of baby Kale crown area.

Week	Spearman Correlation (ρ)	RMSE%
Week 1	0.581	56.246
Week 2	0.945	24.834
Week 3	0.996	11.733
All Weeks	0.956	26.996

The validation of the automatic labeling model revealed significant differences when compared to manual labeling, as indicated by the Wilcoxon test. The result suggests that the two methods do not produce identical results. However, this does not necessarily invalidate the applicability of automatic labeling for agronomic studies. Instead, it highlights the need for a careful interpretation of the results and potential adjustments depending on the specific application.

The correlation analysis provides further insight into the model's performance across different weeks. When all weeks are combined, the Spearman correlation coefficient ($\rho = 0.956$) indicates a strong association between the manual and automatic measurements, with an RMSE % of 26.996. This suggests that, despite some discrepancies, the automatic labeling method follows a similar trend to manual measurements and may be useful for general crop monitoring and trend analysis.

A week-by-week analysis, however, reveals differences in performance. In Week 1, the model shows a relatively weak correlation ($\rho = 0.581$) and a high RMSE% (56.246), suggesting that the automatic labeling method struggles to accurately capture the crown area at this stage. This discrepancy could be attributed to misinterpretation by the segmentation algorithm. As shown in Fig. 3, the algorithm erroneously segmented the plant's growth base along with the plant itself, resulting in an overestimation of the results.

In contrast, the model's performance demonstrably improves during Weeks 2 and 3. Week 2 exhibits a stronger correlation ($\rho = 0.945$, RMSE % = 24.834), while Week 3 shows near-perfect agreement between automatic labeling and manual measurements ($\rho = 0.996$, RMSE% = 11.733). Despite the limited data available for this latter growth stage, the consistent trend suggests high reliability.

The practical implications of these findings suggest that automatic labeling data can be utilized in agronomic studies, particularly for monitoring crop development over time. The combined dataset (all weeks) offers a broad overview and may be useful for general assessments and large-scale analyses. However, if higher precision is required, a week-specific approach is recommended, where Week 1 data should be adjusted by image preprocessing, while Week 2 and Week 3 data can be directly applied, probably with minimal correction depending on the case.

Furthermore, the variability in accuracy across weeks raises the question of whether the automatic labeling data can be used for predicting the planting week of kale. Given that each week exhibits distinct statistical properties, it is feasible to develop a predictive model that estimates plant age based on crown area measurements.

Grounding DINO has been applied in agriculture to enhance plant detection and health assessment. In a study, researchers utilized the model to detect plants with an accuracy of 99.994 % in a phenotyping dataset. This high accuracy facilitated effective segmentation of plant regions, leading to more precise NDVI calculations and improved plant health classification [24]. Likewise, Grounding DINO has been employed in the detection and sizing of durian fruits. Its impressive accuracy in detecting durian fruits, combined with the ability to generate high-quality segmentation masks, makes it a powerful toolset for agricultural applications [25].

Furthermore, the integration of Grounding DINO with models like the Segment Anything Model (SAM) has been explored to enhance segmentation tasks in agriculture. This combination aids in the detection and segmentation of crop regions based on arbitrary text inputs, facilitating more efficient analysis of agricultural imagery [26].

While we acknowledge that training data is essential for developing robust models with high-quality labels, we recommend continuing this approach whenever sufficient data is available for both training and validation. Well-trained models generally yield superior accuracy and reliability in agricultural applications. However, in cases where obtaining large annotated datasets is impractical, alternative methods—such as the automatic labeling approach demonstrated in our study—can serve as viable solutions. Nonetheless, these alternatives come with limitations, particularly in terms of predictive precision, and should be used with careful consideration of their inherent trade-offs.

5. Conclusion

The fully automated pipeline provides an approach that requires an understanding of its limitations, particularly in the early growth stages of Kale. The framework can be applied to real agricultural data without prior training and manual annotation, addressing the challenge of data scarcity in agricultural research where data limitations are common. However, accuracy varies across growth stages. The largest discrepancies were observed in the first week, likely due to challenges in differentiating plant structures. As plants matured, segmentation performance improved.

This study provides a foundation for the use of AI-based automation in plant phenotyping and growth analysis. While improvements in early-stage segmentation are needed, the results indicate that multimodal AI models can improve the efficiency of crop monitoring. Future research could refine the methodology by incorporating additional data preprocessing techniques to further improve accuracy. In addition, exploring the potential for using automated tagging data in predictive models, such as estimating plant age based on canopy area measurements, is an interesting area for further investigation.

CRediT authorship contribution statement

Gianmarco Goycochea Casas: Conceptualization, Data curation, Formal analysis, Methodology, Writing – original draft, Writing – review & editing. **Zool Hilmi Ismail:** Conceptualization, Methodology, Formal analysis, Supervision. **Mohd Ibrahim Shapiai:** Conceptualization, Methodology, Formal analysis, Supervision. **Ettikan Kandasamy Karuppiah:** Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.atech.2025.100903.

Data availability

Data will be made available on request.

References

[1] L. Brigato, L. Iocchi, A close look at deep learning with small data, in: Proceedings of the 2020 25th International Conference on Pattern Recognition (ICPR), IEEE, 2021, pp. 2490–2497, https://doi.org/10.1109/ICPR48806.2021.9412492.

[2] A. Hakami, Strategies for overcoming data scarcity, imbalance, and feature selection challenges in machine learning models for predictive maintenance, Sci. Rep. 14 (2024) 9645, https://doi.org/10.1038/s41598-024-59958-9.

[3] R. Schäfer, T. Nicke, H. Höfener, A. Lange, D. Merhof, F. Feuerhake, V. Schulz, J. Lotz, F. Kiessling, Overcoming data scarcity in biomedical imaging with a foundational multi-task model, Nat. Comput. Sci. 4 (2024) 495–509, https://doi.org/10.1038/s43588-024-00662-z.

[4] Tsourounis D., Theodorakopoulos I., Zois E.N., Economou G., 2023. Leveraging Expert Models For Training Deep Neural Networks in Scarce Data Domains: Application to Offline Handwritten Signature Verification. arXiv:2308.01136v1. 10.48550/arXiv.2308.01136.

[5] A. Cravero, S. Pardo, S. Sepúlveda, L. Muñoz, Challenges to use machine learning in agricultural big data: a systematic literature review, Agronomy 12 (2022) 748, https://doi.org/10.3390/agronomy12030748.

[6] M. Albahar, A survey on deep learning and its impact on agriculture: challenges and opportunities, Agriculture 13 (2023) 540, https://doi.org/10.3390/agriculture13030540.

[7] T. Ali, S.U. Rehman, S. Ali, K. Mahmood, S.A. Obregon, R.C. Iglesias, T. Khurshaid, I. Ashraf, Smart agriculture: utilizing machine learning and deep learning for drought stress identification in crops, Sci. Rep. 14 (2024) 30062, https://doi.org/10.1038/s41598-024-74127-8.

[8] I. Attri, L.K. Awasthi, T.P. Sharma, P. Rathee, A review of deep learning techniques used in agriculture, Ecol. Inform. 77 (2023) 102217, https://doi.org/10.1016/j.ecoinf.2023.102217.

[9] M. Woźniak, M.F. Ijaz, Editorial: recent advances in big data, machine, and deep learning for precision agriculture, Front. Plant Sci. 15 (2024), https://doi.org/10.3389/fpls.2024.1367538.

[10] N. Fei, Z. Lu, Y. Gao, G. Yang, Y. Huo, J. Wen, H. Lu, R. Song, X. Gao, T. Xiang, H. Sun, J.R. Wen, Towards artificial general intelligence via a multimodal foundation model, Nat. Commun. 13 (2022) 3094, https://doi.org/10.1038/s41467-022-30761-2.

[11] Y. Lu, X. Lu, L. Zheng, M. Sun, S. Chen, B. Chen, T. Wang, J. Yang, C. Lv, Application of multimodal transformer model in intelligent agricultural disease detection and question-answering systems, Plants 13 (2024) 972, https://doi.org/10.3390/plants13070972.

[12] Liu S., Zeng Z., Ren T., Li F., Zhang H., Yang J., Jiang Q., Li C., Yang J., Su H., Zhu J., Zhang L., 2024. Grounding DINO: Marrying DINO With Grounded Pre-Training For Open-Set Object Detection. arXiv:2303.05499v5. 10.48550/arXiv.2303.05499.

[13] Ren T., Jiang Q., Liu S., Zeng Z., Liu W., Gao H., Huang H., Ma Z., Jiang X., Chen Y., Xiong Y., Zhang H., Li F., Tang P., Yu K., Zhang L., 2024a. Grounding DINO 1.5: Advance the "Edge" of Open-Set Object Detection. arXiv:2405.10300v2. 10.48550/arXiv.2405.10300.

[14] Kirillov A., Mintun E., Ravi N., Mao H., Rolland C., Gustafson L., Xiao T., Whitehead S., Berg A.C., Lo W.Y., Dollár P., Girshick R., 2023. Segment Anything. arXiv:2304.02643v1. 10.48550/arXiv.2304.02643.

[15] Zhang C., Cho J., Puspitasari F.D., Zheng S., Li C., Qiao Y., Kang T., Shan X., Zhang C., Qin C., Rameau F., Lee L.H., Bae S.H., Hong C.S., 2023. A survey on segment anything model (SAM): vision Foundation model meets prompt engineering.

[16] Ren T., Liu S., Zeng A., Lin J., Li K., Cao H., Chen J., Huang X., Chen Y., Yan F., Zeng Z., Zhang H., Li F., Yang J., Li H., Jiang Q., Zhang L., 2024b. Grounded SAM: assembling open-world models for diverse visual tasks.

[17] I. Zauli, E. Rossini, G. Pennisi, M. Martin, A. Crepaldi, G. Gianquinto, F. Orsini, The perfect match: testing the effect of increasing red and blue ratio on baby-leaf kale growth, yield and physiology, Horticulturae 10 (2024) 1134, https://doi.org/10.3390/horticulturae10111134.

[18] A. Korus, M. Witczak, J. Korus, L. Juszczak, Dough rheological properties and characteristics of wheat bread with the addition of lyophilized kale (Brassica oleracea L. var. sabellica) powder, Appl. Sci. 13 (2022) 29, https://doi.org/10.3390/app13010029.

[19] W. Khalid, Iqra, F. Afzal, M.A. Rahim, A. Abdul Rehman, H. Faiz ul Rasul, M. S. Arshad, S. Ambreen, M. Zubair, S. Safdar, A. Al-Farga, M. Refai, Industrial applications of kale (*Brassica oleracea var. sabellica*) as a functional ingredient: a review, Int. J. Food Prop. 26 (2023) 489–501, https://doi.org/10.1080/10942912.2023.2168011.

[20] U. Subedi, S. Raychaudhuri, S. Fan, O. Ogedengbe, D.N. Obanda, Fermenting kale (*Brassica oleracea* L.) enhances its functional food properties by increasing accessibility of key phytochemicals and reducing antinutritional factors, Food Sci. Nutr. 12 (2024) 5480–5496, https://doi.org/10.1002/fsn3.4195.

[21] J.,.R.. Nidamanuri, Deep learning-based prediction of plant height and crown area of vegetable crops using LiDAR point cloud, Sci. Rep. 14 (2024) 14903, https://doi.org/10.1038/s41598-024-65322-8.

[22] Y. Xie, D. Plett, H. Liu, The promise of hyperspectral imaging for the early detection of crown rot in wheat, AgriEngineering 3 (2021) 924–941, https://doi.org/10.3390/agriengineering3040058.

[23] C.S. Parr, D.G. Lemay, C.L. Owen, M.J. Woodward-Greene, J. Sun, Multimodal AI to improve agriculture, IT Prof. 23 (2021) 53–57, https://doi.org/10.1109/MITP.2020.2986122.

[24] A. Balasundaram, A. Sharma, S. Kumaravelan, A. Shaik, M.S. Kavitha, An improved normalized difference vegetation index (NDVI) estimation using grounded dino and segment anything model for plant health classification, IEEe Access 12 (2024) 75907–75919, https://doi.org/10.1109/ACCESS.2024.3403520.

[25] M. Barakat, G.C. Chung, I.E. Lee, W.L. Pang, K.Y. Chan, Detection and sizing of Durian using zero-shot deep learning models, Int. J. Technol. 14 (2023) 1206, https://doi.org/10.14716/ijtech.v14i6.6640.

[26] S. Rana, M. Crimaldi, D. Barretta, P. Carillo, V. Cirillo, A. Maggio, F. Sarghini, S. Gerbino, GobhiSet: dataset of raw, manually, and automatically annotated RGB images across phenology of Brassica oleracea var. Botrytis, Data Br. 54 (2024) 110506, https://doi.org/10.1016/j.dib.2024.110506.