2025 IEEE/SICE International Symposium on System Integration (SII)
January 21-24, 2025, Munich, Germany

Crop Detection Method using Relative Positional Relationships for Small Weeding Robots

Yusuke Iuchi[1], Atsuki Koshigoe[2], Soki Nishiwaki[3] and Takanori Emaru[4]

Abstract— **Automation in agriculture is increasingly critical for addressing global food security and sustainability challenges. This paper presents a novel crop detection method using relative positional relationships, specifically designed for small weeding robots. Unlike traditional approaches that rely heavily on visual characteristics and large annotated datasets, our method leverages the spatial arrangement of plants to distinguish crops from weeds, thereby reducing the dependency on extensive data collection and annotation efforts. We implemented a multi-stage detection system that first identifies all plants using an object detection algorithm and then classifies them based on their positional and size information. Experimental results on soybean datasets demonstrate that our approach achieves AP of 71.3% for soy crops and 27.2% for weeds in environments not included in the training dataset, showing comparable effectiveness to traditional visual-based detection methods in scenarios with limited data. This advancement offers potential for enhancing the adaptability and efficiency of agricultural automation technologies.**

I. INTRODUCTION

Automation in agriculture is essential for achieving both global food supply stability and environmental protection. As population growth and climate change continue to affect food production, there is an increasing demand for technologies that enhance efficiency and sustainability.

In particular, Japan faces significant socio-economic challenges due to its aging population and declining birth rates. Agriculture in Japan has been severely affected by a decreasing labor force, with the average age of farmers now reaching 68.7 years. Since few young people are entering the agricultural sector, the number of available workers continues to decline. As a result, the number of core agricultural workers has dropped from approximately 2 million in 2010 to 1.16 million in 2023. This decline has led to reduced productivity and an increase in abandoned farmland, negatively impacting Japan's food self-sufficiency and local communities. Improving agricultural productivity as a solution to this issue is also expected to contribute to food security and environmental conservation. Therefore, the solutions to this agricultural problem have the potential to address not only Japan's challenges but also global issues related to sustainable agriculture.

(a) Soy image recorded in Hokkaido university

(b) Soy image in CropAndWeed [8] datasets.

Fig. 1. Two images of soybeans of different varieties

To address these challenges, the automation of agricultural tasks has become a critical issue. By implementing automation technologies, it is possible to mitigate labor shortages and improve the efficiency of agricultural operations. In recent years, advances in artificial intelligence (AI) have accelerated research into agricultural automation. In particular, techniques that utilize object detection and classification have been studied, such as detecting leaves, fruits, pests, and diseases[1][2][3][4], as well as applying these techniques to various agricultural tasks [5][6][7]. However, weed control remains a significant challenge. Minimizing pesticide use in weed management is labor-intensive, which underscores the necessity for automation. If crops can be detected accurately, allowing for mechanical weeding of non-crop areas, it could greatly enhance agricultural efficiency.

Specifically, weed control is a physically demanding task, and the development of weeding robots presents a promising solution. However, automating weed control with unmanned ground vehicle (UGV) requires highly accurate crop detection, which presents technical challenges for the precise automation of such intricate tasks.

Previous research on AI for agricultural tasks has often focused on specific crop varieties or individual farms, relying heavily on limited datasets, which restricts their applicability. Even within the same crop species, variations in light conditions and crop varieties can result in different appearances, as shown in Fig. 1. When a model is trained to recognize crops with varying visual features as the same category, it may easily confuse different plant classes. Consequently, the AI's performance may degrade when applied to farmlands in different environments from the experimental fields, or even

*This work was supported by JSPS KAKENHI Grant Numbers 24KJ0262.

[1]Yusuke Iuchi is with Graduate School of Engineering, Hokkaido University, Hokkaido, Japan
convallariamajalis@eis.hokudai.ac.jp

[2]Atsuki Koshigoe is with Graduate School of Engineering, Hokkaido University, Hokkaido, Japan
koshigoe.atsuki.y7@elms.hokudai.ac.jp

[3]Soki Nishiwaki is with Graduate School of Engineering, Hokkaido University, Hokkaido, Japan
nishiwaki.soki.j7@elms.hokudai.ac.jp

[4]Takanori Emaru is with Faculty of Engineering, Hokkaido University, Hokkaido, Japan emaru@eng.hokudai.ac.jp

with different varieties of the same crop.

Plants grow according to the season and exhibit different appearances at each growth stage, posing a technical challenge in accurately representing these variations within datasets. Additionally, there is no fixed pattern in the types of weeds that grow on a farm, and different weeds may emerge in the next season. Moreover, the process of collecting plant data and ensuring its precise annotation demands significant time and resources, making the development of universally applicable crop detection AI an especially challenging task.

Humans, on the other hand, can identify crops to a certain degree even without prior knowledge of what crops are being grown or their visual characteristics. This ability extends even to fields heavily infested with weeds, because humans typically recognize crops based on their organized planting patterns, such as rows, rather than individual features. If it were possible to classify crops and weeds using positional information rather than relying solely on visual characteristics, this could enable AI implementation across various regions and conditions with significantly less data than is currently required. Consequently, an AI model that considers spatial relationships to distinguish between crops and weeds could make agricultural automation more adaptable and practical.

Previous research has primarily focused on detecting plants using visual features extracted by Convolutional Neural Networks (CNNs) [9] or Transformers [10], with little consideration given to the positional relationships within the field. To address this, we provided the neural network with explicit positional information and size data for each detected plant, without categorizing them initially. By calculating the likelihood of these plants forming rows, we classified them as either weeds or crops. In this study, we evaluated the importance of positional relationships in distinguishing between crops and weeds by comparing our position-based algorithm with existing algorithms.

II. RELATED WORKS

A. Object Detection: Diverse Architectures and Strengths

Object detection involves the task of identifying specific objects within an image or video and marking their locations with bounding boxes. Over time, various approaches have been developed for object detection, each with unique architectures and characteristics that cater to different application needs.

B. Single-Stage Detectors

Single-stage detectors perform both object detection and classification in a single pass through the network. Two prominent examples of this approach are the You Only Look Once (YOLO) [11] series and RetinaNet. These models process the entire image at once, allowing them to detect multiple objects simultaneously. The ability to process the image in a single pass is the defining feature of single-stage detectors, making them particularly suitable for real-time applications such as surveillance, autonomous drones, and robotics.

One of the main strengths of single-stage detectors is their speed and efficiency. Due to their lightweight architecture, they can be deployed on devices with limited computational resources without sacrificing too much in terms of performance. For instance, YOLOv8 [12] achieves an optimal trade-off between speed and accuracy.

Similarly, RetinaNet [13] introduces Focal Loss to address the class imbalance problem, allowing it to better detect hard-to-detect objects, especially those that are rare in the dataset.

C. Multi-Stage Detectors

In contrast to single-stage detectors, multi-stage detectors divide the detection and classification processes into several stages. Each stage refines the previous stage's predictions, leading to progressively more accurate object detection. Notable examples of multi-stage detectors include the models from the Region-based Convolutional Neural Networks (R-CNN) [14] family, such as Faster R-CNN [15] and Cascade R-CNN [16]. These models work by first generating a set of rough predictions, which are then refined in subsequent stages to improve accuracy.

Multi-stage detectors are known for their high detection precision. The successive refinement of predictions allows these models to achieve a level of accuracy that is often superior to that of single-stage detectors. This makes them particularly suitable for tasks where precision is critical, such as in medical imaging or other applications where even a slight detection error could have significant consequences. Cascade R-CNN, for example, utilizes multiple regression stages to incrementally improve the detection results, leading to high-precision outputs that are well-suited for such precision-demanding tasks.

D. Transformer-Based Detectors

A recent development in object detection is the use of Transformer architectures. Unlike traditional convolutional neural networks (CNNs), Transformer-based detectors use self-attention mechanisms to model relationships between different parts of the image. This enables these models to capture long-range dependencies and learn more complex patterns, which can be advantageous in scenarios where objects have intricate relationships or are arranged in complex ways within the image.

Transformer-based detectors, such as DINO [17] and Detection Transformers (DETR) [18], offer high precision in object detection tasks. The ability to model long-range dependencies makes these models particularly effective in complex detection scenarios, such as autonomous driving or scene understanding in environments where multiple objects are interacting in various ways. DINO, for instance, leverages the Transformer architecture to capture these dependencies, resulting in a high level of precision that can outperform traditional CNN-based approaches in such contexts.

III. METHOD

In this chapter, we provide an overview of a system that utilizes positional and size information of detected plants as

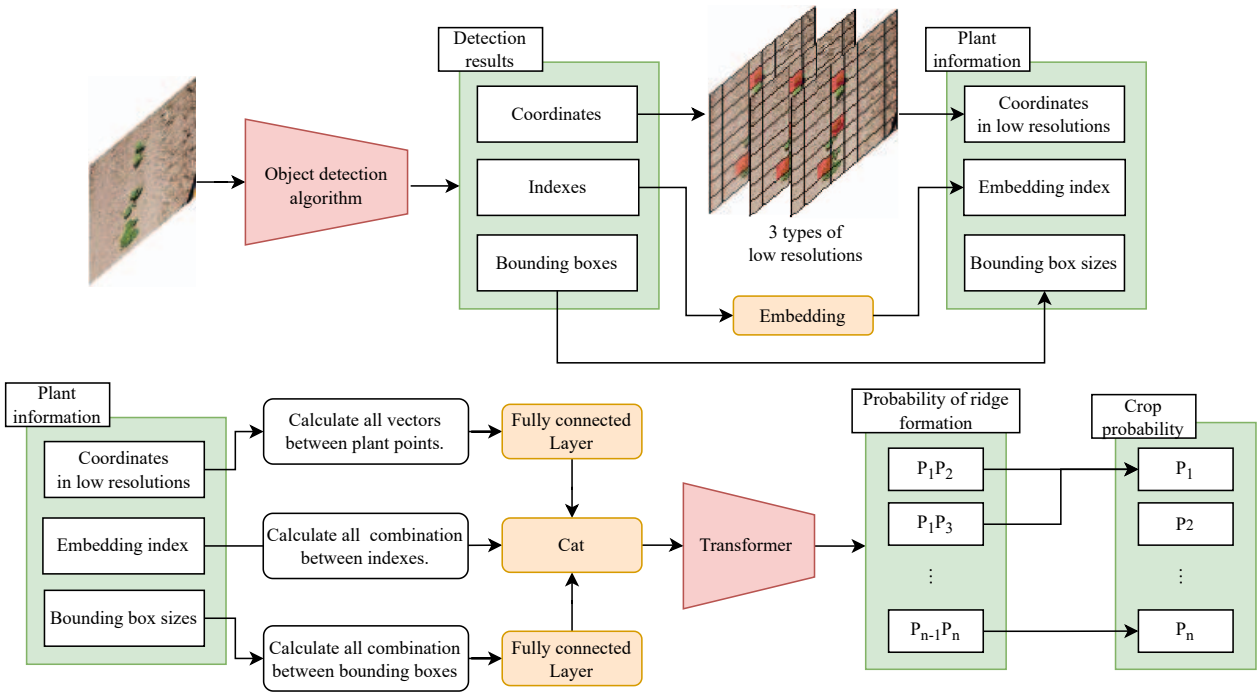

Fig. 2. The diagram illustrates the object detection process, beginning with the detection of plant coordinates and bounding boxes. The plant coordinates are transformed into lower-resolution coordinates, and the sizes of the bounding boxes are calculated. These features are embedded and combined to compute all possible vectors between plant points, which are then processed through a Transformer layer to extract relationships between the point pairs. The resulting features are passed through fully connected layers to estimate probabilities of ridge formation. Finally, class probabilities are calculated based on the extracted features.

input for a neural network. The system is designed to classify plants into crops or non-crops through a series of processing steps, starting with the detection of individual plants. In this study, we used a Multi-Stage approach where classification occurs after the detection of individual plants as shonw in Fig. 2.

First, the object detector is trained using diverse plant data without specifying the target plant species. This approach enables the detection of individual plants in an image based on learned general features, even if the target crop is not included in the training data. The detected positions of the plants are then converted into three types of coordinates by coarsely dividing the image. In this study, the number of divisions was set to 20, 40, and 80.

The plant indices on each coordinate system are embedded. These indices represent the order in which the plants were detected. After embedding, pairs of indices are formed so that all possible combinations of indices are obtained. The embedding layers for these indices utilized PyTorch's embedding layer; however, the parameters of the embedding layer were initialized with a uniform distribution and then fixed, without undergoing training. By passing the index pair information, which was transformed into features through the uniform distribution, through an MLP, pair features were obtained. In this experiment, the embedding layer and the entire MLP were treated as the index embedding layer.

Next, for every combination of plants in the image, vectors between plant pairs are calculated for each coordinate system. The angles of the calculated vectors are then determined. When the UGV crosses ridges, plants are aligned vertically relative to the image, so if the resulting vector is zero, it is considered to be at 0 degrees, which indicates a direction perpendicular to the ridges.

Bounding box sizes are also paired in a similar manner. However, the bounding boxes are normalized so that the largest one in the image is scaled to 1.

The bounding box pairs and vector angle pairs are adjusted by a fully connected layer to match the size of the embedded indices.

By combining the indices, vector orientations, and bounding box sizes into a tensor, we create plant pair information that explicitly includes positional and size information, which is then used as input to the Transformer.

Binary classification is employed on output of the Transformer to determine the probability of whether each plant pair forms a ridge as shown in Equation (3). From the network's output, we count how many times the vector associated with each index is classified as part of a ridge-forming plant pair. The count for each index is then normalized by the maximum count among all indices. Finally, plants with indices whose count exceeds a predefined threshold are treated as crops as described in Equation (6).

$$P_iP_j = \text{BinaryClassifier}(x_{ij}) \tag{3}$$

P_iP_j represents the probability that the plant pair (i, j) forms a ridge. x_{ij} is the feature tensor for the pair (i, j).

$$C_i = \sum_j \mathbb{I}(P_iP_j > \text{threshold}) \tag{4}$$

C_i is the count of how many times index i is part of a ridge-forming pair.

$$C_i^{\text{norm}} = \frac{C_i}{C_{\max}} \tag{5}$$

C_i^{norm} is the normalized count for index i. $C_{\max}$ is the maximum count among all indices.

$$\text{Crop}(i) = \begin{cases} \text{Crop} & \text{if } C_i^{\text{norm}} > \tau \\ \text{Weed} & \text{otherwise} \end{cases} \tag{6}$$

$\text{Crop}(i)$ indicates whether the plant associated with index i is considered a crop or not. τ is the threshold for determining crop status.

IV. EXPERIMENT

We compare the accuracy of object detection using existing methods with that of our proposed method. In this experiment, we trained YOLOv8 with the official repogitory [12] and the others with MMDetection[19].

A. Experiment Overview

Soybeans were selected as the target for accuracy evaluation due to their wide range of applications in Japan, from food products like soy sauce and tofu to medicinal products such as traditional herbal medicines. Additionally, because of their high protein content, soybeans are used as a raw material for alternative meats, making them a highly versatile crop. Therefore, soybeans were chosen as the evaluation target. Furthermore, medicinal soybeans are subject to strict pesticide regulations, necessitating significant manual labor in Japan. The need for effective soybean detection in various environments is another reason they were chosen for this experiment.

For training, we utilized the CropAndWeed [8] dataset alongside additional data collected from various farms in Hokkaido, Japan. The CropandWeed dataset contains over 7,000 images of crops and weeds with annotation, finely categorized by growth stages and crop species. To prevent unnecessary degradation in accuracy, only images containing soybeans and weeds were retained, while those containing other crops were excluded. Additional data collected from within Hokkaido University was also included in the training and validation set. The final training dataset comprised 3,807 images, while the evaluation dataset contained 589 images.

To accurately assess the generalizability of the network, the final evaluation was conducted using data of the same soybean variety as the training data but captured in different years. The data captured in 2024 has slightly different lighting conditions, crop conditions, and soil states due to different shooting dates compared to data captured from 2021 to 2023. The soybean data used for both training and evaluation while model development was captured in Hokkaido between 2021 and 2023, whereas the final evaluation data was captured in 2024. The final evaluation data contains 579 images. For comparison, different detection algorithms, including Cascade R-CNN, YOLOv8, RetinaNet, and DINO, were employed to calculate the detection accuracy for the crops.

For YOLOv8, the model v8x configuration was used, while for the other models, ResNet50 was employed as the backbone, Feature Pyramid Network (FPN) as the neck, and each model's respective detection mechanism was used for the head. The learning rate was set to 0.001, and if the loss diverged, it was reduced by multiplying by -10 until divergence stopped. SGD was adopted as the optimizer, with a learning momentum of 0.9, weight decay of 0.0001, and training was conducted for 100 epochs.

All models were trained using augmentations such as horizontal and vertical flipping and color jittering. All detection confidence thresholds were set to 0.5 in evaluations.

For the proposed method, the YOLOv8 x model was used as the object detection algorithm. The YOLOv8 x part was trained on the same dataset as the existing methods, but the classes for soybeans and weeds were combined into a single plant class. The Transformer used 6 layers and was trained using positional data of soybean, corn and adlay crops, captured in the same fields as the soybean data recorded in 2023. For the positional data training, data augmentation was performed by cropping the edges of images and reconfiguring the crop positions on the low resolution coordinates. The Transformer loss function was binary cross-entropy. Additionally, dummy data was used to align the input dimensions for the Transformer.

In addition to these comparative experiments, we conducted experiments to verify whether the proposed method can detect crops with different planting spatial relationships. For crop images other than soybeans, which were insufficient in quantity for training (approximately 50 images), we performed detection using the proposed method without any additional training, based on comparative experiments with other networks, and evaluated the accuracy. The crops used for verification were potatoes, and soybeans planted with plastic mulch.

B. Result

The experimental results are presented in Table I, and Table II. Figure 3 illustrates a characteristic type of false detection observed with the proposed method, Fig. 4 shows the detection results for other types of crops, and Fig. 5 shows the images of crop detection results using each model.

In weed detection, the proposed method achieved the highest accuracy, resulting in the highest mAP. Moreover, it was possible to detect soybeans or other crops planted at different intervals without any additional training.

As seen in Fig. 3, there were differences in the types of misdetections between classifications based on positional information and conventional object detection methods. The

TABLE I

INFERENCE RESULT OF COMPARISON EXPERIMENT

	AP_{50soy}%	AP_{50weed}% %	mAP_{50}%
Cascade R-CNN	58.3	5.1	31.7
YOLOv8	56.2	11.8	34.0
RetinaNet	71.2	6.5	38.8
DINO	**75.5**	1.0	38.3
YOLOv8 + Ours	71.3	**27.2**	**49.3**

TABLE II

INFERENCE FOR OTHER CROPS WITH PROPOSED METHOD WITHOUT ADDITIONAL TRAINING

	AP_{50crop}%	AP_{50weed}%	mAP_{50}%
Soy with Mulch	69.9	No weed	69.9
Potato	66.8	3.9	35.3

proposed method exhibited unique failures, particularly when a large number of weeds were present alongside the rows.

C. Discussion

In this study, differences in lighting conditions and soil states between the training and evaluation datasets had a significant negative impact on accuracy. Additionally, due to the broad variety of weed species learned from the CropAndWeed dataset, the model may have focused more on learning common features, leading to misclassification of crops as weeds. The presence of a substantial number of weeds smaller than 30 pixels could also explain the generally low detection accuracy for weeds.

The differences in misdetection patterns between conventional and proposed methods indicate that each is not merely a superior version of the other, but rather networks designed to handle different characteristics. In particular, the proposed method experiences a significant decrease in accuracy when the number of crops is two or fewer. Because the method relies solely on angular information and size rather than the distance between plants, it was possible to achieve a certain level of detection for other crops without requiring additional training. Building upon these findings, one potential improvement could be expanding the shooting range. By doing so, the number of crop pairs forming rows can increase, allowing for a higher threshold to be set, which could further improve detection accuracy. Additionally, when the spacing between crops in a field is predetermined, it is possible to enhance detection accuracy by predefining which positions within the camera frame will capture the plants.

Applying this method to multiple rows presents challenges. Crops separated into different rows may not be considered as forming a pair since their vector relationships do not align with the row direction. This necessitates lowering the threshold for how many times a crop pair is identified as forming a row, which can lead to over-detection. In such cases, directly inputting coordinates with vector angle information and bounding box sizes into the Transformer model might be more effective, particularly in situations where generalizability might decrease, but the crop spacing are known information.

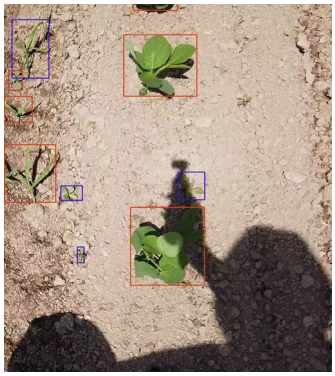

Fig. 3. A characteristic false detection specific to the proposed method

(a) Detection Results for Potatoes

(b) Detection Results for Soybeans with Mulch.

Fig. 4. Detection results for other crops

V. CONCLUSION

From this study, we demonstrated that a network explicitly provided with positional and size information of plants can detect crops with accuracy comparable to that of conventional methods that learn the features of individual crop species. The findings of this study highlight the potential of using relative positional and size information to enhance crop detection models. By integrating these factors with traditional appearance-based features, there is a possibility of developing an object detection model that could be effectively trained with less data compared to conventional crop detection algorithms that require large datasets. Such advancements could make AI applications in agriculture more efficient and widely applicable. Future research should explore the practical challenges of applying these methods across different crop arrangements and field conditions to realize their potential.

ACKNOWLEDGMENT

This work was supported by JSPS KAKENHI Grant Numbers 24KJ0262. The authors would like to express their gratitude to the Japan Society for the Promotion of Science (JSPS) for their generous financial support. We also extend our thanks to all members of our research team and collaborators for their valuable contributions and insightful discussions that greatly enhanced the quality of this work.

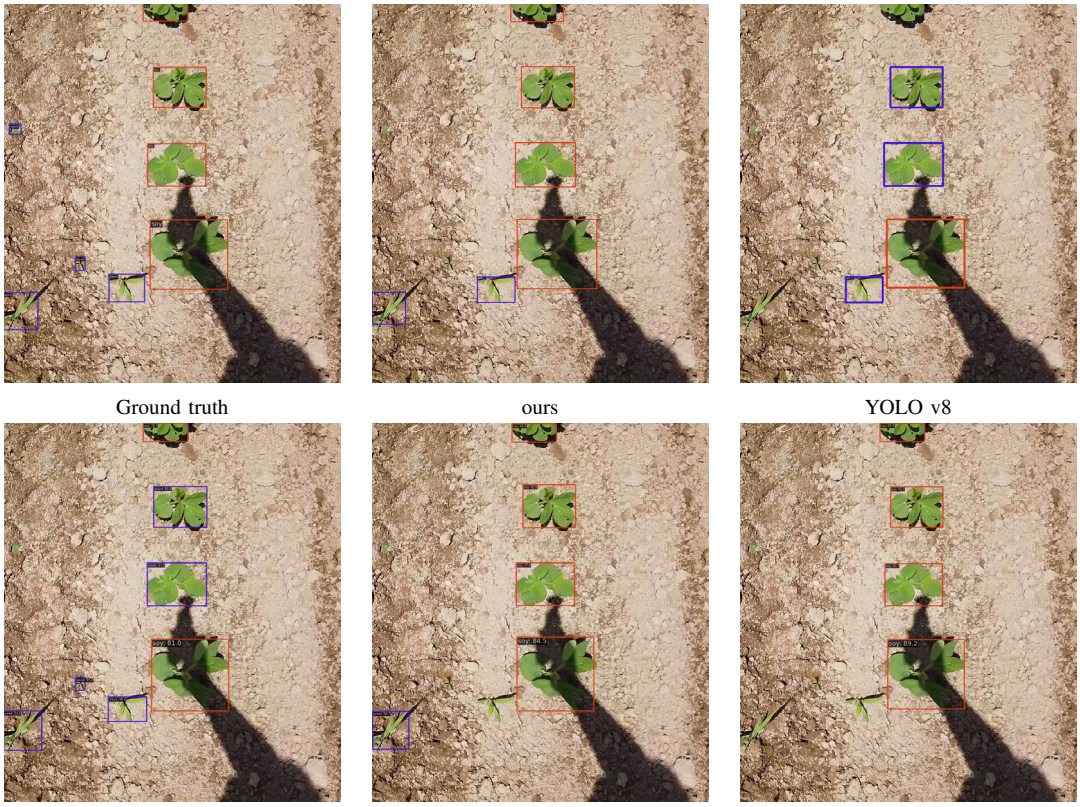

Fig. 5. Qualitative result

REFERENCES

[1] J. Lu, P. Chen, C. Yu, Y. Lan, L. Yu, R. Yang, H. Niu, H. Chang, J. Yuan, and L. Wang, "Lightweight green citrus fruit detection method for practical environmental applications," *Computers and Electronics in Agriculture*, vol. 215, p. 108205, 2023.

[2] T. Chen, H. Li, J. Chen, Z. Zeng, C. Han, and W. Wu, "Detection network for multi-size and multi-target tea bud leaves in the field of view via improved yolov7," *Computers and Electronics in Agriculture*, vol. 218, p. 108700, 2024.

[3] D. Nguyen, A. Tan, R. Lee, W. F. Lim, T. F. Hui, and F. Suhaimi, "Early detection of infestation by mustard aphid, vegetable thrips and two-spotted spider mite in bok choy with deep neural network (dnn) classification model using hyperspectral imaging data," *Computers and Electronics in Agriculture*, vol. 220, p. 108892, 2024.

[4] W. Chen, L. Zheng, and J. Xiong, "Algorithm for crop disease detection based on channel attention mechanism and lightweight upsampling operator," *IEEE Access*, vol. 12, pp. 109 886–109 899, 2024.

[5] X. Jin, X. Zhu, L. Xiao, M. Li, S. Li, B. Zhao, and J. Ji, "Yolo-rds: An efficient algorithm for monitoring the uprightness of seedling transplantation," *Computers and Electronics in Agriculture*, vol. 218, p. 108654, 2024.

[6] Z. Gao, C. Lu, H. Li, J. He, Q. Wang, Q. Wang, Z. Wang, C. Zhai, Z. Zhang, G. Wu, S. Liu, and H. Zhao, "A corn seed spacing detection method based on image stitching and yolox," *Computers and Electronics in Agriculture*, vol. 222, p. 109087, 2024.

[7] B. Wang, Y. Yan, Y. Lan, M. Wang, and Z. Bian, "Accurate detection and precision spraying of corn and weeds using the improved yolov5 model," *IEEE Access*, vol. 11, pp. 29 868–29 882, 2023.

[8] D. Steininger, A. Trondl, G. Croonen, J. Simon, and V. Widhalm, "The cropandweed dataset: A multi-modal learning approach for efficient crop and weed manipulation," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2023, pp. 3729–3738.

[9] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems*, vol. 25, 2012.

[10] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[11] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 779–788.

[12] G. Jocher, A. Chaurasia, and J. Qiu, "Ultralytics YOLO," 2023. [Online]. Available: https://github.com/ultralytics/ultralytics

[13] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 2999–3007.

[14] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Region-based convolutional networks for accurate object detection and segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 1, pp. 142–158, 2016.

[15] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 6, pp. 1137–1149, 2017.

[16] Z. Cai and N. Vasconcelos, "Cascade r-cnn: Delving into high quality object detection," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 6154–6162.

[17] H. Zhang, F. Li, S. Liu, L. Zhang, H. Su, J. Zhu, L. M. Ni, and H.-Y. Shum, "Dino: Detr with improved denoising anchor boxes for end-to-end object detection," 2022.

[18] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *Computer Vision – ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I*, 2020, p. 213–229.

[19] K. Chen, J. Wang, J. Pang, Y. Cao, Y. Xiong, X. Li, S. Sun, W. Feng, Z. Liu, J. Xu, Z. Zhang, D. Cheng, C. Zhu, T. Cheng, Q. Zhao, B. Li, X. Lu, R. Zhu, Y. Wu, J. Dai, J. Wang, J. Shi, W. Ouyang, C. C. Loy, and D. Lin, "MMDetection: Open mmlab detection toolbox and benchmark," *arXiv preprint arXiv:1906.07155*, 2019.