remote sensing

Article

Improved Detection and Location of Small Crop Organs by Fusing UAV Orthophoto Maps and Raw Images

Huaiyang Liu [1], Huibin Li [2], Haozhou Wang [3], Chuanghai Liu [1], Jianping Qian [2], Zhanbiao Wang [4] and Changxing Geng [1,*]

[1] School of Mechanical and Electrical Engineering, Soochow University, Suzhou 215000, China; 20224229025@stu.suda.edu.cn (H.L.); 20245229098@stu.suda.edu.cn (C.L.)
[2] State Key Laboratory of Efficient Utilization of Arid and Semi-arid Arable Land in Northern China, Institute of Agricultural Resources and Regional Planning, Chinese Academy of Agricultural Sciences, Beijing 100086, China; lihuibin@caas.cn (H.L.); qianjianping@caas.cn (J.Q.)
[3] Graduate School of Agricultural and Life Sciences, The University of Tokyo, Tokyo 188-0002, Japan; haozhou-wang@g.ecc.u-tokyo.ac.jp
[4] Institute of Western Agriculture, Chinese Academy of Agricultural Sciences, Changji 831100, China; wangzhanbiao@caas.cn
* Correspondence: chxgeng@suda.edu.cn

Abstract: Extracting the quantity and geolocation data of small objects at the organ level via large-scale aerial drone monitoring is both essential and challenging for precision agriculture. The quality of reconstructed digital orthophoto maps (DOMs) often suffers from seamline distortion and ghost effects, making it difficult to meet the requirements for organ-level detection. While raw images do not exhibit these issues, they pose challenges in accurately obtaining the geolocation data of detected small objects. The detection of small objects was improved in this study through the fusion of orthophoto maps with raw images using the EasyIDP tool, thereby establishing a mapping relationship from the raw images to geolocation data. Small object detection was conducted by using the Slicing-Aided Hyper Inference (SAHI) framework and YOLOv10n on raw images to accelerate the inferencing speed for large-scale farmland. As a result, comparing detection directly using a DOM, the speed of detection was accelerated and the accuracy was improved. The proposed SAHI-YOLOv10n achieved precision and mean average precision (mAP) scores of 0.825 and 0.864, respectively. It also achieved a processing latency of 1.84 milliseconds on 640×640 resolution frames for large-scale application. Subsequently, a novel crop canopy organ-level object detection dataset (CCOD-Dataset) was created via interactive annotation with SAHI-YOLOv10n, featuring 3986 images and 410,910 annotated boxes. The proposed fusion method demonstrated feasibility for detecting small objects at the organ level in three large-scale in-field farmlands, potentially benefiting future wide-range applications.

Keywords: remote sensing data fusion; object localization; georeferencing information; aerial photogrammetry; detection dataset

Academic Editor: Yuxin Miao

Received: 23 December 2024
Revised: 27 February 2025
Accepted: 28 February 2025
Published: 4 March 2025

Citation: Liu, H.; Li, H.; Wang, H.; Liu, C.; Qian, J.; Wang, Z.; Geng, C. Improved Detection and Location of Small Crop Organs by Fusing UAV Orthophoto Maps and Raw Images. *Remote Sens.* **2025**, *17*, 906. https://doi.org/10.3390/rs17050906

1. Introduction

Precision agriculture technologies have significantly enhanced crop production efficiency and resource utilization [1]. UAV-based monitoring is a key tool in precision agriculture, offering high-resolution imagery and non-destructive data collection [2,3]. Additionally, photogrammetry techniques, such as Structure-from-Motion and Multi-View

Stereo (SfM-MVS) [4], are widely used to improve georeferenced analysis in computer vision methods, enabling more accurate large-scale assessments of crop health and distribution [5].

Traditional computer vision methods face challenges such as being labor-intensive and lacking generalizability in large-scale, diversified agricultural phenotypic analyses. For example, feature detection methods like Scale-Invariant Feature Transform (SIFT) [6], Speeded-Up Robust Features (SURF) [7], and Oriented FAST and Rotated BRIEF (ORB) [8] have limited robustness to image noise, occlusion, and complex backgrounds. Template matching techniques [9] rely on pixel-by-pixel similarity but are sensitive to rotation, scaling, and lighting variations. Edge detection methods like Canny and Sobel [10] extract object contours but are prone to false detections in the case of images with complex backgrounds.

With rapid development from traditional computer vision to deep learning, object detection algorithms have transformed complex algorithm design processes into "data-driven" model training workflows [11]. Unlike traditional methods, which rely heavily on manually designed features, deep learning techniques leverage numerous annotated datasets to train models, enabling them to automatically learn relevant features from data. The performance of deep learning depends on diverse, high-quality annotated datasets (e.g., bounding boxes and class labels) [12]. Deep learning can be used to extract multi-level features, from low-level information such as edges and colors to high-level semantic features like crop health and disease detection [13]. Deep learning achieves significant improvements for large-scale and diverse agricultural phenotyping tasks by leveraging diverse and extensive datasets, combined with optimized model architectures and regularization strategies [14]. Additionally, the emergence of manual and semi-automated annotation methods has facilitated data processing for deep learning [15,16]. Despite these advancements, accurately detecting organ-level crop features, such as canopy flowers and fruits, remains technically challenging.

The integrating of raw images and DOMs for organ-level object detection has advanced, but challenges remain in balancing detection precision [17] and geolocation [18]. This study tackles three key issues to enhance agricultural target detection precision and efficiency for small objects:

(1) Low quality issues of DOMs: DOMs provide macro-level insights into the georeferenced distribution of crops [19,20], but they suffer from limitations inherent to the reconstruction process. Blurred edges, seamline distortions, and low resolution often arise due to the use of low-angle images and the geometric corrections requiring interpolation during reconstruction [21–23], as shown in Figure 1a–c with yellow circles. While prior studies have enhanced DOM resolution using preprocessing methods such as super-resolution [24,25] and histogram equalization [26,27], these approaches demand extensive data and often produce unsatisfactory results for small objects [28].

(2) Efficiency optimization for detection models within raw images: Raw UAV images provide higher resolution than DOMs, but organ-level objects like flowers and fruits occupy only a few pixels [29,30], making feature extraction challenging. While some studies have tailored algorithms to small objects [31–33], these adjustments often forgo pretrained weights, requiring training from scratch. Oversampling small objects with segment-wise annotations [34] is incompatible with standard object detection datasets, while cropping and enlarging small-object areas improves detection but increases computational costs. For instance, Yu et al. [35] achieved 3.06 mm georeferencing resolution using a DJI M600-Pro drone with a Sony 7αII camera but at the cost of higher computational demands, reducing practicality. Moreover, agricultural UAV surveys produce thousands of large images, further straining

computational and memory resources [36]. Optimizing raw image detection for organ-level objects remains necessary to enhance accuracy and efficiency.

(3) Missing geolocation information on raw images: Raw UAV images lack precise geolocation data, making it difficult to accurately map detected objects to their actual field position. Yang et al. [37] detected potato canopy leaves in specific areas by manually operating a UAV to capture images at 2 m above the canopy to improve image distribution, but this limits the availability of raw images of specific areas and hinders 3D reconstruction from 2D images. Such methods reduce the practical applicability of detection results in agricultural management [38]. Establishing effective mapping between DOMs and raw UAV images is essential for improving locating precision [39] and enabling georeferencing on orthophoto maps and reverse-geocoding for raw images, thus directly linking the detection results to georeferencing information [40,41].

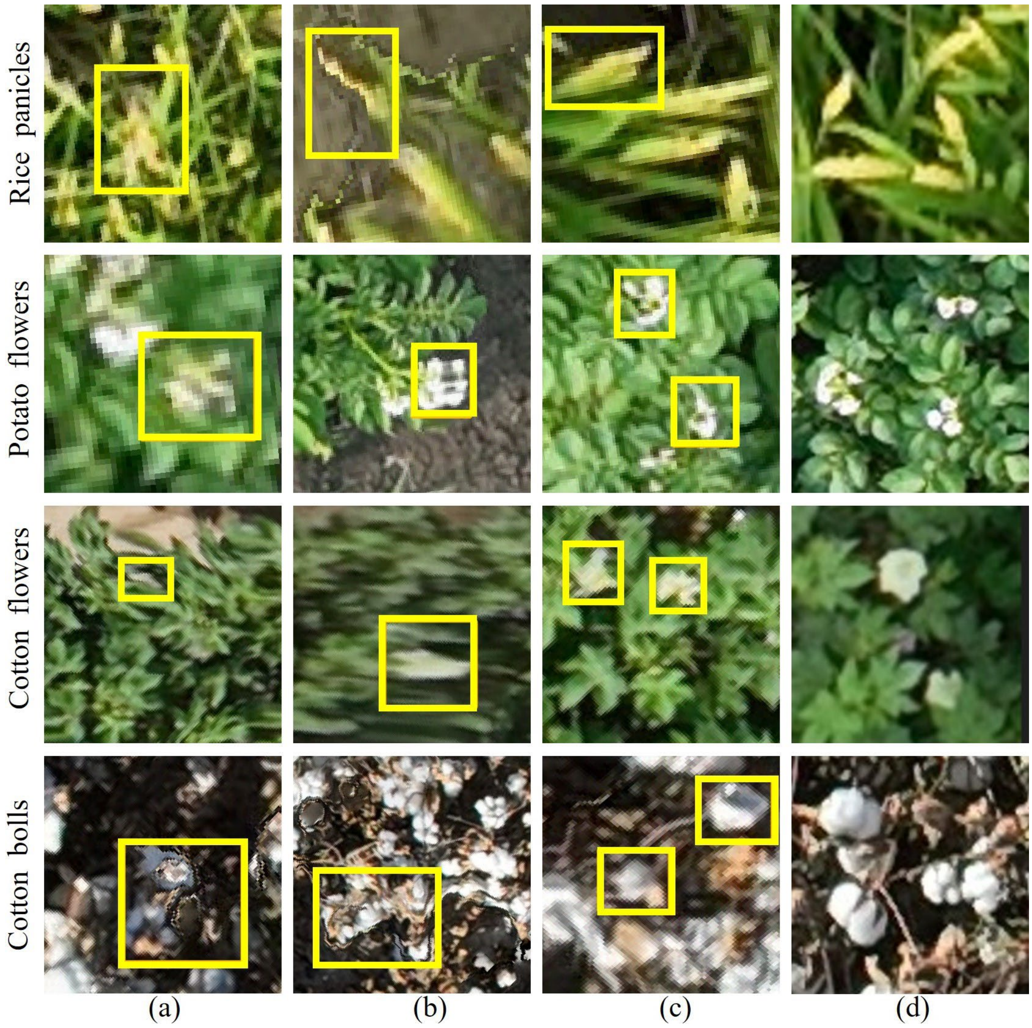


Figure 1. Low quality issues of DOM: (**a**) blurred edges, (**b**) distortions, and (**c**) low resolution; (**d**) raw image.

This study proposes a robust geolocation mapping strategy by precisely linking organ-level objects detected in raw UAV images to georeferencing within the DOM to address the challenges in detecting organ-level objects within crop canopies. Specifically, we integrate the SAHI framework [42] with YOLOv10n to detect objects in UAV raw images, effectively mitigating the issue of missed detections caused by the low quality of DOMs. Once the detection boxes in raw UAV images are obtained, EasyIDP is employed to establish a geographic mapping relationship between raw UAV images and the DOM, thereby addressing the lack of georeferencing information in raw UAV images. This integrated approach enhances both the detection accuracy and spatial localization, providing technical support and practical significance for crop growth monitoring in precision agriculture. In contrast to other methods, our approach directly detects raw UAV images using SAHI, eliminating the computational cost associated with techniques such as super-resolution and histogram equalization for DOM reconstruction. Additionally, SAHI does not participate in the algorithm's training phase, which further reduces training costs. During the inference phase, SAHI is performed on image slices, thereby reducing the memory demands for a single inference. In Section 2.3, we propose an interactive annotation method, which significantly reduces annotation costs compared to manual annotation methods [43,44]. Existing methods, such as manually flying the UAV at lower altitudes [37] or replacing expensive lenses to ensure image clarity [35], result in images that cannot be used to generate a complete DOM with geographical information. Furthermore, the process of generating DOMs often leads to the loss of features from small targets [22,23]. To address these challenges, we introduce EasyIDP, which maps detection boxes from raw UAV images onto the DOM, enabling the acquisition of a complete distribution of target geographical locations within the study area. The specific novelties and contributions of this paper are as follows:

1. Detecting small objects in raw images with higher quality to guarantee the precision and fusion with orthophoto maps to obtain the missing geolocation data;
2. Using SAHI to accelerate the model inferencing for large-scale applications;
3. Using interactive annotation to decrease the workload of data annotation and releasing a small object detection dataset.

The structure of the remaining parts of this paper is organized as follows: Section 2 covers the collection of aerial imagery of crop canopy objects, the creation of the CCOD-Dataset, and the framework for the fusion of DOMs and raw UAV images. Section 3 presents and analyzes the training results of the CCOD-Dataset and evaluates the effectiveness of mapping the detected crop canopy objects from the raw images back to the DOM. Section 4 discusses the detection and mapping results of more complex scenarios involving cotton bolls and rice panicles back to the DOM. Section 5 concludes the paper.

2. Materials and Methods

The proposed method integrates DOMs and raw UAV images to detect and localize organ-level agricultural objects. Figure 2 illustrates the workflow for detecting and localizing these objects, which is divided into five main components: (1) data acquisition and processing; (2) interactive annotation of the CCOD-Dataset; (3) base detection model selection; (4) crop organ detection using the SAHI framework; and (5) georeferencing of the detection results. Sections 2.1 through 2.8 provide a detailed explanation of the implementation process. A brief overview of the workflow is provided below:

1. Raw UAV images were acquired following automated flight planning and processed into a DOM and a Digital Elevation Model (DEM) (see Sections 2.1 and 2.2).
2. The raw UAV images were cropped into 640×640 pixel patches. These patches were then annotated using interactive annotation techniques to efficiently generate the CCOD-Dataset at a low cost (see Section 2.3).
3. The CCOD-Dataset was employed to train an object detection model. Section 2.4 details the selection process, wherein the YOLOv10n model (https://github.com/THU-MIG/yolov10)(accessed on 15 July 2024) is trained for crop organ detection. Additionally, a performance comparison of various detection models is presented in Appendices A and B.

4. The SAHI framework was integrated with YOLOv10n to form the SAHI-YOLOv10n (https://github.com/Nirvana557/SAHI_YOLOv10) (accessed on 11 August 2024) detection framework, which was used to detect objects in the raw UAV images and obtain their corresponding bounding boxes (see Section 2.6).
5. EasyIDP (https://github.com/UTokyo-FieldPhenomics-Lab/EasyIDP) (accessed on 21 August 2024) [45,46] enabled bidirectional mapping between raw images and the DOM, allowing for the precise georeferencing localization of detected objects and distribution analysis of the DOM (see Section 2.7).

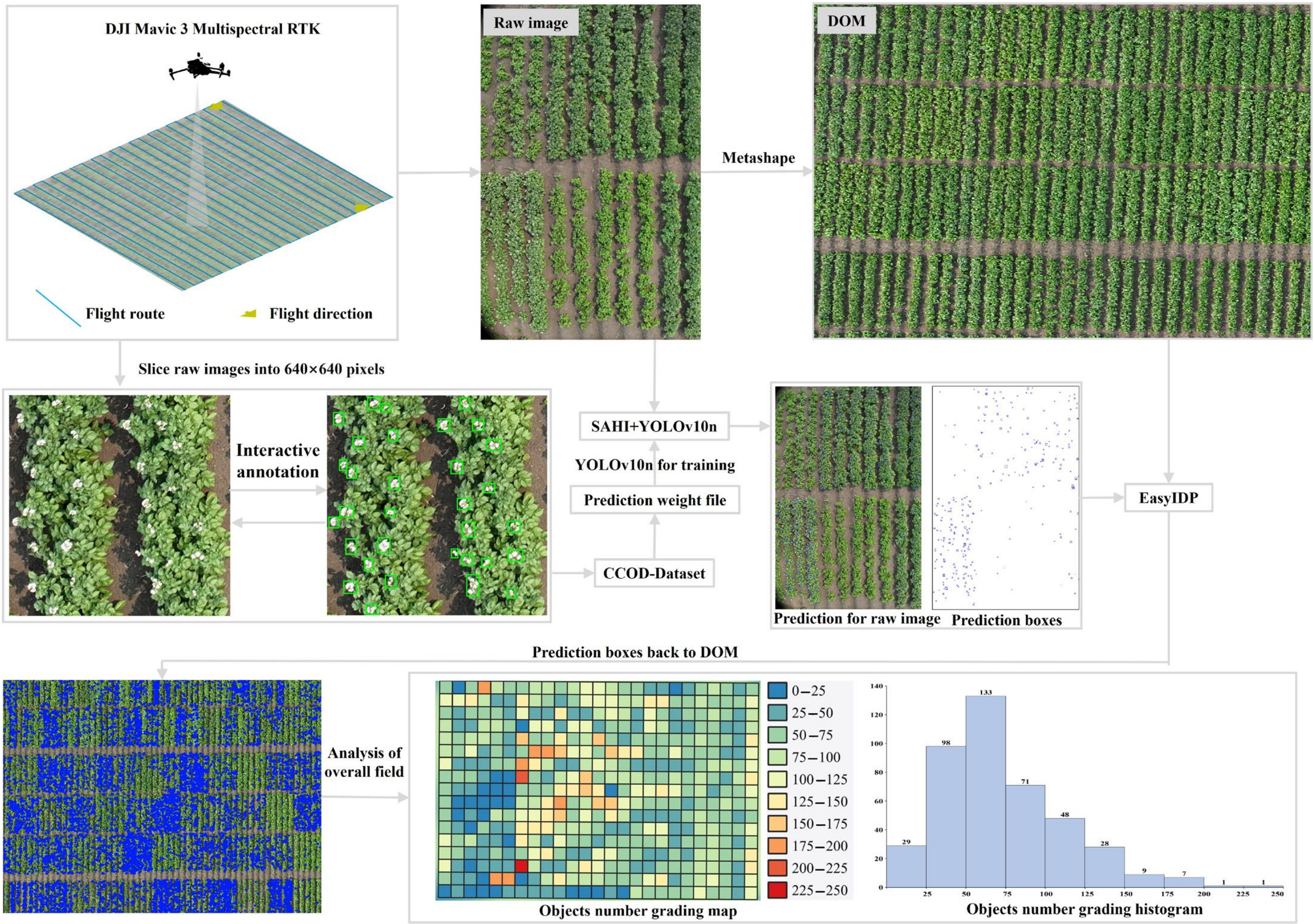


Figure 2. Workflow of detection and location of organ-level objects.

2.1. Experimental Field

The experiments utilized data from three kinds of crop canopy: cotton, potato, and rice. Rice panicle images were collected at the Future Smart Pastoral in Luja, Kunshan, Suzhou, Jiangsu, China (longitude 120.976E, latitude 31.386N, Figure 3c), to validate the method proposed in this study. All data were collected in 2024, specifically, images of cotton canopy flowers and cotton canopy bolls were gathered in the Xinjiang Uygur Autonomous Region, China (44.263N, 87.332E, Figure 3a), while images of potato canopy flowers were collected at the Keshan Farm in Heilongjiang, China (48.033N, 125.864E, Figure 3b).

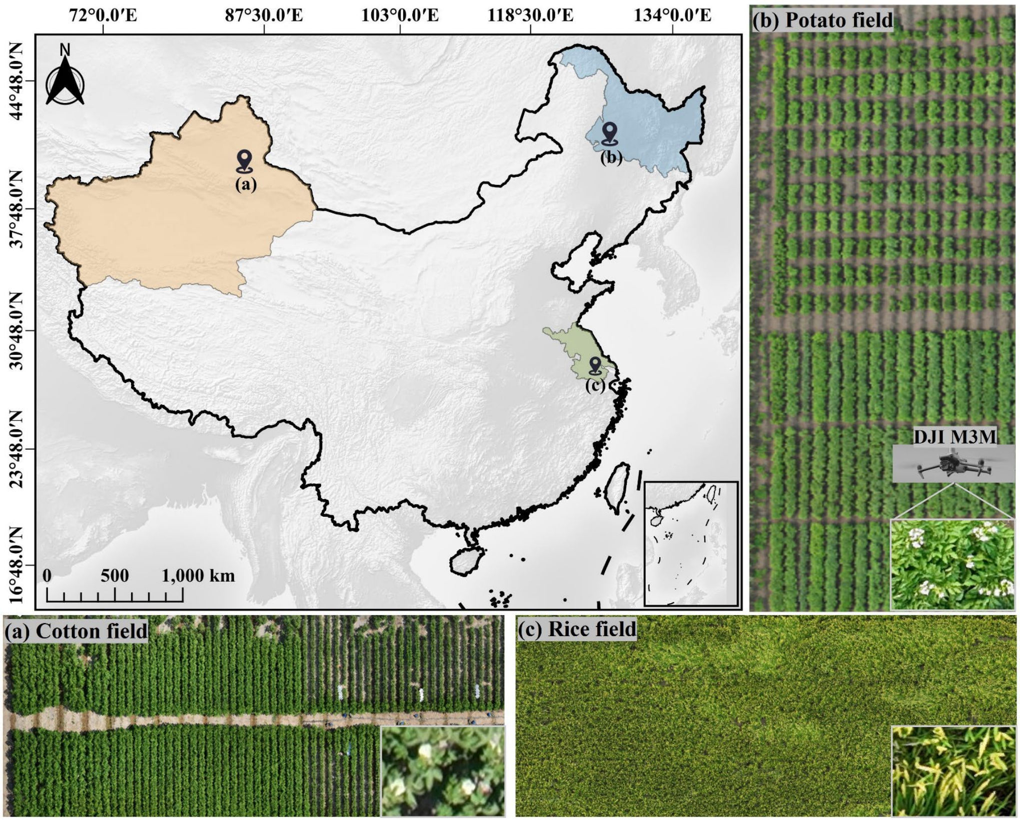


Figure 3. Aerial images captured at three locations.

2.2. Data Acquisition and Processing

The data acquisition locations are shown in Figure 3. The equipment used was a quadcopter unmanned aerial vehicle (DJI Mavic 3 Multispectral RTK, DJI M3M), equipped with a 20-megapixel RGB camera. The flight altitude was set to 12 m, with frontal and lateral overlaps adjusted to 85% and 75%, respectively. The flight routes were automatically planned using DJI Pilot 2, and during the flight, the DJI M3M camera remained perpendicular to the target surfaces. The orthophoto images covered cotton (approximately 50 m in length and 50 m in width), potato (approximately 130 m in length and 85 m in width), and rice (approximately 40 m in length and 80 m in width) fields that served as the research areas of this study. All data were collected during the flowering or maturation stages of each crop. The weather conditions during data collection were clear and windless, with the data acquisition taking place around midday.

2.3. Interactive Annotation of CCOD-Dataset

To reduce the need for manual annotation, rice panicle data from the DRPD dataset [47], which was collected in 2021 and 2022, were utilized in this study. This dataset was used for training purposes. Furthermore, to support the detection of crop canopy organs, the first CCOD-Dataset based on aerial images was constructed in this study by interactively annotating the collected data, including the DRPD dataset.

First, we cropped the aerial orthophotos of the crop canopy organs into 640×640 pixel images. After screening, we ultimately collected 3986 high-quality images. The CCOD-Dataset was subsequently annotated using an interactive approach similar to weakly supervised learning, but more simplistic [48,49]. As is shown in Figure 4a, some images in the CCOD-Dataset were initially annotated using Labelme (https://github.com/wkentaro/labelme) (accessed on 07 June 2024) [50] for training purposes. The trained model was then used to predict unannotated images, after which the predicted bounding boxes were corrected and fed back into the model for further training. This process was repeated, doubling the number of annotated images in each cycle.

To further streamline annotation, we used a pretrained model based on a publicly available dataset containing a small number of cotton boll samples (https://github.com/linzhe67/Cotton_Boll_Counting) (accessed on 12 July 2024) and then repeated the aforementioned prediction, correction, and feedback process. This allowed the annotation of a large quantity of cotton boll data in only two cycles. Examples 1 to 3 in the black box in Figure 4b illustrate the cotton boll annotations achieved with the pretrained model, while examples 1 to 3 in the blue box in Figure 4b demonstrate the predictions made after training on these annotated cotton boll data. With minimal effort, an initial set of annotated samples was obtained. We then refined examples 1 to 3 in the light green box in Figure 4b by correcting annotation errors indicated by red circles and omissions indicated by yellow circles, ultimately yielding high-quality annotated data. Compared to manual annotation, this method decreases the annotation time from 1.5 h per image to an average of 15 min per item.

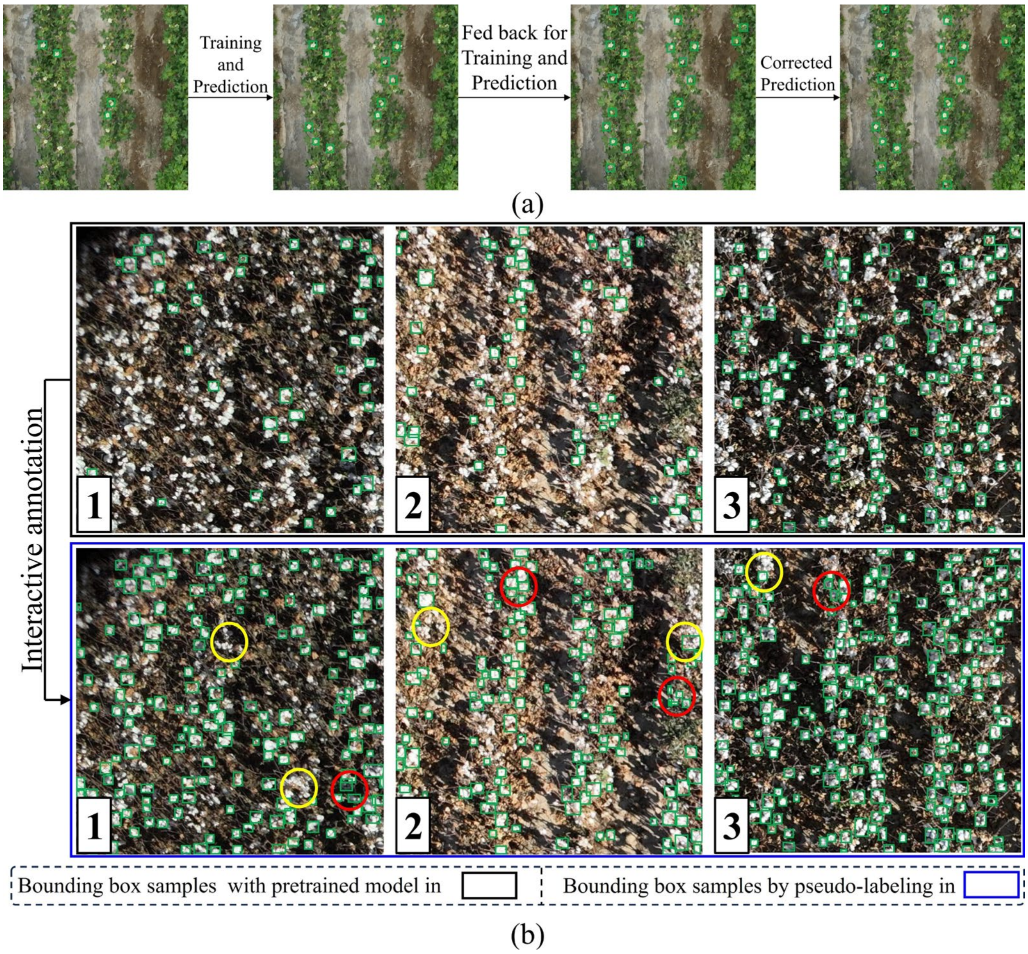


Figure 4. Workflow for constructing the CCOD-Dataset: (**a**) initial annotation and iterative model training using Labelme; (**b**) efficient cotton boll annotation process using a pretrained model, showing annotated and corrected samples. red circles mark annotation errors; yellow circles highlight omissions.

Through this iterative approach, we obtained annotations for all images with minimal manual effort. Finally, we obtained a CCOD-Dataset containing 410,910 bounding boxes. We divided the CCOD-Dataset into a training set, validation set, and testing set in a 7:2:1 ratio for training and validation purposes. The information of the CCOD-Dataset is shown in Table 1.

Table 1. Distribution of crop canopy organs in the CCOD-Dataset.

Crop Canopy Organs	Number of Images	Number of Bounding Boxes
Potato canopy flowers	992	77,034
Rice canopy panicles	994	82,108
Cotton canopy flowers	1000	15,490
Cotton canopy bolls	1000	236,278

The data annotation for the CCOD-Dataset is illustrated as follows: Figure 5a represents potato canopy flowers, Figure 5b shows rice canopy panicles, Figure 5c illustrates cotton canopy flowers, and Figure 5d depicts cotton canopy bolls. The CCOD-Dataset link is publicly available at Hugging Face at https://huggingface.co/datasets/Nirvana123/CCOD-Dataset (accessed on 15 December 2024).

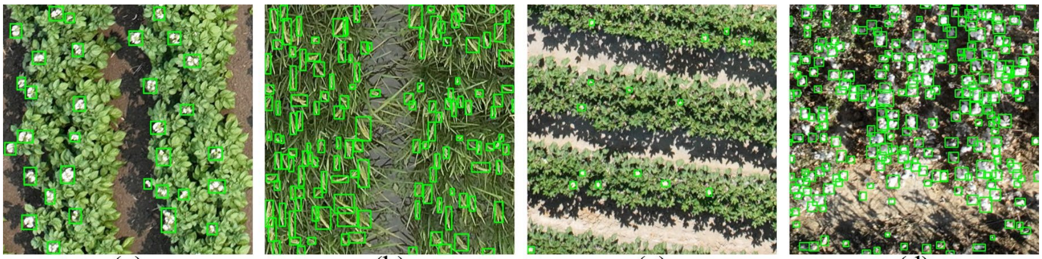


Figure 5. Annotation samples of crop canopy organs in the CCOD-Dataset: (**a**) potato canopy flowers; (**b**) rice canopy panicles; (**c**) cotton canopy flowers; (**d**) cotton canopy bolls. Green boxes indicate the annotation of bounding boxes.

2.4. Base Detection Model Selection

In recent years, advancements in object detection algorithms have markedly improved the precision of small object detection. Convolutional Neural Networks (CNNs) and Visual Transformers (ViTs), as two fundamental neural network architectures, have given rise to a variety of visual detection models, including VGG [51], ResNet [52], MobileNet [53–55], YOLO [56], and DETR [57]. The YOLO series, representing classic object detection approaches, exemplifies the effectiveness of these advancements. Specifically, the YOLOv8 series employs multi-scale detection layers to enhance its ability to recognize objects of varying sizes, while the YOLOv10 series incorporates large-kernel depth-wise convolutions within its Cross Information Block (CIB) and utilizes structural reparameterization techniques in deeper layers, thereby demonstrating superior performance in detecting organ-level objects [58,59]. The detection precision and computational resource requirements of typical algorithms based on CNNs and ViTs were systematically compared in this study to optimize the detection of organ-level objects in crop canopies, and the model that offers the best balance of precision and efficiency was ultimately selected as the foundational detection algorithm.

2.5. Evaluation Metrics

To comprehensively assess detection performance, classical metrics were employed in this research as shown in Equations (1) to (3). Specifically, precision (P)measures the precision of detected objects, recall (R) focuses on assessing the model's ability to identify actual objects, and mean average precision (mAP) provides a comprehensive metric that quantifies detection precision by averaging the precision values across multiple recall levels, offering an overall assessment of the model's performance in detecting objects across different confidence thresholds. The specific formulas for these metrics are as follows:

$$P = \frac{TP}{TP + FP} \quad (1)$$

$$R = \frac{TP}{TP + FN} \quad (2)$$

$$mAP = \frac{\sum_{i=1}^{K} AP_i}{K} \quad (3)$$

In the above metrics, K is the number of crop canopy object types, and True Positive (*TP*) indicates the number of positive samples correctly detected by the model, reflecting the precision of target recognition. False Negative (*FN*) represents the number of objects missed by the model, indicating the extent of actual target omission. True Negative (*TN*) denotes the number of negative samples correctly identified by the model, showcasing its ability to recognize instances that do not belong to the target class. Conversely, False Positive (*FP*) refers to the number of negative samples incorrectly identified as positive, highlighting instances where the model mistakenly identifies a target.

In this paper, "*Ground truth*" represents the annotated reference values, "*Prediction*" indicates the number of predicted bounding boxes, "Omission" denotes undetected objects, and "*Error*" refers to incorrect detections. Based on the definitions, Ground truth satisfies Equation (4), *TP* conforms to Equation (5), *FP* corresponds to Error, *FN* equals Omission, and *TN*, being nonexistent, was uniformly set to 0.

$$Ground\ truth = Prediction + Omission - Error \quad (4)$$

$$TP = Prediciotn - Error \quad (5)$$

2.6. Crop Organ Detection via SAHI

To address the issues of reduced detection precision and high computational costs, SAHI was employed as an auxiliary tool for detecting organ-level objects within UAV images of crop canopies. As illustrated in Figure 6, the original query image P was first sliced into I overlapping patches of size $M \times N$, represented as P_1, P_2, …, P_I, and each patch was resized while preserving the aspect ratio. An object detection forward pass was then independently applied to each overlapping patch. An optional full inference (FI) could be applied to the raw image to detect larger objects. Finally, the overlapping predictions, along with the FI results if used, were merged back to the original size using Non-Maximum Suppression (NMS). During NMS, boxes with Intersection over Union (IoU) ratios exceeding the predefined matching threshold T_m were matched, and for each match, detections with probabilities below the threshold T_d were removed.

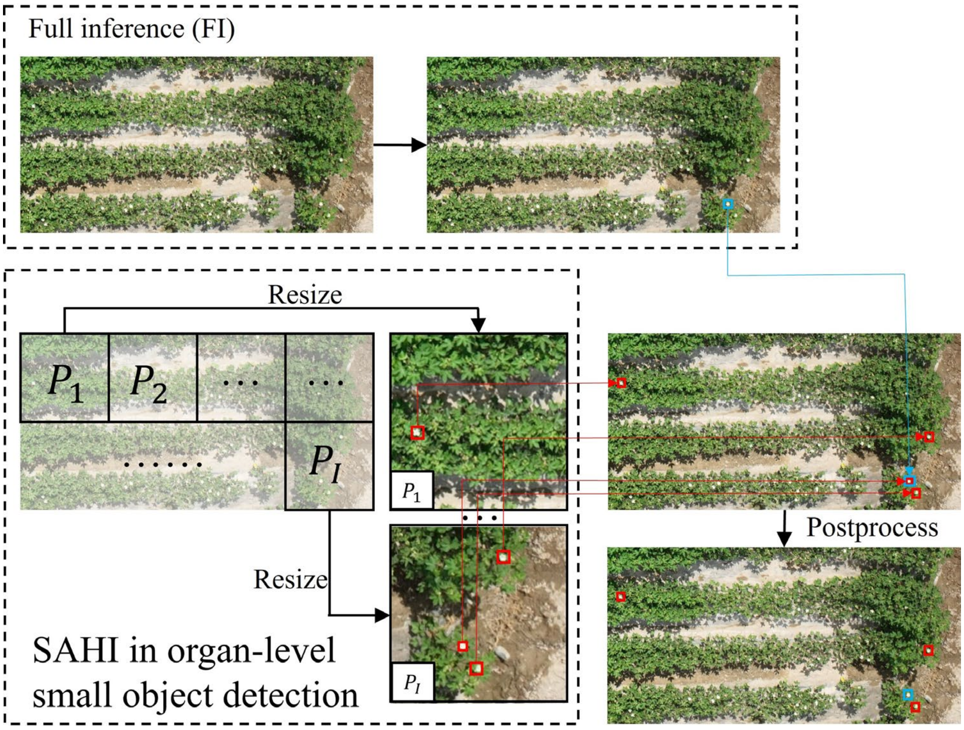


Figure 6. SAHI in organ-level object detection.

2.7. Geo-Mapping of Detections

The ultimate objective of this study extended beyond merely detecting crop canopy objects within raw UAV images; the aim was to accurately map these target locations back onto a digital orthophoto map (DOM). To accomplish this, the EasyIDP framework was employed. As illustrated in Figure 7, a portion of the DOM was selected and divided into nine regions of interest (ROIs), labeled ROI_1 through ROI_9, with ROI_1 used as a demonstration example. EasyIDP's Reversing Module first mapped ROI_1 back to the raw image, where prediction boxes for crop canopy objects within ROI_1 (derived from the SAHI results) were calculated and identified. Then, using EasyIDP's forward mapping function, the ROI, along with its prediction boxes, was mapped back onto the DOM. By repeating this process, the precise locations of all crop canopy objects within the DOM were obtained. Insights into crop growth status and developmental stages were further evaluated based on the distribution density and quantity of these objects.

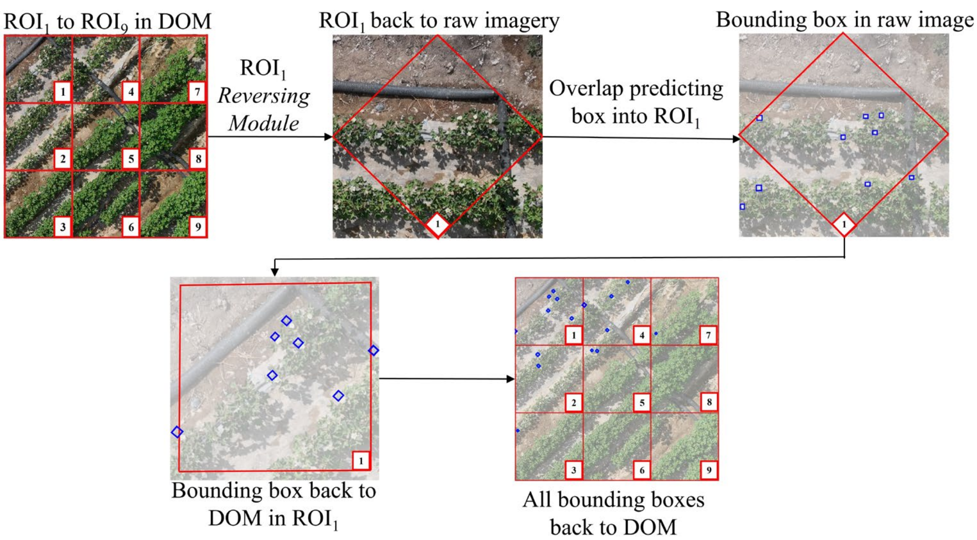


Figure 7. Mapping of crop canopy objects from raw image to DOM using EasyIDP framework.

2.8. Implementation

The hardware platform and software environment used for training and testing are detailed as follows: The experiments were conducted on a system equipped with a Tesla V100-SXM-2 GPU with 32 GB memory, an Intel Xeon(R) Gold 6226R v4 CPU operating at 2.90 GHz with 64 cores, and 256 GB of RAM. The operating system was Ubuntu 20.04 LTS, with PyTorch 1.12 serving as the tensor library and CUDA 11.5 enabling GPU acceleration. Python 3.8 was used as the programming language. Appendix A compares the training parameters across the YOLOv8 to YOLOv10 series, and Appendix B compares the training parameters across the classic models; these comparisons ultimately led to the selection of YOLOv10n as the base detection model in this study.

3. Results

3.1. Comparison of SAHI-YOLOv10n in Detecting Canopy Targets on Raw Images and DOM

Appendix A Figure A1 illustrates the resolution limitations for detecting organ-level objects in the DOM. To further validate and compare the differences in detecting canopy objects on raw images and DOMs, SAHI was applied to detect the four types of crop canopy objects collected in this study. In Figure 8, panels (a), (c), (e), (g) and (b), (d), (f), (h) represent the detection results for cotton flowers, potato flowers, rice panicles, and cotton bolls in the DOM and raw images of the same regions, respectively. Sub-figures (a1) to (h6) in Figure 8 display selected subregions from panels (a) to (h), respectively.

Figure 8a–h illustrate increasing density and complexity in the distribution of cotton flowers, potato flowers, rice panicles, and cotton bolls. In general, cotton flowers in the canopy are the most sparsely distributed and have the smallest target size; however, after defoliation treatment during the cotton maturity phase, the absence of leaves causes cotton bolls to become more densely clustered. Compared to cotton flowers, potato flowers exhibit a higher density with a relatively even distribution. Rice panicles display the most complex distribution pattern due to their slender morphology, presenting varying shapes in the orthogonal view. Overall, SAHI achieves higher detection precision for crop canopy objects in raw images than in the DOM, mainly because the higher image quality of the raw images provides clearer target features within the crop canopy.

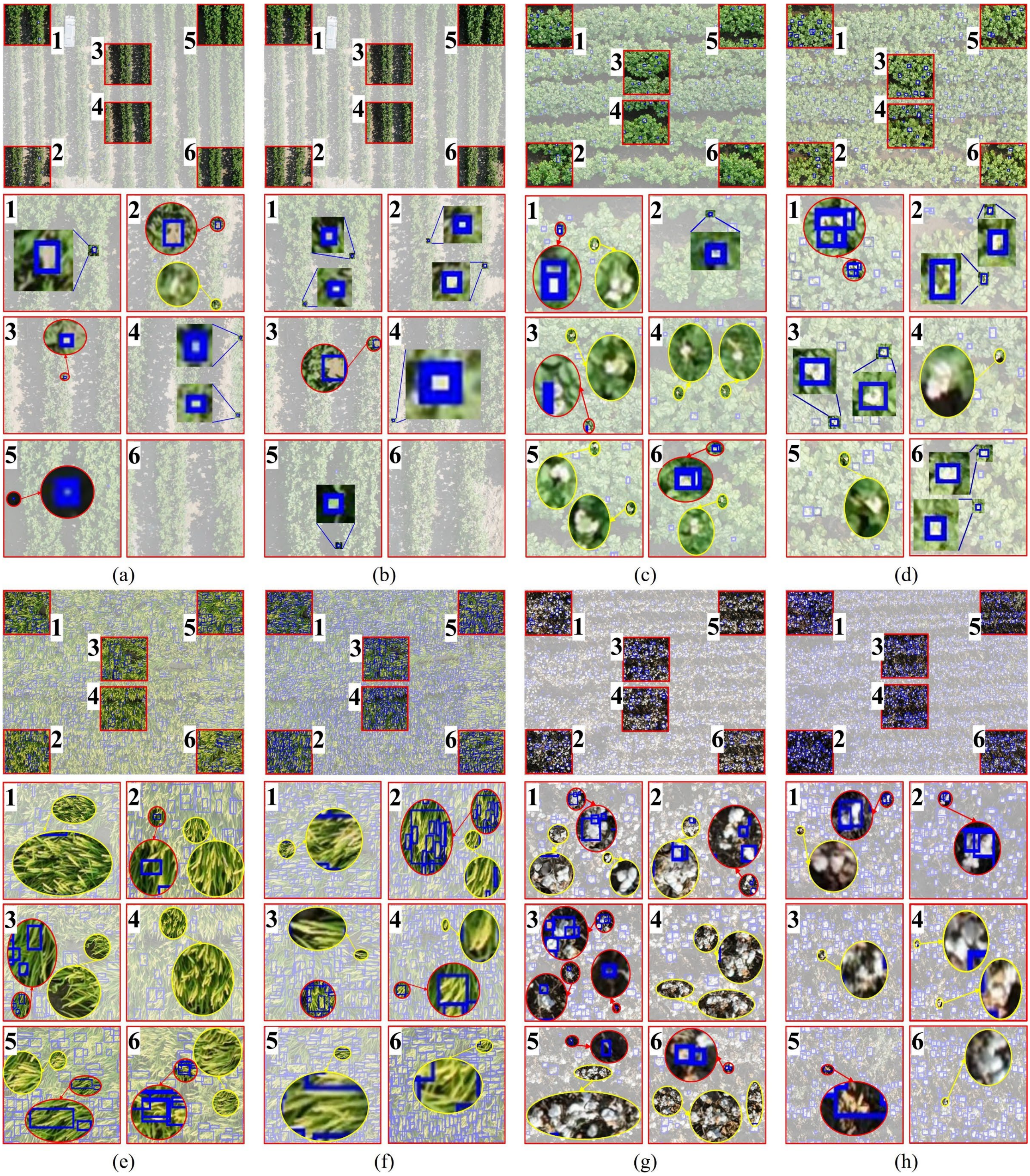


Figure 8. Comparison of detecting canopy objects on raw images and DOM with SAHI: (**a**) DOM and (**b**) raw image of cotton canopy flowers; (**c**) DOM and (**d**) raw image of potato canopy flowers; (**e**) DOM and (**f**) raw image of rice panicles; (**g**) DOM and (**h**) raw image in cotton canopy bolls. Numerical labels 1–6 within each subfigure (**a**–**h**) indicate the selected sub-regions.

Figure 8(a1–h6) illustrate the comparison of crop canopy target detection via SAHI in the DOM and raw images. Red circles indicate false detections, while yellow circles denote missed detections. Through comparative analysis, the following primary causes of false detections were identified:

1. Low quality and image distortion introduced during the DOM reconstruction process. Sub-figures (a2) to (a5) show instances where light spots and fallen leaves are

misidentified as cotton flowers. In sub-figures (c1) to (c5), a single potato flower is mistakenly detected as two, while in sub-figures (e1) to (e6), rice leaves are misidentified as rice panicles. In sub-figures (g1) to (g6), cotton bolls are incompletely detected, and a single cotton boll is sometimes detected as two.

2. These issues are effectively mitigated in raw images; however, occasional duplicate detections occur when crop canopy objects are densely clustered. In such cases, objects detected multiple times may have a larger detection frame encompassing them, mainly because the algorithm misinterprets the combined edges of densely packed objects as a single feature.

The primary causes of missed detections are as follows:

3. Low quality and image distortion in DOM reconstruction lead to blurred features of crop canopy objects. As target complexity increases, missed detections become more pronounced, particularly for rice panicles and cotton bolls, as shown in sub-figures (e1) to (e6) and (g1) to (g6).
4. In raw images, missed detections are mainly attributed to the complexity of the crop canopy objects. Although missed detections are generally less frequent in raw images compared to the DOM, partial omissions of rice panicles and cotton bolls can still be observed in sub-figures (f2) to (f5) and (h4).

We qualitatively analyzed the detection performance of organ-level objects in the DOM and raw images as detailed in the previous Section. Table 2 further quantifies these results by summarizing the number of organ-level objects identified in sub-figures (a1) through (h6) in Figure 8. Figure 9 presents the error matrices for the four types of organ-level objects derived from these data. The analysis of Table 2 and Figure 9 yields the following results:

1. The limited number of cotton and potato flowers in individual sub-figures led to extreme cases where omission (Omission) and error (Error) rates reached 100%, as observed in sub-figures (a3) and (b3). However, such extreme cases do not adequately represent the overall detection accuracy of the DOM and raw images. From an overall perspective, the omission count for cotton flowers in sub-figures (a1) to (a6) was 2, accounting for 22.22% of the total Ground truth, while the error count was 4, accounting for 44.44%. In contrast, the corresponding omission and error rates for cotton flowers in sub-figures (b1) to (b6) were 0% and 10%, respectively. Similar trends were observed for potato flowers: sub-figures (c1) to (c6) exhibited omission and error rates of 27.69% and 4.62%, while sub-figures (d1) to (d6) showed rates of 3.08% and 1.54%, respectively. With the increase in complexity and density of the detection objects, omission and error rates increased significantly, especially in the DOM. For example, the omission rate of rice panicles in DOM sub-figure (e1) reached 88.67%, whereas the highest omission rate in raw image sub-figure (f5) was 21.88%, a reduction of 66.79% compared to the DOM. This analysis highlights that detection based on raw images achieves significantly higher precision.
2. Figure 9 illustrates the normalized error matrices derived from the data in Table 2. The comparison shows that, compared to the DOM, the TP rates for cotton flowers, potato flowers, rice panicles, and cotton bolls in raw images increased by 22.22%, 24.31%, 63.52%, and 52.28%, respectively. Simultaneously, the FP rates decreased by 34.44%, 3.08%, 4.71%, and 6.79%, and the FN rates decreased by 22.22%, 24.61%, 63.52%, and 52.28%, respectively. These results confirm that raw images deliver superior detection accuracy, particularly for organ-level objects with higher complexity and density.

Table 2. The number of organ-level objects identified in sub-figures in Figure 8.

Type	DOM						Raw image					
Cotton flowers	(a1)	(a2)	(a3)	(a4)	(a5)	(a6)	(b1)	(b2)	(b3)	(b4)	(b5)	(b6)
Ground truth	1	5	1	2	0	0	2	3	1	2	2	0
Prediction	1	5	1	2	2	0	2	3	2	2	2	0
Omission	0	1	1	0	0	0	0	0	0	0	0	0
Error	0	1	1	0	2	0	0	0	1	0	0	0
Potato flowers	(c1)	(c2)	(c3)	(c4)	(c5)	(c6)	(d1)	(d2)	(d3)	(d4)	(d5)	(d6)
Ground truth	19	4	13	14	9	6	15	9	11	11	10	9
Prediction	12	4	10	12	7	5	16	9	11	10	9	9
Omission	8	0	4	2	2	2	0	0	0	1	1	0
Error	1	0	1	0	0	1	1	0	0	0	0	0
Rice panicles	(e1)	(e2)	(e3)	(e4)	(e5)	(e6)	(f1)	(f2)	(f3)	(f4)	(f5)	(f6)
Ground truth	256	223	169	202	172	147	214	218	182	212	224	195
Prediction	44	61	49	33	66	51	182	178	156	195	179	166
Omission	227	174	130	173	130	117	39	44	33	23	49	34
Error	15	12	10	4	24	21	7	4	7	6	4	5
Cotton bolls	(g1)	(g2)	(g3)	(g4)	(g5)	(g6)	(h1)	(h2)	(h3)	(h4)	(h5)	(h6)
Ground truth	89	80	87	85	85	91	143	137	147	152	122	157
Prediction	52	51	67	43	27	37	142	140	143	152	120	153
Omission	52	40	33	48	63	59	5	5	7	9	5	10
Error	15	11	13	6	5	5	4	8	3	9	3	6

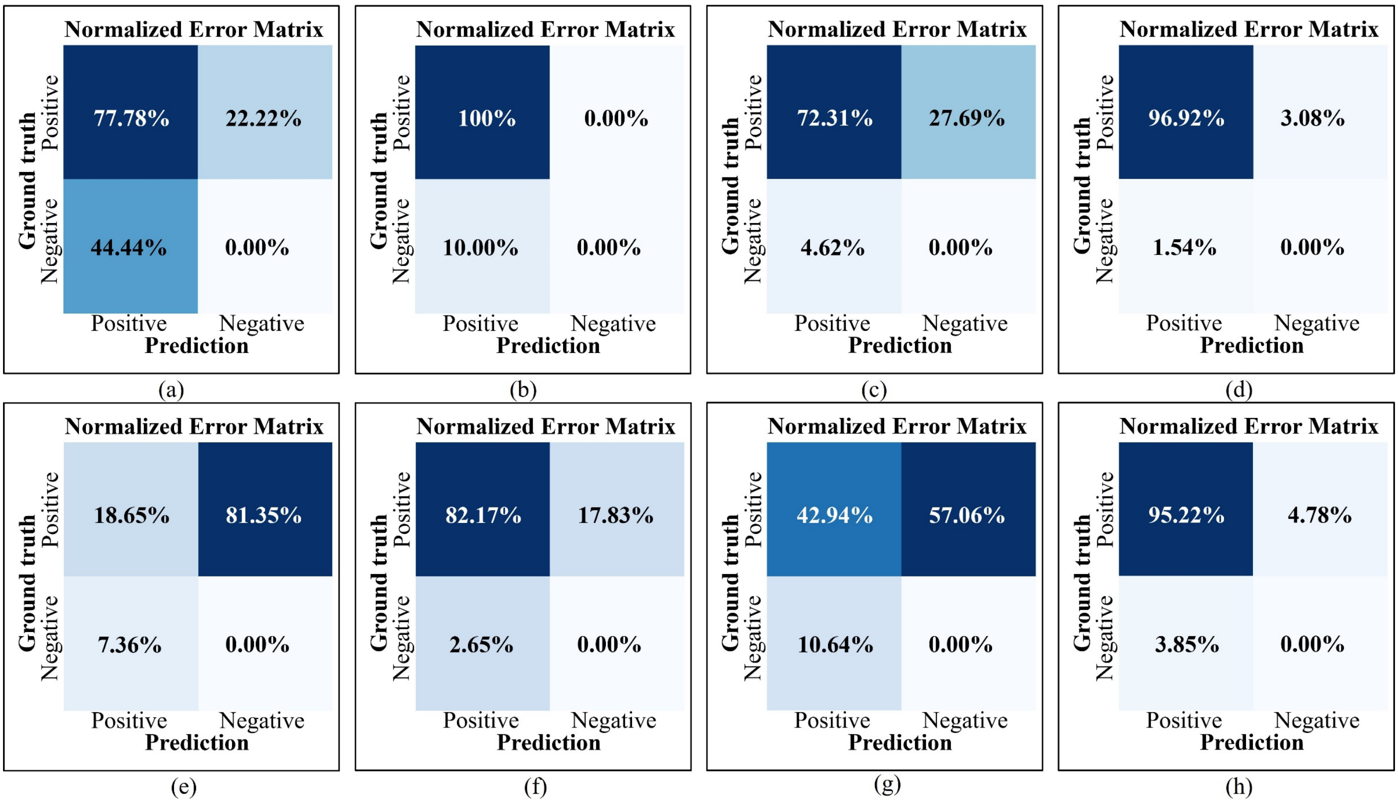


Figure 9. Normalized error matrix of Table 2: (**a**) DOM and (**b**) raw image of cotton canopy flowers; (**c**) DOM and (**d**) raw image of potato canopy flowers; (**e**) DOM and (**f**) raw image in rice panicles; (**g**) DOM and (**h**) raw image in cotton canopy bolls.

3.2. Evaluating EasyIDP for Mapping Raw Image-Based Detection Results to DOM

In Section 3.1, the detection performance of crop canopy objects in both DOM and raw images was analyzed, with the finding that detection precision was higher for crop canopy objects in raw images. This Section analyzes the effectiveness of EasyIDP in mapping crop canopy target prediction boxes from the raw images to the DOM. Cotton flowers, which exhibit the lowest density distribution, were selected as the analysis case shown in Figure 10 to facilitate a clear and concise analysis of EasyIDP's performance. Figure 10b is derived from the nine ROIs in Figure 10a. Figure 10(b1) to Figure 10(b9) correspond to ROI_1 to ROI_9, with yellow boxes representing the local magnified DOM images and light blue boxes indicating the corresponding regions in raw images. Based on the analysis, the following results were drawn:

1. Mapping displacement: As shown in Figure 10(b1,b4), EasyIDP effectively maps the crop canopy target prediction boxes onto the DOM; however, slight positional deviations are observed. The causes of these minor deviations were analyzed, revealing that the fisheye lens on the DJI M3M UAV introduces lens distortion, which impacts mapping accuracy. This distortion causes unavoidable georeferencing discrepancies when converting ROIs and raw images into geographic coordinates, leading to mapping errors.
2. Compensation for the DOM's resolution deficiencies by EasyIDP: As shown in the yellow boxes in Figure 10(b3,b6,b7–b9), the DOM exhibits resolution defects, as shown in Figure 3, which cause the cotton flower features within the yellow boxes to be severely blurred or even disappear during the reconstruction process. In this case, these features could not be detected in the DOM. However, EasyIDP successfully maps the corresponding cotton flower prediction boxes from the raw image to the same locations in the DOM, effectively compensating for the resolution deficiencies of the DOM.
3. Mapping of objects at ROI edges: As shown in the yellow boxes in Figure 10(b2,b5,b6), cotton flowers are located on the outer edge of ROI_2, the inner side of ROI_5, and the edge of ROI_6. The corresponding prediction boxes are placed around the cotton flowers, indicating that EasyIDP can correctly process cotton flowers located at the edges of ROIs in a DOM and accurately map the corresponding prediction boxes from the raw image back to the DOM.

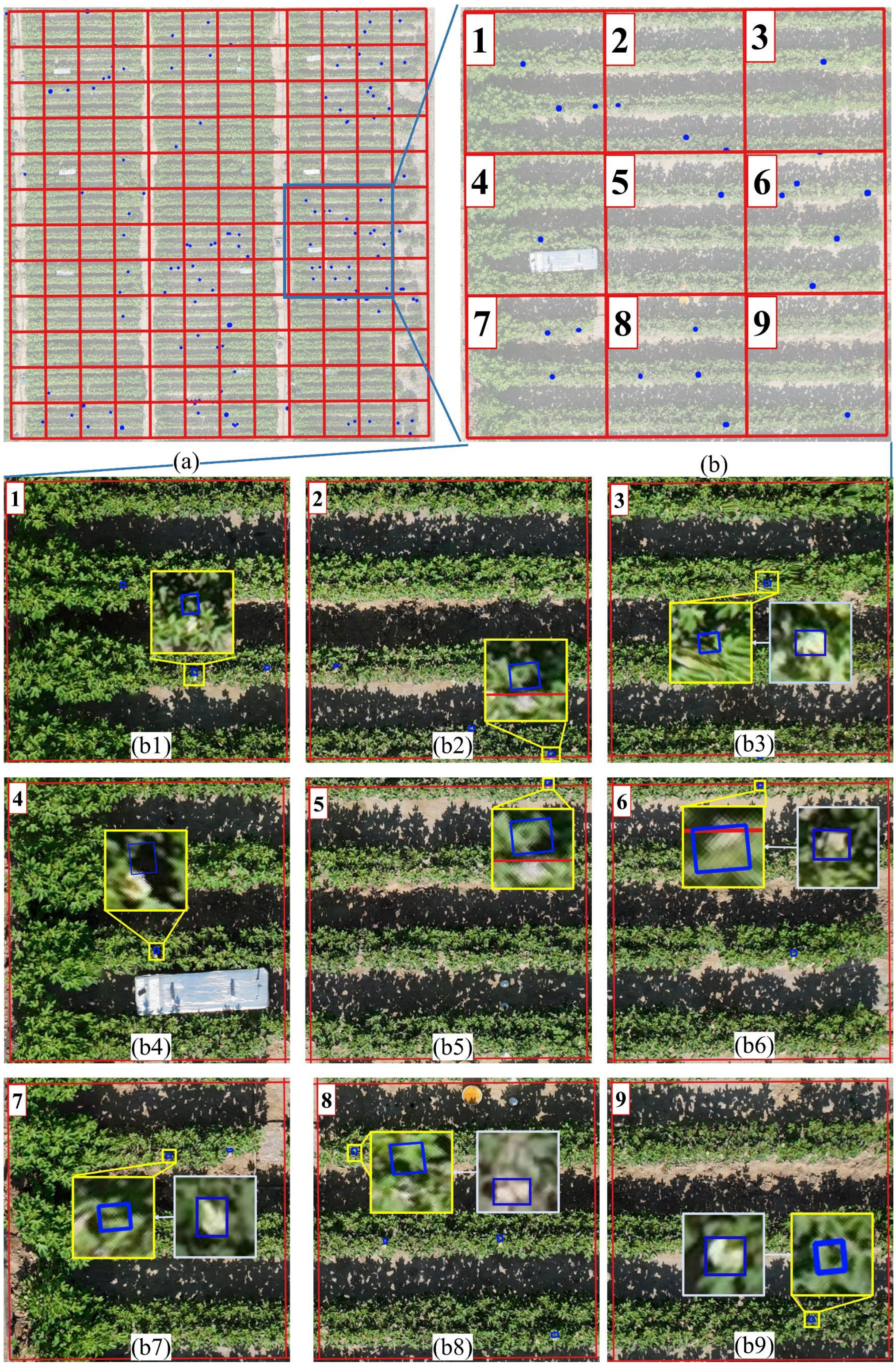


Figure 10. Application of EasyIDP in cotton flower detection: (**a**) overall results; (**b**) sample ROIs, with ROI1 to ROI9 shown in (**b1**–**b9**). The light gray box represents the object in the raw image, which corresponds to the object position in the yellow box in the DOM. (b) Composite image divided into 9 zones, with numerically labeled subregions (1-9) corresponding to subfigures (b1-b9). The numbering follows clockwise from the top-left (1) to the bottom-right (9)

3.3. Analysis of Crop Phenotypic Information Based on Mapped Detection Boxes in DOM

In Sections 3.1 and 3.2, the performance of SAHI and EasyIDP in detecting crop canopy objects was discussed. We selected potato flowers, which exhibit moderate density and sample area, as an analysis case to further illustrate the extraction of crop phenotypic information based on mapped prediction boxes in the DOM. As shown in Figure 11a, we examined a potato field with an area of 10,625 square meters. The x- and y-axes in Figure 11 reference the EPSG:2413-Beijing 1954/3-degree Gauss-Kruger zone 37 coordinate system, with units in meters. From Figure 11a, it can be determined that this canopy region contains 31,385 potato flowers. Figure 11b further indicates significant variability in flower density across different ROIs, which provides insights into potato growth vigor and stages of the growth cycle. This information can, in turn, inform targeted agronomic practices such as optimizing irrigation schedules, adjusting fertilization rates, and predicting harvest times to improve crop yields.

By analyzing the distribution of prediction box counts within each ROI, crop phenotypic information based on mapped prediction boxes in the DOM can be further clarified. As shown in Figure 12a, the number of prediction boxes within each ROI was calculated, grouped at intervals of 25, and visualized as a graded distribution map of potato flower counts. This graded map provides a clear view of the distribution of potato flowers across the sample field. Additionally, Figure 12b presents a histogram of the count distribution across these intervals. For example, using the ROI size in this study ($5m \times 5m$), the area with potato flower counts in the range of (50, 75] covered the largest blooming area at 3325 m^2, while areas within the (150, 250] range occupied only 450 m^2.

Together, the results from Figures 11 and 12 offer a more holistic understanding of the crop growth status and development stages at both macro and micro levels, providing valuable data support for precise agricultural management decisions.

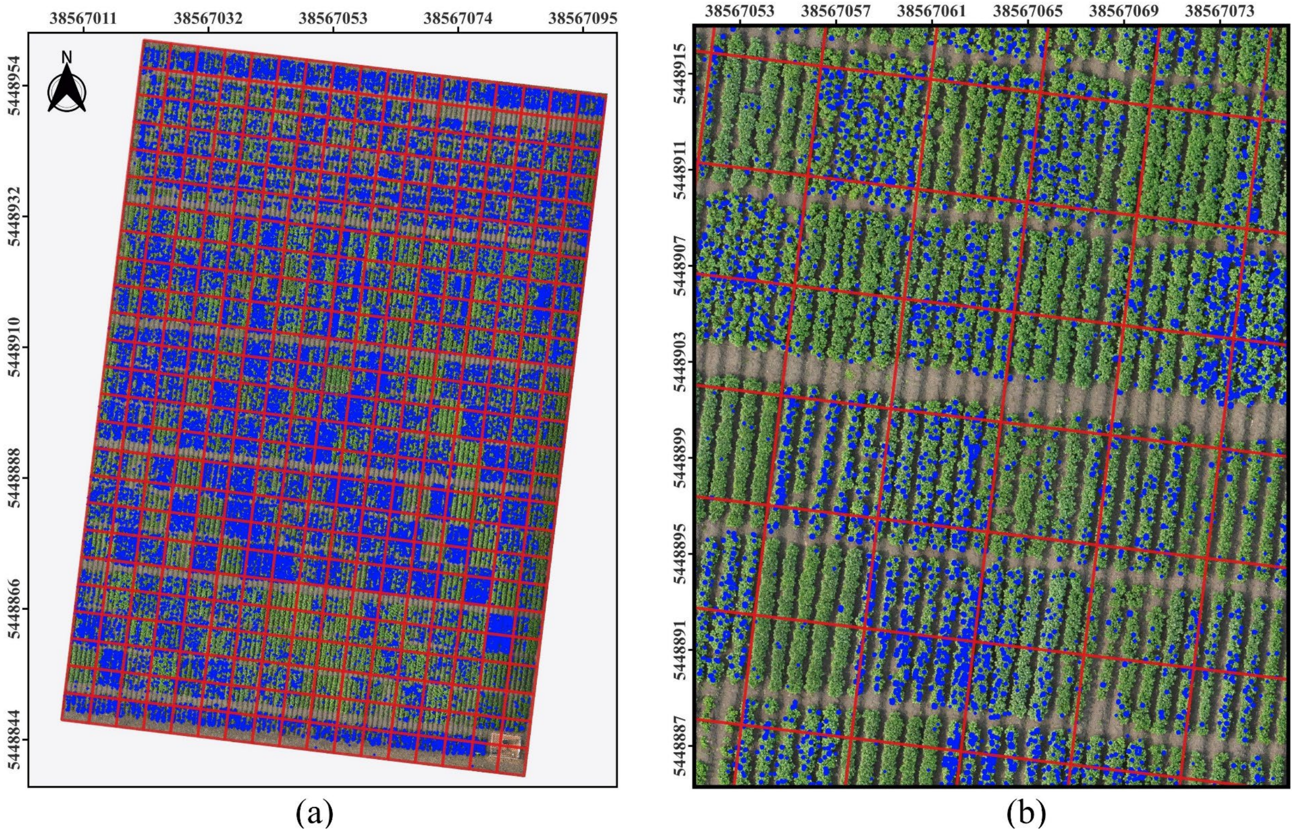
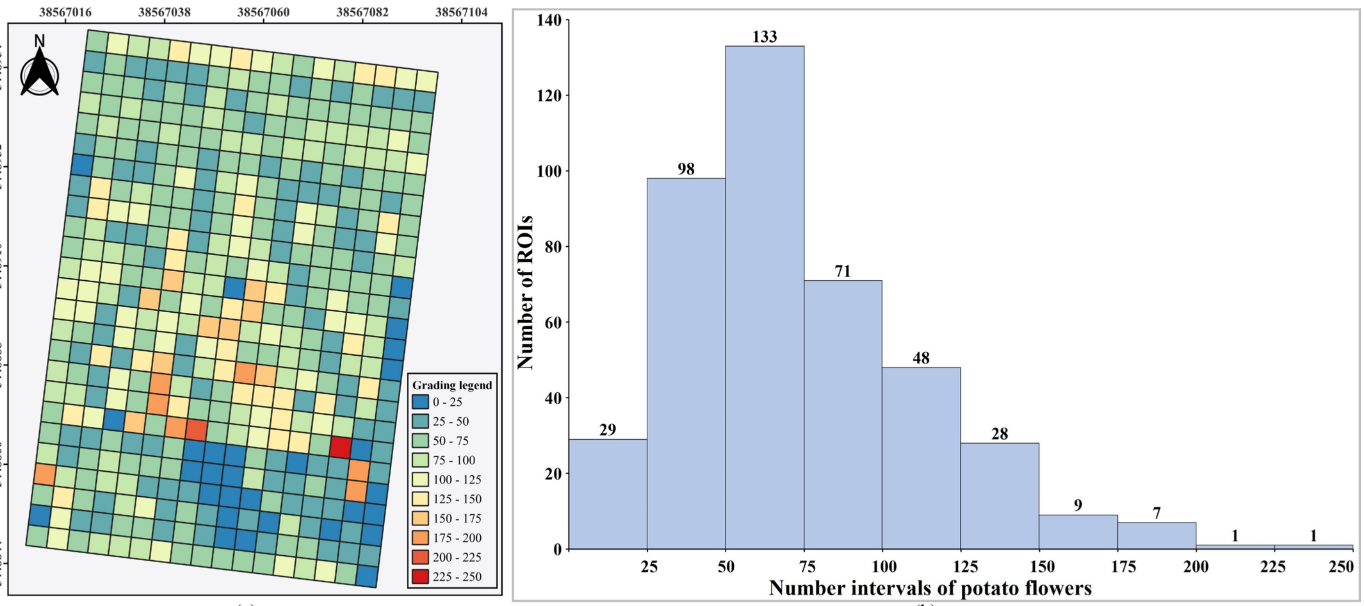


Figure 11. Mapped prediction boxes in the DOM for potato flowers: (**a**) overview of results; (**b**) detailed view of selected DOM regions.



(a) (b)

Figure 12. Analysis of potato flowers by ROIs: (**a**) potato flowers grading; (**b**) histogram of distributions.

4. Discussion

In this Section, the limitations of the proposed method in complex scenarios, as well as directions for future improvements and applications, are discussed. As shown in Figure 13, a high-density cotton boll canopy sample plot in the DOM was selected, and different ROI sizes of 5 m, 4 m, 3 m, 2 m, 1 m, and 0.5 m squares were defined to compare the performance of EasyIDP in mapping prediction boxes from the raw image to the DOM. It was observed that as the ROI size decreased, the computation time for mapping prediction boxes significantly increased, with times recorded as 0.81, 2.17, 3.41, 13.6, 40.14, and 153.91 min for ROI sizes of 5 m, 4 m, 3 m, 2 m, 1 m, and 0.5 m, respectively. Figure 13a illustrates the mapping results for ROI sizes of 5 m, enabling the extraction of cotton boll counts within each individual ROI.

Further extract phenotypic information was extracted from the ROIs. Figure 13b–g present the canopy cotton flower density grading maps for the different ROI sizes. The grading_legend in these figures represents the count of cotton flowers within each ROI. It was found by comparing the density grading maps of varying ROI sizes in Figure 13b–g with the density histograms in Figure 14 that smaller ROIs yield more precise density statistics. Specifically, with a ROI size of 0.5 m in Figure 13g, the density grading map effectively highlights gaps within the cotton fields, while the corresponding density histogram in Figure 14f shows a clearer normal distribution. This observation suggests that smaller ROI sizes produce density grading maps that more accurately reflect the growth patterns of the cotton canopy. As the ROI size decreases, the grading_legend becomes more detailed, further underscoring the advantage of smaller ROIs in capturing the refined density distribution. Consequently, appropriate ROI sizes may be selected in different agricultural monitoring applications according to specific phenotypic information requirements.

The proposed method has been validated for its flexibility, practicality, and ability to effectively address the challenge of obtaining large-scale target geographical information via UAVs. These findings differ from those of other methods, such as that of Peng et al. [60], who used the segmentation algorithm DeNet to generate heatmaps for detecting wheat plant density. Their approach only examined nine raw UAV images per ROI and did not generate a complete geographical distribution map of wheat plants in the study area through the use of the DOM. Similarly, Bai et al. [61] developed RiceNet, which utilizes a dual-branch detection head to detect rice plants' bounding boxes and determine their relative positions in patches from UAV raw images, but it also lacks a comprehensive geographical location analysis. Although Lyu et al. [62] employed improved YOLOv7-tiny for dynamic detection of lotus flowers and seedpods, the use of manual stitching to assemble images of the study area remains highly inefficient. Likewise, Wu et al. [63] improved YOLOv8 to detect apricot flowers in raw UAV images but relied on high-cost cameras to capture super-resolution images, thus increasing computational and labor costs. In contrast, the method proposed in this paper provides a practical approach for detecting and monitoring small crop canopy targets across large areas.

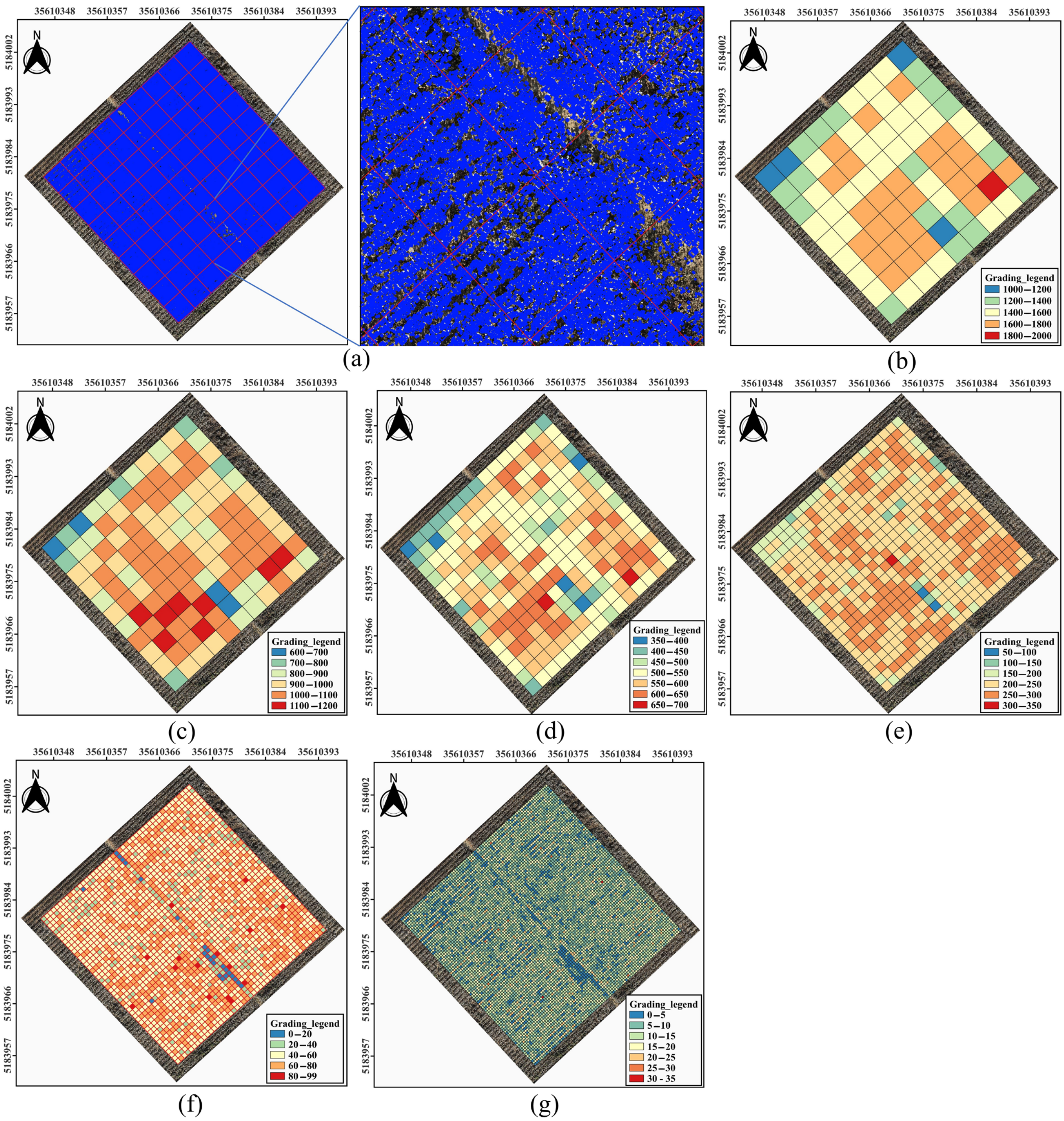


Figure 13. Analysis of cottons bolls by ROIs: (**a**) mapping prediction boxes in DOM; cottons bolls grading in (**b**–**g**).

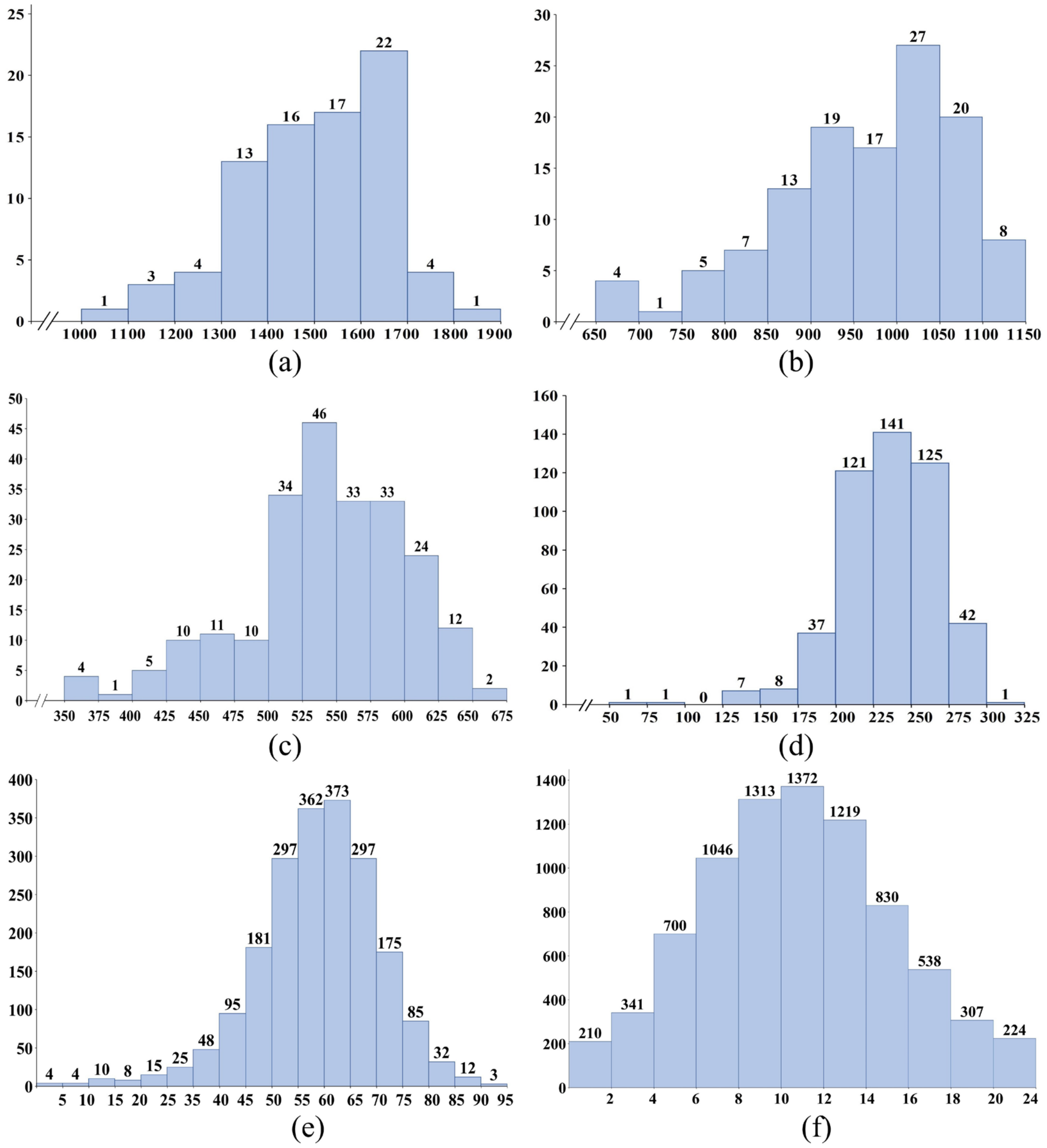


Figure 14. Density histograms for cotton boll detection at different ROI sizes: (**a**) 5 m, (**b**) 4 m, (**c**) 3 m, (**d**) 2 m, (**e**) 1 m, and (**f**) 0.5 m. The horizontal and vertical axes of the coordinate system represent number intervals of cotton bolls and the number of ROIs, respectively.

High-density distribution is one of the complex scenarios in small crop canopy target detection. Additionally, the morphology of crop canopy objects significantly affects detection precision. To examine this, we selected morphologically complex rice panicles for analysis. Figure 15a illustrates the detection of rice panicles in the raw image, where the panicles exhibit slender forms with varying orientation, size, and shape. Figure 15b shows areas with relatively high detection precision, while Figure 15c–f illustrate how rice panicle morphology affects YOLOv10n's detection performance. As shown in Figure 15b, distinct features, high resolution, and relatively isolated panicle distribution contribute to the high detection precision. In contrast, the red-circled rice panicles in Figure 15c have low quality and close proximity, which impairs detection, while those in Figure 15d are less distinguishable due to an unfavorable angle despite the high resolution, highlighting that both resolution and target distinctiveness are critical to YOLOv10n's detection precision. Furthermore, Figure 15e,f reveal another challenge: even with high resolution and clear features, overlapping panicles and close spacing can cause multiple rice panicles to be detected as a single object, leading to inaccuracies in the panicle count. These issues warrant further research for improved detection precision.

In summary, we propose two directions for future research based on this study's detection of crop canopy organs in raw images and mapping of prediction boxes to the DOM as well as the analysis of YOLOv10n's performance in detecting canopy organs. First, the detection algorithm was not specifically improved for morphologically complex rice panicles in this study; future research could focus on developing general algorithms targeting complex canopy organs. Second, the data annotation process still requires manual handling, especially in high-density and complex scenarios, which demands significant time and effort. Thus, developing automated tools for data annotation would be beneficial to enhancing research efficiency.

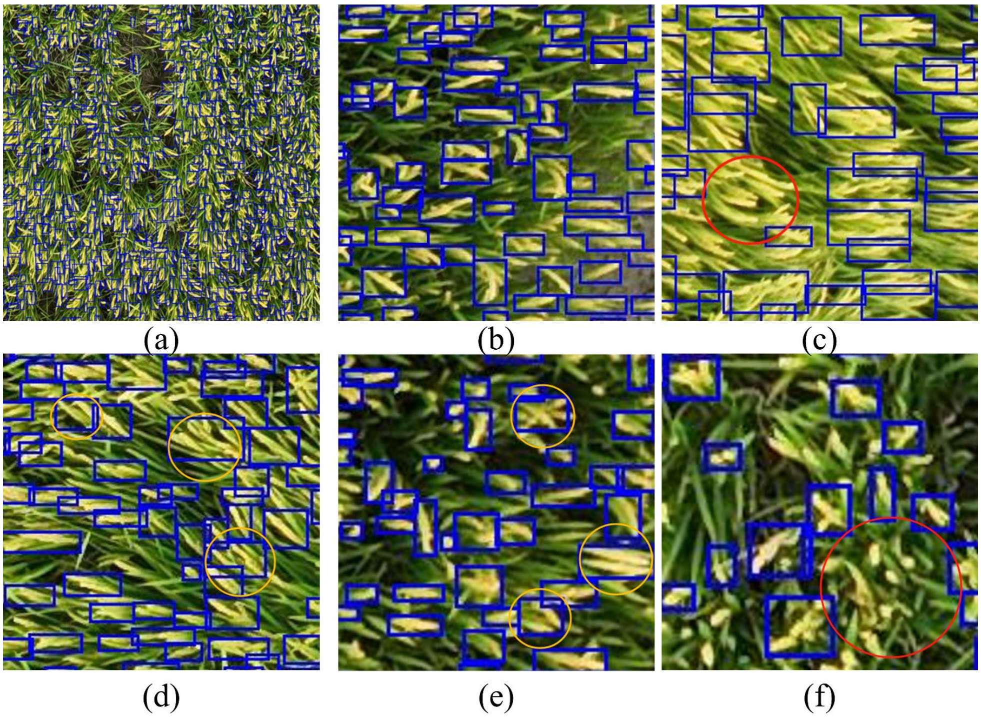


Figure 15. Detection of rice panicles with complex morphologies: (**a**) overall detection in raw image; (**b**) example of clear detection; (**c**–**f**) cases illustrating detection challenges due to resolution, proximity, orientation, and overlapping panicles. The blue boxes indicate prediction boxes, the yellow circles indicate error prediction , the red circles indicate omissions of prediction.

5. Conclusions

This study addresses two key challenges in UAV-based organ-level object detection. The first is seamline distortion and ghost effects in DOMs, which reduce their suitability for detection tasks. The second is the difficulty in obtaining georeferenced information for detected objects in raw images. To solve these problems, we proposed a fusion method. It combines orthophoto maps and raw UAV images using the EasyIDP tool. This approach establishes a georeferencing mapping relationship between raw images and the DOM. Small object detection was performed using the SAHI framework and YOLOv10n. This significantly improved detection speed and precision for large-scale farmlands. The experimental results show that the proposed method outperforms direct detection with DOMs. YOLOv10n with SAHI achieved a precision of 0.825, a mAP of 0.864, and a processing latency of 1.84 milliseconds on 640×640 resolution frames. Additionally, we developed the CCOD-Dataset for crop canopy organ-level object detection. It contains 3986 images and 410,910 annotated boxes, generated through interactive annotation with SAHI-YOLOv10n. The dataset is publicly available to support further research and detection tasks. The proposed method provides a robust solution for detecting and geolocating small organ-level objects in crop canopies. Future work will focus on expanding the dataset to include more crop types, improving detection accuracy for complex objects, and developing a user-friendly canopy target detection tool.

Author Contributions: Conceptualization, H.L. (Huaiyang Liu) and H.W.; Data curation, H.L. (Huaiyang Liu); Formal analysis, H.L. (Huaiyang Liu) and C.L.; Funding acquisition, J.Q. and Z.W.; Investigation, H.L. (Huaiyang Liu) and C.L.; Methodology, H.L. (Huaiyang Liu), H.L. (Huibin Li), and H.W.; Project administration, C.G.; Resources, J.Q. and Z.W.; Software, H.L. (Huaiyang Liu) and H.W.; Supervision, C.G.; Validation, H.L. (Huibin Li) and H.W.; Visualization, H.L. (Huaiyang Liu), H.W., and C.L.; Writing—original draft, H.L. (Huaiyang Liu); Writing—review and editing, H.L. (Huibin Li), H.W., and C.G. All authors have read and agreed to the published version of the manuscript.

Funding: This work received funding support from the following projects: Key Technology Development for Agricultural Condition Parameter Acquisition and Integrated Application of Sensing Equipment by 2022LQ02004, Research on Key Technologies for Sky-Ground Integrated Crop Production Diagnosis and Precision Farming by 2023B02014-2, Research on Field Agricultural Condition Information Acquisition and Agricultural Machinery Operation Decision Technology by CAAS-CAE-202301, Research on Macro-Micro Precision Integrated Navigation and Target Recognition Methods by CAAS-CAE-202302, and the Science and Technology Development Program of the Pilot Zone for Innovation-Driven Development along the Silk Road Economic Belt and the Wu-Chang-Shi National Innovation Demonstration Zone by 2023LQJ03.

Data Availability Statement: The CCOD-Dataset is available at https://huggingface.co/datasets/Nirvana123/CCOD-Dataset, Source code available upon request to corresponding authors.

Acknowledgments: We would like to express our gratitude to the technical staff at Future Smart Pastoral in Luja, the Changji Laolonghe Experimental Base, and Keshan Farm for their invaluable support in land management, fertilization, and other essential tasks, which greatly facilitated the progress of this study.

Conflicts of 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.

Appendix A

This Section presents the training results and relevant metrics of the CCOD-Dataset for the YOLOv8 and YOLOv10 series. To conserve computational resources, an early stopping mechanism was employed, whereby training was halted if the best weight file (best.pth) had not been updated for 100 consecutive epochs. The training was conducted with a batch size of 32 and a learning rate of 0.001, while other parameters followed the default YOLO settings. Appendix Figure A1 illustrates the trend of the algorithm's loss curve during the training process, with all algorithms ceasing training at approximately 350 epochs. The results revealed that the differences among the loss value, mAP@50 value, precision, and recall metrics were minimal. For instance, the precision metric indicated that YOLOv10l achieved a maximum value of 0.829, while YOLOv10b recorded a minimum value of 0.823, resulting in a negligible difference of 0.006. This observation suggested that the detection performance of the YOLOv8 and YOLOv10 series on the CCOD-Dataset was closely aligned, leading to two key insights: first, investing considerable effort solely for marginal improvements in detection precision yields low cost-effectiveness; second, in the agricultural sector, the emphasis is placed on the sparsity of crop target distribution. Therefore, within an acceptable range of precision, obtaining an overall distribution of objects in the DOM assumes greater significance.

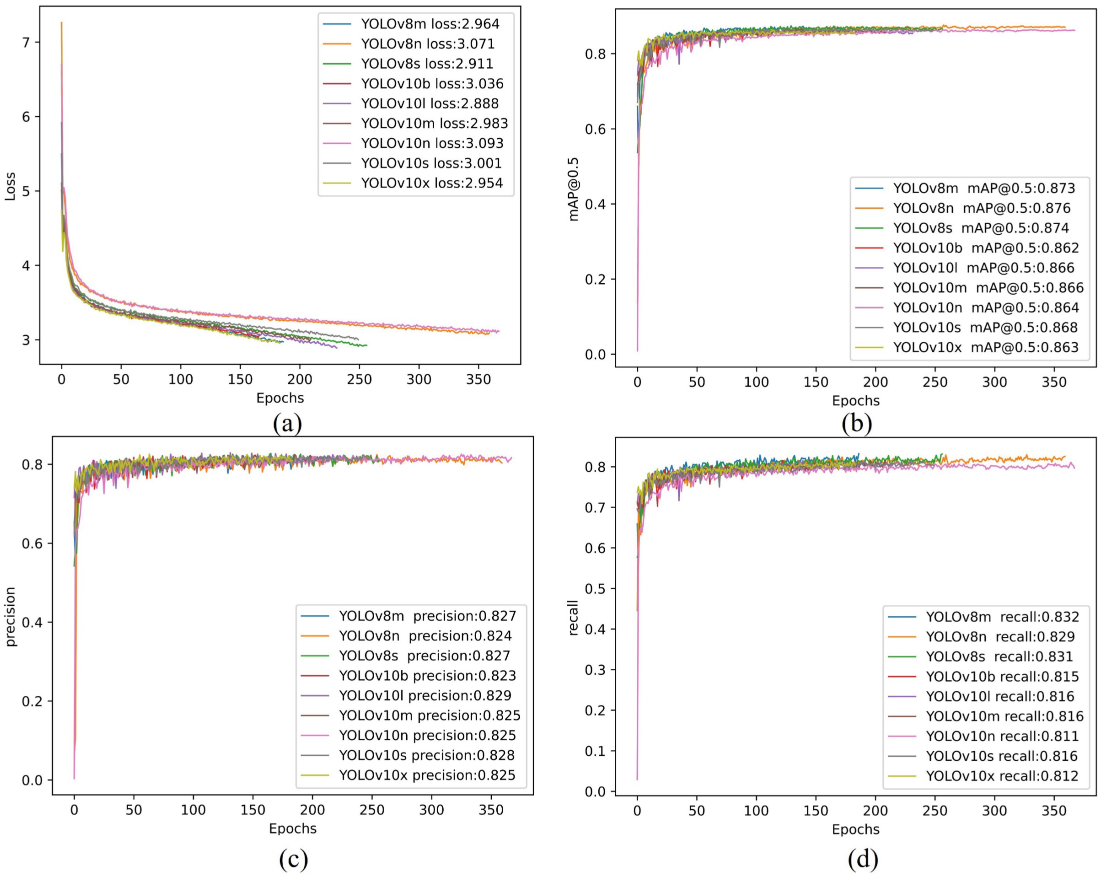


Figure A1. Metric curves with YOLOv8 and YOLOv10: (**a**) loss value curves; (**b**) mAP@0.5 value curves; (**c**) precision value curves; (**d**) recall value curves.

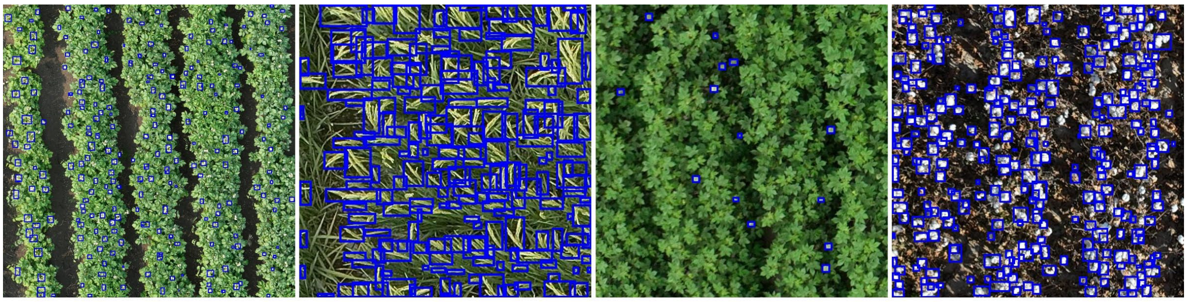
In the agricultural domain, balancing the computational resources of the model with detection precision is crucial. Lower computational demands facilitate rapid deployment to smart agricultural platforms and devices, thus enhancing the practicality of the models and methods. In Appendix A Table A1, a comprehensive comparison of the computational resource consumption parameters for the YOLOv8 and YOLOv10 series alongside the training results from the CCOD-Dataset is presented, with optimal metrics highlighted in bold font. It was observed that YOLOv10n exhibited the fewest parameters (Param.) and FLOPs, with a latency of only 1.840 ms, indicating minimal resource consumption during computation. In contrast, YOLOv10l, which demonstrated the lowest loss value and the highest precision value, achieved reductions of 90.5%, 94.4%, and 74.7% in Param, FLOPs, and latency, respectively. However, the differences in loss and precision values were limited to 6.6% and 0.4%. Consequently, YOLOv10n was selected as the target detection model for this study to establish an optimal balance between computational resources and detection precision.

Table A1. Comparison of metrics between YOLOv8 and YOLOv10 series.

Model	Param. (M)	FLOPs (G)	Latency (ms)	Loss	mAP@0.5	Precision	Recall
YOLOv8n	3.200	8.700	6.160	3.071	**0.876**	0.824	0.829
YOLOv10n	**2.300**	**6.700**	**1.840**	3.093	0.864	0.825	0.811
YOLOv8s	11.200	26.400	7.070	2.911	0.874	0.827	0.831
YOLOv10s	7.200	21.600	2.490	3.001	0.868	0.828	0.816
YOLOv8m	25.900	78.900	9.500	2.964	0.873	0.827	**0.832**
YOLOv10m	15.400	59.100	4.740	2.983	0.866	0.825	0.816
YOLOv10b	19.100	92.000	5.470	3.036	0.862	0.823	0.815

YOLOv10l	24.400	120.300	7.280	**2.888**	0.866	**0.829**	0.816
YOLOv10x	29.500	160.400	10.700	2.954	0.863	0.825	0.812

Appendix A Figure A2 displays the detection results of YOLOv10n on crop canopy objects within the CCOD-Dataset test set. The results indicate that YOLOv10n performs effectively in detecting the four types of crop canopy objects.



(a) (b) (c) (d)

Figure A2. Results of YOLOv10n on crop canopy objects: (**a**) potato flowers; (**b**) rice panicles; (**c**) cotton flowers; (**d**) cotton bolls.

Appendix B

This Section details the experimental outcomes and performance metrics of several classical models evaluated on the CCOD-Dataset. All models were trained with a batch size of 16 and a learning rate of 0.001, while default configurations were maintained for the remaining hyperparameters. Training dynamics are illustrated in Appendix Figures A3 and A4, and key parameters are summarized in Table A2.

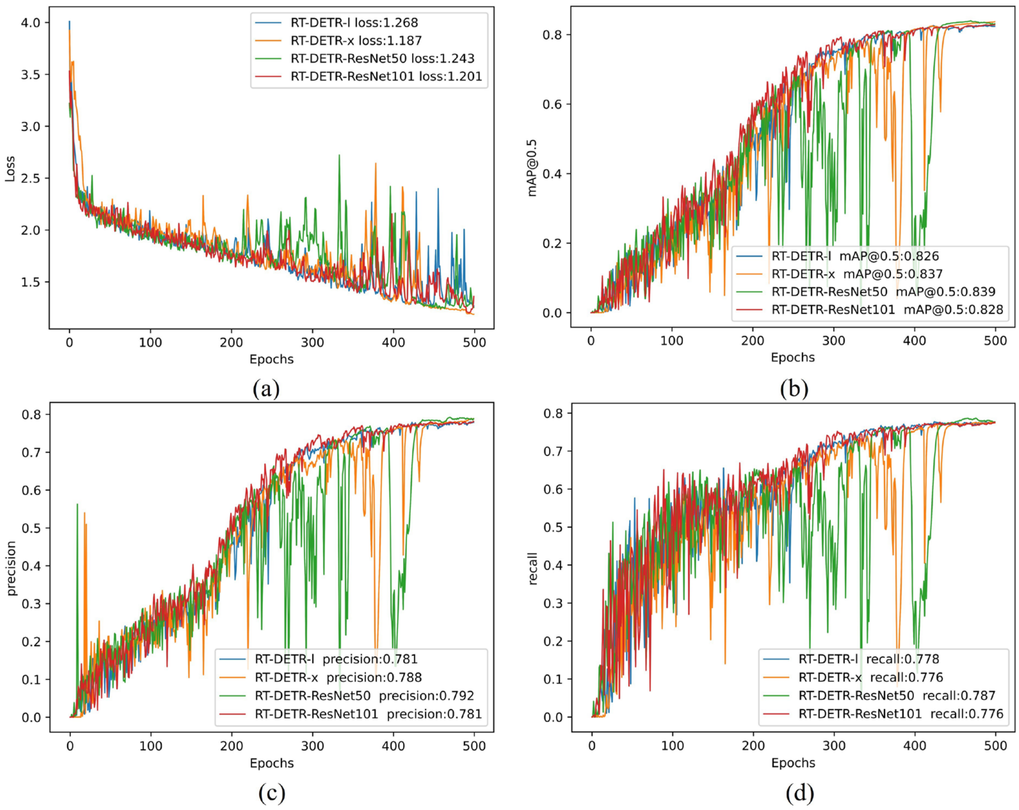


Figure A3. Metric curves with RT-DETR series: (**a**) loss value curves; (**b**) mAP@0.5 value curves; (**c**) precision value curves; (**d**) recall value curves.

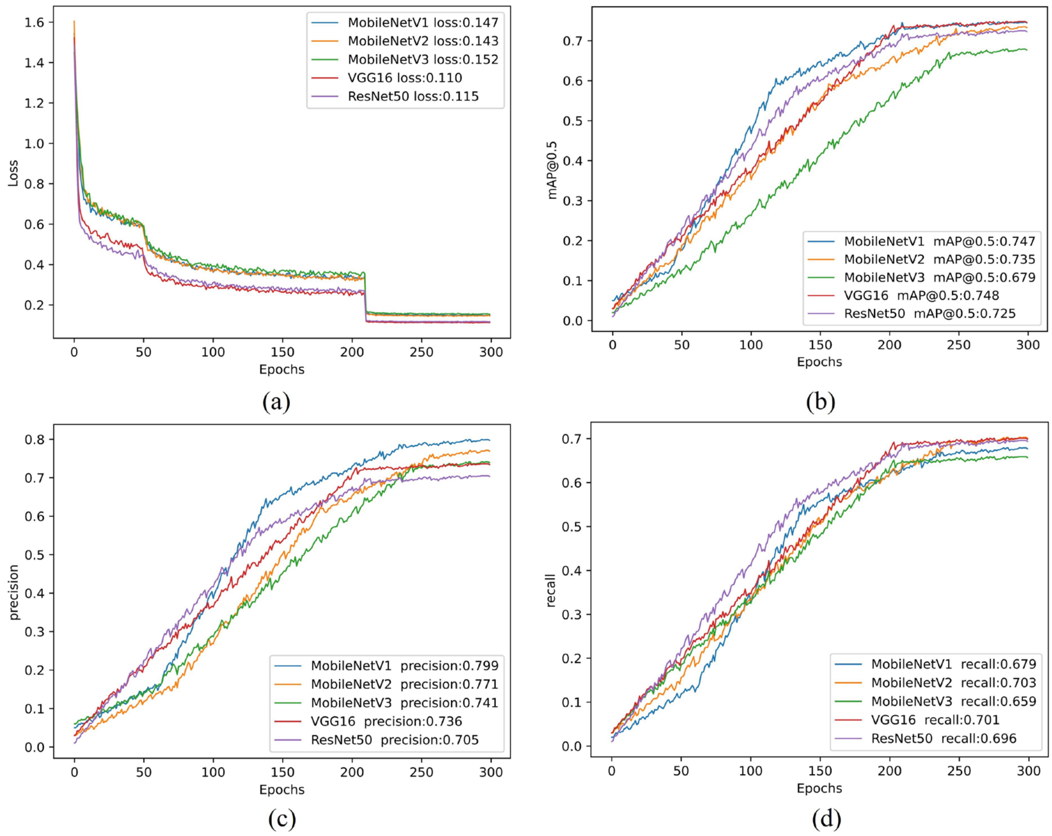


Figure A4. Metric curves with MobileNet series, VGG16, and ResNet50: (**a**) loss value curves; (**b**) mAP@0.5 value curves; (**c**) precision value curves; (**d**) recall value curves.

Table A2. Comparison of metrics between classic models.

Model	Param.(M)	FLOPs(G)	Latency(ms)	Loss	mAP@0.5	Precision	Recall
RT-DETR-l	31.987	103.400	87.825	1.268	0.826	0.781	0.778
RT-DETR-x	65.471	222.500	93.025	1.187	0.837	0.788	0.776
RT-DETR-ResNet50	41.939	125.600	87.351	1.243	0.839	0.792	0.787
RT-DETR-ResNet101	60.905	186.200	93.165	1.201	0.828	0.781	0.776
MobileNetV1	12.272	23.870	19.751	0.147	0.747	0.799	0.679
MobileNetV2	10.381	18.270	25.402	0.143	0.735	0.771	0.703
MobileNetV3	11.309	16.882	32.216	0.152	0.679	0.741	0.659
VGG16	23.518	264.620	27.656	0.110	0.748	0.736	0.701
ResNet50	33.261	82.985	29.812	0.115	0.725	0.705	0.696

As depicted in Figure A3, the RT-DETR series exhibits significant oscillations in its metric curves during training in the detection of small crop organs. This instability likely results from the transformer-based architecture, which typically requires large-scale data and meticulous optimization to stabilize attention mechanisms. Additionally, the high number of parameters (approximately 32 M to 65 M) and considerable computational complexity (roughly 103 G to 223 G FLOPs) further exacerbate optimization challenges for small object tasks, leading to convergence difficulties. In contrast, Figure A4 reveals that models based on the MobileNet series, VGG16, and ResNet50 produce smoother training curves, suggesting enhanced stability in small-object detection. However, their overall performance—characterized by a mean average precision (mAP) ranging from 0.679 to 0.748 and precision/recall metrics below 0.8—remains inferior to that of the RT-DETR series. This trade-off is likely due to the simpler architectures; for example, while lightweight designs (e.g., MobileNetV2 with approximately 10.4 M parameters) facilitate optimization stability, they also constrain the feature representation capacity needed to accurately detect small, texture-rich agricultural targets.

Table A2 further quantifies the computational inefficiency of these classical models for edge deployment. For instance, VGG16 demands an excessive number of FLOPs (approximately 265 G), despite its moderate parameter count (around 23.5 M). Similarly, the RT-DETR variant employing a ResNet101 backbone suffers from high inference latency (93.2 ms), rendering it impractical for real-time applications. Moreover, all classical models underperform compared to modern architectures (e.g., the YOLOv8 and YOLOv10 series), which achieve precision and recall metrics above 0.8—highlighting limitations in both efficiency and small-object detection accuracy. In summary, classical models face significant challenges in balancing computational efficiency with detection performance for small agricultural objects. While transformer-based architectures like RT-DETR offer marginally higher accuracy, their training instability and resource demands hinder practical deployment. Conversely, lightweight CNNs (MobileNet, VGG, and ResNet) provide greater stability but at the expense of discriminative power.

References

1. Liu, J.; Xiang, J.; Jin, Y.; Liu, R.; Yan, J.; Wang, L. Boost Precision Agriculture with Unmanned Aerial Vehicle Remote Sensing and Edge Intelligence: A Survey. *Remote Sens.* **2021**, *13*, 4387. https://doi.org/10.3390/rs13214387.
2. Zahra, S.; Ruiz, H.; Jung, J.; Adams, T. UAV-Based Phenotyping: A Non-Destructive Approach to Studying Wheat Growth Patterns for Crop Improvement and Breeding Programs. *Remote Sens.* **2024**, *16*, 3710. https://doi.org/10.3390/rs16193710.
3. Wang, D.; Zhao, M.; Li, Z.; Xu, S.; Wu, X.; Ma, X.; Liu, X. A survey of unmanned aerial vehicles and deep learning in precision agriculture. *Eur. J. Agron.* **2025**, *164*, 127477. https://doi.org/10.1016/j.eja.2024.127477.
4. Vong, A.; Matos-Carvalho, J.P.; Toffanin, P.; Pedro, D.; Azevedo, F.; Moutinho, F.; Garcia, N.C.; Mora, A. How to Build a 2D and 3D Aerial Multispectral Map?—All Steps Deeply Explained. *Remote Sens.* **2021**, *13*, 3227. https://doi.org/10.3390/rs13163227.
5. Aszkowski, P.; Kraft, M.; Drapikowski, P.; Pieczyński, D. Estimation of corn crop damage caused by wildlife in UAV images. *Precis. Agric.* **2024**, *25*, 2505–2530. https://doi.org/10.1007/s11119-024-10180-7.
6. Jyothi, V.K.; Aradhya, V.N.M.; Sharath Kumar, Y.H.; Guru, D.S. Retrieval of flower videos based on a query with multiple species of flowers. *Artif. Intell. Agric.* **2021**, *5*, 262–277. https://doi.org/10.1016/j.aiia.2021.11.001.
7. Bay, H.; Ess, A.; Tuytelaars, T.; Van Gool, L. Speeded-Up Robust Features (SURF). *Comput. Vis. Image Underst.* **2008**, *110*, 346–359. https://doi.org/10.1016/j.cviu.2007.09.014.
8. Rublee, E.; Rabaud, V.; Konolige, K.; Bradski, G. ORB: An efficient alternative to SIFT or SURF. In Proceedings of the 2011 International Conference on Computer Vision, Barcelona, Spain, 6–13 November 2011; pp. 2564–2571. https://doi.org/10.1109/ICCV.2011.6126544.
9. Liu, W.; Sun, H.; Xia, Y.; Kang, J. Real-Time Cucumber Target Recognition in Greenhouse Environments Using Color Segmentation and Shape Matching. *Appl. Sci.* **2024**, *14*, 1884. https://doi.org/10.3390/app14051884.
10. Liang, W.-z.; Kirk, K.R.; Greene, J.K. Estimation of soybean leaf area, edge, and defoliation using color image analysis. *Comput. Electron. Agric.* **2018**, *150*, 41–51. https://doi.org/10.1016/j.compag.2018.03.021.
11. Linaza, M.T.; Posada, J.; Bund, J.; Eisert, P.; Quartulli, M.; Döllner, J.; Pagani, A.; Olaizola, I.G.; Barriguinha, A.; Moysiadis, T.; et al. Data-Driven Artificial Intelligence Applications for Sustainable Precision Agriculture. *Agronomy* **2021**, *11*, 1227.
12. Istiak, M.A.; Syeed, M.M.M.; Hossain, M.S.; Uddin, M.F.; Hasan, M.; Khan, R.H.; Azad, N.S. Adoption of Unmanned Aerial Vehicle (UAV) imagery in agricultural management: A systematic literature review. *Ecol. Inform.* **2023**, *78*, 102305. https://doi.org/10.1016/j.ecoinf.2023.102305.
13. Wachs, J.P.; Stern, H.I.; Burks, T.; Alchanatis, V. Low and high-level visual feature-based apple detection from multi-modal images. *Precis. Agric.* **2010**, *11*, 717–735. https://doi.org/10.1007/s11119-010-9198-x.
14. Li, L.; Liu, L.; Peng, Y.; Su, Y.; Hu, Y.; Zou, R. Integration of multimodal data for large-scale rapid agricultural land evaluation using machine learning and deep learning approaches. *Geoderma* **2023**, *439*, 116696. https://doi.org/10.1016/j.geoderma.2023.116696.
15. Zhu, H.; Lin, C.; Liu, G.; Wang, D.; Qin, S.; Li, A.; Xu, J.-L.; He, Y. Intelligent agriculture: Deep learning in UAV-based remote sensing imagery for crop diseases and pests detection. *Front. Plant Sci.* **2024**, *15*, 1435016. https://doi.org/10.3389/fpls.2024.1435016.

16. Rasmussen, C.B.; Kirk, K.; Moeslund, T.B. The Challenge of Data Annotation in Deep Learning—A Case Study on Whole Plant Corn Silage. *Sensors* **2022**, *22*, 1596. https://doi.org/10.3390/s22041596.
17. Nethala, P.; Um, D.; Vemula, N.; Montero, O.F.; Lee, K.; Bhandari, M. Techniques for Canopy to Organ Level Plant Feature Extraction via Remote and Proximal Sensing: A Survey and Experiments. *Remote Sens.* **2024**, *16*, 4370. https://doi.org/10.3390/rs16234370.
18. Bao, W.; Liu, W.; Yang, X.; Hu, G.; Zhang, D.; Zhou, X. Adaptively spatial feature fusion network: An improved UAV detection method for wheat scab. *Precis. Agric.* **2023**, *24*, 1154–1180. https://doi.org/10.1007/s11119-023-10004-0.
19. Lin, Y.-C.; Zhou, T.; Wang, T.; Crawford, M.; Habib, A. New Orthophoto Generation Strategies from UAV and Ground Remote Sensing Platforms for High-Throughput Phenotyping. *Remote Sens.* **2021**, *13*, 860. https://doi.org/10.3390/rs13050860.
20. Petti, D.; Li, C. Weakly-supervised learning to automatically count cotton flowers from aerial imagery. *Comput. Electron. Agric.* **2022**, *194*, 106734. https://doi.org/10.1016/j.compag.2022.106734.
21. Wan, L.; Zhang, G. Super-resolution reconstruction of unmanned aerial vehicle image based on deep learning. *J. Phys. Conf. Ser.* **2021**, *1948*, 012028. https://doi.org/10.1088/1742-6596/1948/1/012028.
22. Shoab, M.; Singh, V.K.; Ravibabu, M.V. High-Precise True Digital Orthoimage Generation and Accuracy Assessment based on UAV Images. *J. Indian Soc. Remote Sens.* **2022**, *50*, 613–622. https://doi.org/10.1007/s12524-021-01364-z.
23. Zhang, J.; Xu, S.; Zhao, Y.; Sun, J.; Xu, S.; Zhang, X. Aerial orthoimage generation for UAV remote sensing: Review. *Inf. Fusion* **2023**, *89*, 91–120. https://doi.org/10.1016/j.inffus.2022.08.007.
24. Wang, H.; Li, Y.; Minh Dang, L.; Moon, H. An efficient attention module for instance segmentation network in pest monitoring. *Comput. Electron. Agric.* **2022**, *195*, 106853. https://doi.org/10.1016/j.compag.2022.106853.
25. Hu, G.; Ye, R.; Wan, M.; Bao, W.; Zhang, Y.; Zeng, W. Detection of Tea Leaf Blight in Low-Resolution UAV Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 1–18. https://doi.org/10.1109/TGRS.2023.3339765.
26. Zhou, J.; Xi, W.; Li, D.; Huang, H. Image enhancement and image matching of UAV based on histogram equalization. In Proceedings of the 2021 28th International Conference on Geoinformatics, Nanchang, China, 3–5 November 2021; pp. 1–5. https://doi.org/10.1109/IEEECONF54055.2021.9687650.
27. Panduangnat, L.; Posom, J.; Saikaew, K.; Phuphaphud, A.; Wongpichet, S.; Chinapas, A.; Sukpancharoen, S.; Saengprachatanarug, K. Time-efficient low-resolution RGB aerial imaging for precision mapping of weed types in site-specific herbicide application. *Crop Prot.* **2024**, *184*, 106805. https://doi.org/10.1016/j.cropro.2024.106805.
28. Magoulianitis, V.; Ataloglou, D.; Dimou, A.; Zarpalas, D.; Daras, P. Does Deep Super-Resolution Enhance UAV Detection? In Proceedings of the 2019 16th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), Taipei, Taiwan, 18–21 September 2019; pp. 1–6. https://doi.org/10.1109/AVSS.2019.8909865.
29. Tang, G.; Ni, J.; Zhao, Y.; Gu, Y.; Cao, W. A Survey of Object Detection for UAVs Based on Deep Learning. *Remote Sens.* **2024**, *16*, 149. https://doi.org/10.3390/rs16010149.
30. Su, J.; Qin, Y.; Jia, Z.; Liang, B. MPE-YOLO: Enhanced small target detection in aerial imaging. *Sci. Rep.* **2024**, *14*, 17799. https://doi.org/10.1038/s41598-024-68934-2.
31. Wu, H.; Wang, Y.; Zhao, P.; Qian, M. Small-target weed-detection model based on YOLO-V4 with improved backbone and neck structures. *Precis. Agric.* **2023**, *24*, 2149–2170. https://doi.org/10.1007/s11119-023-10035-7.
32. Li, J.; Wang, E.; Qiao, J.; Li, Y.; Li, L.; Yao, J.; Liao, G. Automatic rape flower cluster counting method based on low-cost labelling and UAV-RGB images. *Plant Methods* **2023**, *19*, 40. https://doi.org/10.1186/s13007-023-01017-x.
33. Tan, C.; Li, C.; He, D.; Song, H. Anchor-free deep convolutional neural network for tracking and counting cotton seedlings and flowers. *Comput. Electron. Agric.* **2023**, *215*, 108359. https://doi.org/10.1016/j.compag.2023.108359.
34. Abekasis, D.; Sadka, A.; Rokach, L.; Shiff, S.; Morozov, M.; Kamara, I.; Paz-Kagan, T. Explainable machine learning for revealing causes of citrus fruit cracking on a regional scale. *Precis. Agric.* **2024**, *25*, 589–613. https://doi.org/10.1007/s11119-023-10084-y.
35. Yu, X.; Yin, D.; Xu, H.; Pinto Espinosa, F.; Schmidhalter, U.; Nie, C.; Bai, Y.; Sankaran, S.; Ming, B.; Cui, N.; et al. Maize tassel number and tasseling stage monitoring based on near-ground and UAV RGB images by improved YoloV8. *Precis. Agric.* **2024**, *25*, 1800–1838. https://doi.org/10.1007/s11119-024-10135-y.
36. Vélez, S.; Ariza-Sentís, M.; Panić, M.; Ivošević, B.; Stefanović, D.; Kaivosoja, J.; Valente, J. Speeding up UAV-based crop variability assessment through a data fusion approach using spatial interpolation for site-specific management. *Smart Agric. Technol.* **2024**, *8*, 100488. https://doi.org/10.1016/j.atech.2024.100488.
37. Yang, C.; Baireddy, S.; Cai, E.; Crawford, M.; Delp, E.J. Field-Based Plot Extraction Using UAV RGB Images. In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW), Montreal, BC, Canada, 11–17 October 2021; pp. 1390–1398. https://doi.org/10.1109/ICCVW54120.2021.00160.

38. Chen, Y.; Jiang, J. An Oblique-Robust Absolute Visual Localization Method for GPS-Denied UAV With Satellite Imagery. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 1–13. https://doi.org/10.1109/TGRS.2023.3342142.
39. Luo, X.; Tian, Y.; Wan, X.; Xu, J.; Ke, T. Deep learning based cross-view image matching for UAV geo-localization. In Proceedings of the 2022 International Conference on Service Robotics (ICoSR), Chengdu, China, 10–12 June 2022; pp. 102–106. https://doi.org/10.1109/ICoSR57188.2022.00028.
40. Gómez-Candón, D.; De Castro, A.I.; López-Granados, F. Assessing the accuracy of mosaics from unmanned aerial vehicle (UAV) imagery for precision agriculture purposes in wheat. *Precis. Agric.* **2014**, *15*, 44–56. https://doi.org/10.1007/s11119-013-9335-4.
41. Feng, A.; Vong, C.N.; Zhou, J.; Conway, L.S.; Zhou, J.; Vories, E.D.; Sudduth, K.A.; Kitchen, N.R. Developing an image processing pipeline to improve the position accuracy of single UAV images. *Comput. Electron. Agric.* **2023**, *206*, 107650. https://doi.org/10.1016/j.compag.2023.107650.
42. Cagatay Akyon, F.; Onur Altinuc, S.; Temizel, A. Slicing Aided Hyper Inference and Fine-tuning for Small Object Detection. *arXiv* **2022**, arXiv:2202.06934. https://doi.org/10.48550/arXiv.2202.06934.
43. Reedha, R.; Dericquebourg, E.; Canals, R.; Hafiane, A. Transformer Neural Network for Weed and Crop Classification of High Resolution UAV Images. *Remote Sens.* **2022**, *14*, 592. https://doi.org/10.3390/rs14030592.
44. Hosoya, Y.; Suganuma, M.; Okatani, T. Rethinking Annotation for Object Detection: Is Annotating Small-size Instances Worth Its Cost? *arXiv* **2024**, arXiv:2412.05611. https://doi.org/10.48550/arXiv.2412.05611.
45. Wang, H.; Duan, Y.; Shi, Y.; Kato, Y.; Ninomiya, S.; Guo, W. EasyIDP: A Python Package for Intermediate Data Processing in UAV-Based Plant Phenotyping. *Remote Sens.* **2021**, *13*, 2622. https://doi.org/10.3390/rs13132622.
46. Wang, H.; Guo, W. EasyIDP V2.0: An Intermediate Data Processing Package for Photogrammetry-Based Plant Phenotyping. In *Harnessing Data Science for Sustainable Agriculture and Natural Resource Management*; Springer: Singapore, 2024; pp. 149–172. https://doi.org/10.1007/978-981-97-7762-4_7.
47. Teng, Z.; Chen, J.; Wang, J.; Wu, S.; Chen, R.; Lin, Y.; Shen, L.; Jackson, R.; Zhou, J.; Yang, C. Panicle-Cloud: An Open and AI-Powered Cloud Computing Platform for Quantifying Rice Panicles from Drone-Collected Imagery to Enable the Classification of Yield Production in Rice. *Plant Phenomics* **2023**, *5*, 0105. https://doi.org/10.34133/plantphenomics.0105.
48. Peng, J.; Sun, M.; Lim, E.G.; Wang, Q.; Xiao, J. Prototype Guided Pseudo Labeling and Perturbation-based Active Learning for domain adaptive semantic segmentation. *Pattern Recognit.* **2024**, *148*, 110203. https://doi.org/10.1016/j.patcog.2023.110203.
49. Lawal, O.M. Real-time cucurbit fruit detection in greenhouse using improved YOLO series algorithm. *Precis. Agric.* **2024**, *25*, 347–359. https://doi.org/10.1007/s11119-023-10074-0.
50. Russell, B.C.; Torralba, A.; Murphy, K.P.; Freeman, W.T. LabelMe: A Database and Web-Based Tool for Image Annotation. *Int. J. Comput. Vis.* **2008**, *77*, 157–173. https://doi.org/10.1007/s11263-007-0090-8.
51. Simonyan, K.; Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv* **2014**, arXiv:1409.1556. https://doi.org/10.48550/arXiv.1409.1556.
52. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2015; pp. 770–778. https://doi.org/10.1109/CVPR.2016.90..
53. Howard, A.G.; Zhu, M.; Chen, B.; Kalenichenko, D.; Wang, W.; Weyand, T.; Andreetto, M.; Adam, H. MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications. *arXiv* **2017**, arXiv:1704.04861. https://doi.org/10.48550/arXiv.1704.04861.
54. Sandler, M.; Howard, A.; Zhu, M.; Zhmoginov, A.; Chen, L.-C. MobileNetV2: Inverted Residuals and Linear Bottlenecks. *arXiv* **2018**, arXiv:1801.04381. https://doi.org/10.48550/arXiv.1801.04381.
55. Howard, A.; Sandler, M.; Chu, G.; Chen, L.-C.; Chen, B.; Tan, M.; Wang, W.; Zhu, Y.; Pang, R.; Vasudevan, V.; et al. Searching for MobileNetV3. *arXiv* **2019**, arXiv:1905.02244. https://doi.org/10.48550/arXiv.1905.02244.
56. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You Only Look Once: Unified, Real-Time Object Detection. *arXiv* **2015**, arXiv:1506.02640. https://doi.org/10.48550/arXiv.1506.02640.
57. Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; Zagoruyko, S. End-to-End Object Detection with Transformers. *arXiv* **2020**, arXiv:2005.12872. https://doi.org/10.48550/arXiv.2005.12872.
58. Varghese, R.; S, M. YOLOv8: A Novel Object Detection Algorithm with Enhanced Performance and Robustness. In Proceedings of the 2024 International Conference on Advances in Data Engineering and Intelligent Computing Systems (ADICS), Chennai, India, 18–19 April 2024; pp. 1–6. https://doi.org/10.1109/ADICS58448.2024.10533619.

59. Wang, A.; Chen, H.; Liu, L.; Chen, K.; Lin, Z.; Han, J.; Ding, G. YOLOv10: Real-Time End-to-End Object Detection. *arXiv* **2024**, arXiv:2405.14458. https://doi.org/10.48550/arXiv.2405.14458.
60. Peng, J.; Rezaei, E.E.; Zhu, W.; Wang, D.; Li, H.; Yang, B.; Sun, Z. Plant Density Estimation Using UAV Imagery and Deep Learning. *Remote Sens.* **2022**, *14*, 5923. https://doi.org/10.3390/rs14235923.
61. Bai, X.; Liu, P.; Cao, Z.; Lu, H.; Xiong, H.; Yang, A.; Cai, Z.; Wang, J.; Yao, J. Rice Plant Counting, Locating, and Sizing Method Based on High-Throughput UAV RGB Images. *Plant Phenomics* **2023**, *5*, 0020. https://doi.org/10.34133/plantphenomics.0020.
62. Lyu, Z.; Wang, Y.; Huang, C.; Zhang, G.; Ding, K.; Tang, N.; Zhao, Z. Dynamic monitoring and counting for lotus flowers and seedpods with UAV based on improved YOLOv7-tiny. *Comput. Electron. Agric.* **2024**, *225*, 109344. https://doi.org/10.1016/j.compag.2024.109344.
63. Wu, Z.; Wang, X.; Jia, M.; Liu, M.; Sun, C.; Wu, C.; Wang, J. Dense object detection methods in RAW UAV imagery based on YOLOv8. *Sci. Rep.* **2024**, *14*, 18019. https://doi.org/10.1038/s41598-024-69106-y.