sustainability

MDPI

Article

APD-YOLOv7: Enhancing Sustainable Farming through Precise Identification of Agricultural Pests and Diseases Using a Novel Diagonal Difference Ratio IOU Loss

Jianwen Li [1,2,†], Shutian Liu [3,4,†], Dong Chen [1,2], Shengbang Zhou [1,2,*] and Chuanqi Li [1,2,*]

[1] Guangxi Key Laboratory of Functional Information Materials and Intelligent Information Processing, Nanning Normal University, Nanning 530001, China; ljwaassslw@163.com (J.L.)
[2] Smart Logistics Industry School of Nanning Normal University, Demonstrative Modern Industrial School of Guangxi University, Nanning 530001, China
[3] Guangxi Geographical Indication Crops Research Center of Big Data Mining and Experimental Engineering Technology, Nanning Normal University, Nanning 530001, China
[4] Guangxi Key Laboratory of Earth Surface Processes and Intelligent Simulation, Nanning Normal University, Nanning 530001, China
* Correspondence: zhoushengbang@foxmail.com (S.Z.); lchq@nnnu.edu.cn (C.L.)
† These authors contributed equally to this work.

Abstract: The diversity and complexity of the agricultural environment pose significant challenges for the collection of pest and disease data. Additionally, pest and disease datasets often suffer from uneven distribution in quantity and inconsistent annotation standards. Enhancing the accuracy of pest and disease recognition remains a challenge for existing models. We constructed a representative agricultural pest and disease dataset, FIP6Set, through a combination of field photography and web scraping. This dataset encapsulates key issues encountered in existing agricultural pest and disease datasets. Referencing existing bounding box regression (BBR) loss functions, we reconsidered their geometric features and proposed a novel bounding box similarity comparison metric, DDRIoU, suited to the characteristics of agricultural pest and disease datasets. By integrating the focal loss concept with the DDRIoU loss, we derived a new loss function, namely Focal-DDRIoU loss. Furthermore, we modified the network structure of YOLOV7 by embedding the MobileViTv3 module. Consequently, we introduced a model specifically designed for agricultural pest and disease detection in precision agriculture. We conducted performance evaluations on the FIP6Set dataset using mAP75 as the evaluation metric. Experimental results demonstrate that the Focal-DDRIoU loss achieves improvements of 1.12%, 1.24%, 1.04%, and 1.50% compared to the GIoU, DIoU, CIoU, and EIoU losses, respectively. When employing the GIoU, DIoU, CIoU, EIoU, and Focal-DDRIoU loss functions, the adjusted network structure showed enhancements of 0.68%, 0.68%, 0.78%, 0.60%, and 0.56%, respectively, compared to the original YOLOv7. Furthermore, the proposed model outperformed the mainstream YOLOv7 and YOLOv5 models by 1.86% and 1.60%, respectively. The superior performance of the proposed model in detecting agricultural pests and diseases directly contributes to reducing pesticide misuse, preventing large-scale pest and disease outbreaks, and ultimately enhancing crop yields. These outcomes strongly support the promotion of sustainable agricultural development.

Keywords: loss function design; MobileViTv3; YOLOv7; agricultural diseases and pests; sustainable agricultural development

Citation: Li, J.; Liu, S.; Chen, D.; Zhou, S.; Li, C. APD-YOLOv7: Enhancing Sustainable Farming through Precise Identification of Agricultural Pests and Diseases Using a Novel Diagonal Difference Ratio IOU Loss. *Sustainability* **2024**, *16*, 8855. https://doi.org/10.3390/su16208855

Academic Editors: Yang Li and Jing Nie

Received: 8 September 2024
Revised: 3 October 2024
Accepted: 7 October 2024
Published: 13 October 2024

1. Introduction

Globally, agriculture plays a crucial role in contributing to the health of the population, maintaining social stability, and even contributing to national security. Therefore, there has been a consistent demand for continuous innovation in agricultural technology to enhance the efficiency of the agricultural industry and maximize food production to meet the needs

of a growing population [1]. However, due to the abundance of harmful organisms and microorganisms in the cultivation environment, crops are increasingly susceptible to pests and diseases. Infestations of pests and diseases seriously threaten agricultural production safety and the sustainable supply of food. Therefore, accurate identification of crop pests and diseases and effective early warning of their outbreaks can help prevent agricultural disasters and ensure the quality and yield of farmland [2,3].

However, traditional methods for pest and disease identification typically depend on the extensive experience and professional knowledge of agricultural experts [4], which can pose efficiency challenges when handling large-scale datasets. While expert diagnosis is crucial for ensuring accuracy, the increasing diversity and complexity of crop diseases may strain the availability of experts' time and resources. Therefore, it is particularly important to explore more efficient and accurate methods for pest and disease identification, in order to assist agricultural experts in enhancing overall identification efficiency, support agricultural practitioners in boosting crop yields, and promote sustainable agricultural development.

With the advancement of image processing technology, crop pest and disease identification methods based on machine learning have emerged [5,6]. However, these methods still struggle with extracting abstract features and improving accuracy to a level suitable for practical applications. Moreover, improving accuracy to a practical application level remains challenging. Subsequently, with the development of graphics processing units (GPUs), deep learning technology, which relies on the powerful processing capabilities of GPUs, has rapidly evolved [7] and is widely used in agricultural pest and disease identification [8]. For large-scale agricultural pest and disease datasets, deep learning technologies have demonstrated superior accuracy in identifying crop pests and diseases compared to traditional methods and other machine learning-based approaches [9]. Therefore, exploring deep learning models that are tailored to agricultural pest and disease identification can significantly enhance the accuracy and efficiency of diagnosis. This, in turn, has the potential to prevent large-scale outbreaks of crop pests and diseases, reduce excessive pesticide usage, foster healthy crop growth, increase crop yields, and ultimately steer agriculture towards sustainable development.

Although many studies provide references and demonstrate the feasibility of using deep learning neural networks to identify agricultural pests and diseases, existing deep learning models still face challenges in addressing issues such as high computational costs and ensuring high accuracy in agricultural pest and disease identification.

The bounding box regression (BBR) module is crucial for determining object locations in advanced object detectors. During training, BBR adjusts predicted bounding box coordinates to better align with the ground truth bounding box, thereby improving object localization accuracy—a pivotal aspect of object detection tasks. The design of the loss function is crucial for BBR's success, as it directly affects training efficiency and the model's overall performance. So far, BBR-based loss functions are mostly classified into two categories: ℓn-norm-based loss functions and intersection-over-union (IoU)-based loss functions.

In the field of object detection, the YOLO series [10–16] models, particularly YOLOv7 [15], have occupied a mainstream position in real-time object detection tasks due to their excellent inference speed and high detection accuracy. We selected YOLOv7 as the foundation for our research due to its demonstrated balance between high speed and accuracy across multiple public datasets, particularly its proficiency in detecting small targets and managing complex scenes. Additionally, YOLOv7 offers a rich array of configuration options and good hardware compatibility, facilitating model portability and subsequent improvements.

From YOLOv1 [10] to YOLOv7, each generation of the YOLO series has meticulously optimized its network structure and loss function in pursuit of higher detection efficiency and accuracy. Meanwhile, successors to the RCNN [17] model, such as Fast R-CNN [18], Mask R-CNN [19], and Dynamic R-CNN [20], have each improved their loss functions according to their respective functional emphases. Furthermore, the DETR [21] model is a direct set prediction object detection model based on the transformer architecture and

bipartite matching loss, with one of its core designs being the computation of loss through a bipartite matching process. Despite their application in various domains, the loss function remains a key component common to all these models. Enhancing the loss function is thus of paramount importance for improving the overall performance.

However, existing BBR loss functions sometimes have the same value under different prediction results, which reduces the convergence speed and accuracy of BBR. Additionally, the current BBR loss functions are not fully suitable for precise identification of agricultural pests and diseases. A well-designed loss function can assist the model in performing agricultural pest and disease recognition tasks with greater accuracy and efficiency.

Therefore, building upon the YOLOv7 model, we introduce Agricultural Pest Detection-YOLOv7 (APD-YOLOv7), a model specifically tailored for the detection of agricultural pests and diseases. The core aim of this study is to devise a novel model through the development of a new BBR loss function and the enhancement of YOLOv7's network architecture. This approach aims to attain precise identification of crop pests and diseases, thereby contributing to the sustainable development of agriculture.

(1) Existing BBR loss functions have limitations because they sometimes assign identical values to distinct prediction outcomes. To mitigate this issue, we propose a novel BBR loss function, termed Diagonal Difference Ratio IoU (DDRIoU), which exploits the geometric properties inherent in BBR. Our aim is to utilize DDRIoU as a metric to assess the similarity between predicted and ground truth bounding boxes during the BBR procedure.
(2) Furthermore, during the BBR process, an imbalance exists between high-quality and low-quality anchor boxes. To address this, we intend to integrate the DDRIoU loss with the focal loss concept, resulting in a regression-focused loss, termed Focal-DDRIoU loss. We expect that this combined loss function will enhance the contribution of high-quality anchor boxes to model optimization while simultaneously minimizing the impact of irrelevant anchor boxes.
(3) In order to refine the YOLOv7 network architecture for more precise identification of agricultural pests and diseases, we plan to modify the network structure of the YOLOv7 model and incorporate the MobileViTv3 module. We anticipate that this innovative network design will not only reduce the number of model parameters and computational complexity but also enhance its feature extraction and fusion capabilities, ultimately leading to improved model accuracy.

2. Materials and Methods

2.1. Bounding Box Regression Loss Function

2.1.1. Commonly Used Boundary Box Regression Loss Functions

The YOLO series of real-time detectors has been widely accepted by researchers and applied in various scenarios since its inception. For instance, YOLOv1 constructed a loss function weighted by BBR loss, classification loss, and objectness loss. Until now, this approach has remained one of the most effective loss function paradigms in object detection tasks, as the BBR loss directly determines the localization performance of the model. To further enhance the model's localization accuracy, a carefully designed BBR loss is crucial.

Initially, the ℓn-norm [22] loss function was extensively used for BBR. This function is straightforward but sensitive to various scales. Later, to more accurately calculate the differences between predicted and ground truth bounding boxes, the IoU loss function was introduced in Unitbox [22] and has been widely employed since. This is a loss function designed to optimize the position and size of bounding boxes in object detection tasks. To ensure training stability, the Bounded-IoU [23] loss function introduces an upper bound for IoU. In deep learning-based object detection and instance segmentation, IoU-based metrics are considered more consistent than the ℓn-norm.

$$IoU = \frac{A_{gt} \cap A_{pr}}{A_{gt} \cup A_{pr}} \tag{1}$$

Here, A_{gt} denotes the ground truth bounding box and A_{pr} the predicted bounding box.

The original intersection-over-union (IoU) metric computes the union area of two bounding boxes, but it fails to differentiate between nonoverlapping scenarios of the two boxes. To address this limitation, Generalized-IoU (GIoU) [24] was introduced, incorporating a penalty term within its loss function. This ensures that, in cases of nonoverlap, the predicted bounding box converges towards the target box.

$$GIoU = IoU - \frac{|C - (A_{gt} \cap A_{pr})|}{|C|}, \tag{2}$$

where C represents the minimum bounding rectangle enclosing both A_{gt} and A_{pr}, and $|C|$ denotes the area of C.

Nevertheless, GIoU becomes ineffective when the predicted bounding box is completely covered by the ground truth bounding box. To overcome this challenge, Distance-IoU (DIoU) [25] was proposed, considering the distance between the center points of the predicted and ground truth bounding boxes.

$$DIoU = IoU - \frac{\rho^2(b_{gt}, b_{pr})}{h_c^2 + w_c^2} \tag{3}$$

Here, b_{gt} and b_{pr} are the center points of the ground truth and predicted bounding boxes, respectively, and $\rho(b_{gt}, b_{pr})$ is the Euclidean distance between b_{gt} and b_{pr}.

Yet, in situations where the center points of the predicted and ground truth bounding boxes overlap, the metric reduces to the original IoU. To rectify this, Complete-IoU (CIoU) [26] was introduced, which integrates both the center point distance and the aspect ratio.

$$CIoU = IoU - \frac{\rho^2(b_{gt}, b_{pr})}{h_c^2 + w_c^2} - \alpha v, \tag{4}$$

$$v = \frac{4}{\pi^2}\left(arctan\frac{h_{gt}}{w_{gt}} - arctan\frac{h_{pr}}{w_{pr}}\right)^2, \tag{5}$$

$$\alpha = \frac{v}{1 - IoU - v} \tag{6}$$

h_{gt} and w_{gt} can be defined as the height and width of the ground truth bounding box, respectively, and h_{pr} and w_{pr} as the corresponding dimensions of the predicted bounding box.

In CIoU, however, the aspect ratio is defined based on a relative value, not an absolute one. To address this, Efficient-IoU (EIoU) [27] was proposed, building upon DIoU.

$$EIoU = IoU - \frac{\rho^2(b_{gt}, b_{pr})}{h_c^2 + w_c^2} - \frac{\rho^2(w_{gt}, w_{pr})}{w_c^2} - \frac{\rho^2(h_{gt}, h_{pr})}{h_c^2}, \tag{7}$$

Here, h_c and w_c denote the height and width of C, respectively. Similarly, h_{pr} and w_{pr} are the height and width of the predicted bounding box, and h_{gt} and w_{gt} are the corresponding dimensions of the ground truth bounding box.

Notwithstanding, EIoU could exhibit limitations in scenarios where the predicted bounding box and the ground truth bounding box possess an identical aspect ratio yet differ in terms of their widths and heights. Such circumstances have the potential to hinder convergence speed and accuracy. To mitigate this issue, further improvements or alternative metrics might be necessary.

Figure 1 is a schematic diagram of commonly used BBR loss functions.

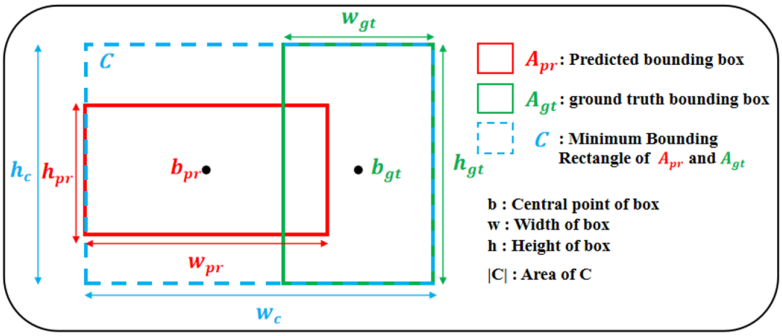

Figure 1. Commonly used BBR loss functions.

2.1.2. Analysis of BBR Loss Function

The existing loss functions fail to fully leverage the geometric attributes of BBR. In scenarios where the predicted and ground truth bounding boxes share an aspect ratio but diverge in terms of their widths and heights, these loss functions prove ineffective, thereby hindering convergence speed and accuracy.

We have observed a significant imbalance in datasets pertaining to agricultural pest and disease identification, as compared to deep learning datasets in other domains. This imbalance contributes to inconsistencies in the quality of agricultural pest and disease datasets currently available.

As depicted in Figure 2, numerous deep learning datasets for agricultural pests and diseases are compiled using varying image acquisition techniques and annotation standards. Consequently, scenarios arise where the predicted and ground truth bounding boxes exhibit matching aspect ratios but differ in their dimensions.

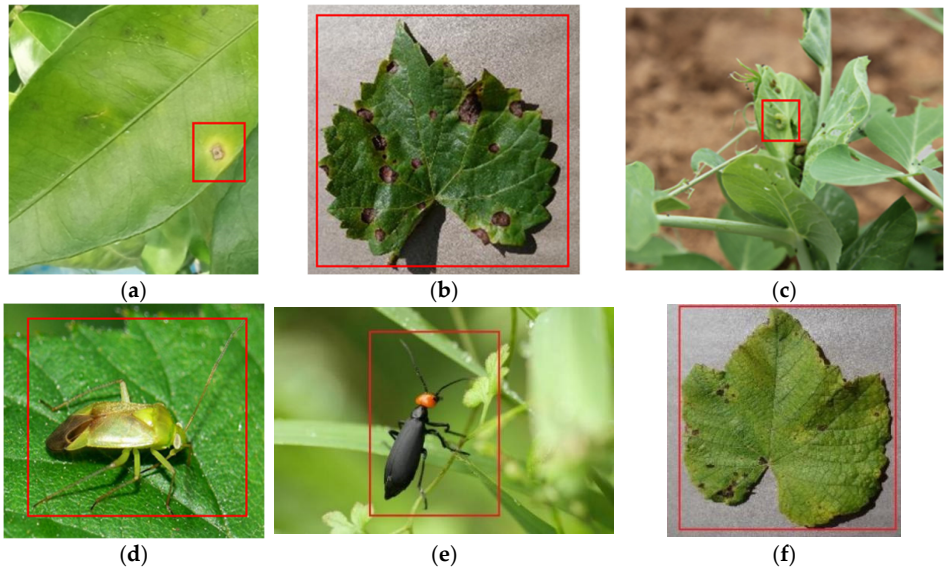

Figure 2. Current status of annotations in agricultural pest and disease datasets: (**a**) Citrus canker disease. (**b**) Black rot of grapes. (**c**) Cabbage caterpillar on snow peas. (**d**) Acridoidea insect. (**e**) Legume blister beetle. (**f**) Leaf blight of grapes.

The aforementioned scenarios can lead to existing BBR loss functions producing uniform values for disparate prediction outcomes, thereby diminishing the convergence rate and precision of BBR. The recurrence of this problem undermines the accurate identification of agricultural pests and diseases. Presently, BBR loss functions possess constraints that hinder precise identification, particularly in datasets bearing similar traits.

Hence, considering the merits of loss functions like GIoU, DIoU, CIoU, and EIoU, we strive to devise a novel loss function for BBR, termed DDRIoU loss, aimed at enhancing the efficacy and precision of BBR.

2.2. Construction of the FIP6Set Dataset

The identification of agricultural pests and diseases faces certain challenges, particularly the imbalance in the dataset caused by the varying availability of data. Some types of data are easy to obtain, resulting in a large quantity, while others are difficult to acquire, leading to a smaller amount. This imbalance affects the training and generalization capabilities of the model. To address this issue, we have constructed a crop pest and disease dataset called FIP6Set.

The FIP6Set dataset originates from three primary sources: Firstly, photos taken by us in Nanning, Guangxi Zhuang Autonomous Region, including images of citrus canker disease and cabbage caterpillars on snow peas; secondly, images from the public dataset IP102 [28], specifically, those of the Acridoidea insect and legume blister beetle; and thirdly, images sourced from the internet, featuring black rot of grapes and leaf blight of grapes.

The construction of the FIP6Set dataset involved three main components. Firstly, on several sunny days and under the guidance of agricultural experts, we captured 821 images of citrus canker disease and 800 images of cabbage caterpillars on snow peas. These images were subsequently cleaned based on clarity and realism, resulting in a selection of 440 images of citrus canker disease and 440 images of cabbage caterpillars on snow peas. These selected images were then annotated under the direction of agricultural experts and reviewed by a second expert. Secondly, we selected a portion of images from the public dataset IP102. Out of 533 images of Acridoidea insects, 470 high-quality images were chosen. Due to the limited number of legume blister beetle images available, all 435 images were included in our dataset. Lastly, we collected 1043 images of black rot of grapes and 1007 images of leaf blight of grapes from the internet. These images were also cleaned based on quality, with 460 images of black rot of grapes and 475 images of leaf blight of grapes being incorporated into our dataset. Similar to the previous components, these images were annotated under the guidance of agricultural experts and reviewed by a second expert. Thus, the new dataset, FIP6Set, was constructed.

During the image selection process, we deliberately selected a combination of complex lesions suitable for broad annotation, as seen in Figure 2d,f as well as simpler lesions amenable to precise annotation, exemplified by Figure 2a. Furthermore, we incorporated pests that bear similar colors to the leaves, as shown in Figure 2c,d and those that contrast with the leaf color, as in Figure 2e.

Based on Saleh Shahinfar's [29] research, which suggests that a minimum of 150–500 images are required for training, and considering the current state of agricultural pest and disease datasets, we maintained the image count for each category in our dataset between 435 and 475. This approach not only ensures the consistency of the dataset but also prevents the quantity of data from adversely affecting training effectiveness.

2.3. Diagonal Difference Ratio Intersection over Union

As shown in Figure 3, the red box represents the predicted bounding box, the green box represents the ground truth bounding box, and the blue box represents the minimum enclosing rectangle encompassing both the red and green boxes. When the predicted bounding box and the ground truth bounding box possess the same aspect ratio but differ in widths and heights, the traditional IoU loss function becomes less effective. Figure 3a,b depict two distinct scenarios; however, the traditional IoU loss function assigns an equivalent loss value in both instances.

To tackle the aforementioned issues and inspired by the geometric characteristics of the bounding box, we introduce DDRIoU, which is based on CIoU.

Figure 4 shows the key components of our proposed DDRIoU.

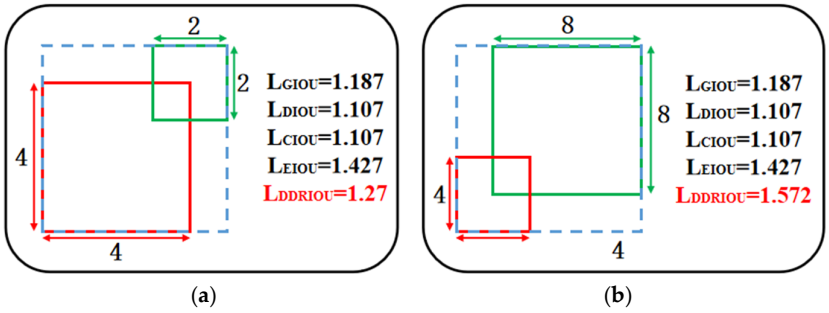

Figure 3. Calculation results across different scenarios, displaying diverse BBR losses such as GIoU, DIoU, CIoU, EIoU, and DDRIoU. Specifically, Figure (**a**) illustrates the values of various BBR losses for a scenario where the predicted bounding box is a 4 × 4 rectangle, while the ground truth bounding box is a 2 × 2 rectangle. Similarly, Figure (**b**) depicts the BBR loss values for a case where the predicted bounding box remains a 4 × 4 rectangle, but the ground truth bounding box is an 8 × 8 rectangle.

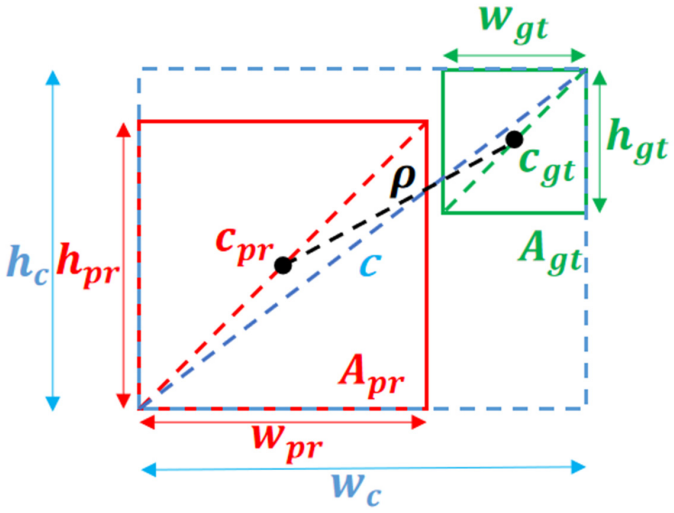

Figure 4. Our DDRIoU component. Specifically, the red box represents the predicted bounding box, the green box represents the ground truth bounding box, and the blue box represents the minimum enclosing rectangle encompassing both the red and green boxes.

Algorithm 1 outlines the computational process for DDRIoU:

Algorithm 1: Intersection over Union with Diagonal Difference Ratio.

Input: Two arbitrary convex shapes: A_{gt}, A_{pr}, width and height of input image: w, h
Output: DDRIoU
1 : Let h_{gt} and w_{gt} represent the height and width of A_{gt}, respectively.
Let h_{pr} and w_{pr} represent the height and width of A_{pr}, respectively.
2 : Let ρ represent the Euclidean distance between the center points of A_{gt} and A_{pr}.
3 : C represents the minimum bounding rectangle enclosing both A_{gt} and A_{pr}, where h_c and w_c denote the height and width of C, respectively.
4 : $IoU = \frac{A_{gt} \cap A_{pr}}{A_{gt} \cup A_{pr}}$
5 : $c_{gt}^2 = h_{gt}^2 + w_{gt}^2$
6 : $c_{pr}^2 = h_{pr}^2 + w_{pr}^2$
7 : $c^2 = h_c^2 + w_c^2$
8 : $DDRIoU = IoU - \frac{\rho^2}{c^2} - \frac{1}{\pi} * arctan\left(\left(\frac{c_{gt}^2 - c_{pr}^2}{c_{pr}^2}\right)^2\right)$
9 : return DDRIoU.

where c_{gt} represents the length of the hypotenuse of the ground truth bounding box, and c_{pr} represents the length of the hypotenuse of the predicted bounding box.

Therefore, the DDRIoU loss is defined as follows:

$$L_{DDRIoU} = 1 - IoU - \frac{\rho^2}{c^2} - \frac{1}{\pi} * \arctan((\frac{c_{gt}^2 - c_{pr}^2}{c_{pr}^2})^2), \tag{8}$$

The core component is the penalty term $\frac{1}{\pi} * \arctan((\frac{c_{gt}^2 - c_{pr}^2}{c_{pr}^2})^2)$. We utilize a function, $\frac{1}{\pi} * \arctan$, to confine the penalty term within the range of $\left[0, \frac{1}{2}\right)$. This confinement aids in stabilizing the training process and prevents gradient explosion due to excessively large loss values. The expression $\frac{c_{gt}^2 - c_{pr}^2}{c_{pr}^2}$ reflects the relative difference in proportion between the predicted bounding box and the ground truth bounding box. This is advantageous for object detection tasks involving significant variations in scale, such as detecting objects of different sizes.

Compared to traditional loss functions, utilizing the square of the hypotenuse of the bounding box as a metric enhances the stability of the loss function. This is because it provides a more consistent and robust measure of the bounding box size, regardless of its aspect ratio. Traditional metrics based on aspect ratio (or width/height difference) can cause instability during training, especially in tasks that require frequent detection of large and small targets.

In our approach, the expression $\frac{c_{gt}^2 - c_{pr}^2}{c_{pr}^2}$ plays a crucial role, where $c_{gt}^2 - c_{pr}^2$ represents the difference or deviation and c_{pr}^2 serves as a normalizing factor. Meanwhile, we specifically use c_{pr}^2 as the denominator in this expression. This approach not only avoids potential instability issues during training, especially when the ground truth bounding box size is too small, but also makes the loss function more sensitive to changes in the predicted bounding box size. This sensitivity encourages the model to better adjust the size of the predicted bounding box to closely approximate the ground truth bounding box.

In this way, we retain the characteristics of CIoU loss. DDRIoU loss, however, goes beyond CIoU by considering both the predicted and actual bounding boxes. This approach effectively addresses the issue of identical loss that arises when the predicted and actual bounding boxes share the same aspect ratio but differ in widths and heights.

2.4. Focal-DDRIoU

In BBR, there exists an issue of imbalance in training examples, where due to the sparsity of target objects in the image, the number of high-quality examples with small regression errors is far fewer than low-quality examples. Outliers can produce excessively large gradients, which are detrimental to the training process. Hence, it is crucial to allow high-quality examples to provide a greater contribution to training. Therefore, we propose the Focal-DDRIoU loss to enhance the performance of the DDRIoU loss.

Based on the research by Yi-Fan Zhang et al. [27], we similarly adopt the strategy of weighting the DDRIoU loss using the IoU value to obtain the Focal-DDRIoU loss, as shown in Formula (9):

$$L_{Focal-DDRIoU} = IoU^{\gamma} L_{DDRIoU}, \tag{9}$$

where $IoU = \frac{A_{gt} \cap A_{pr}}{A_{gt} \cup A_{pr}}$, and γ is a parameter that controls the degree of outlier suppression.

2.5. Overview of the YOLOv7 Network Structure

There are seven different variants of the YOLOv7 network structure, including the standard YOLOv7 and YOLOv7-tiny, among others. In this paper, we focus on the standard YOLOv7 structure for our research.

Figure 5 is a schematic diagram of the YOLOv7 network structure.

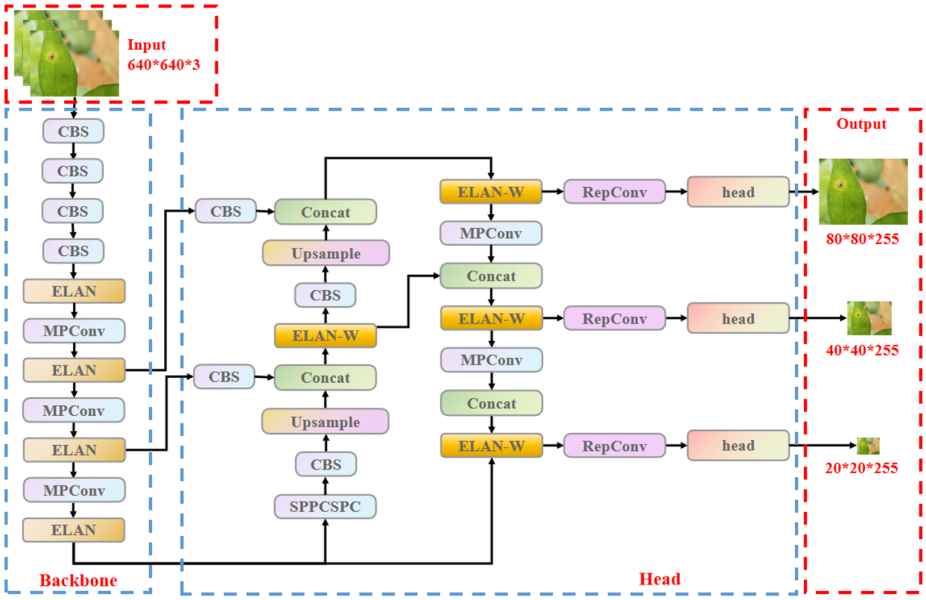

Figure 5. YOLOv7 network structure.

The overall framework of YOLOv7 is largely inherited from YOLOv5 [15], encompassing the overall network architecture and configuration file settings, as well as the training, inference, and validation processes. Furthermore, it incorporates certain aspects from YOLOR [30], such as the design of distinct networks, hyperparameter configurations, and the integration of implicit knowledge learning. Additionally, the SimOTA approach from YOLOX [31] is adopted for positive sample matching.

Apart from the elements adopted from other YOLO iterations, YOLOv7 also integrates several notable advancements, including efficient aggregated networks (e.g., ELAN and E-ELAN), re-parameterized convolution, auxiliary head detection, and model scaling.

2.6. APD-YOLOv7 Network Structure

We have made the following improvements to the YOLOv7 network structure:

(1) Introduction of C3 and MobileViTv3 Modules to Enhance Feature Extraction and Scalability.

The C3 module, a pivotal component in YOLOv5, employs a hierarchical structure for feature extraction. Due to its design, the C3 module is capable of more effectively capturing contextual information, thereby achieving superior performance in handling complex backgrounds.

MobileViT, originally known as MobileViTv1, represents a lightweight model tailored for mobile vision tasks, seamlessly integrating convolutional neural networks (CNNs) and vision transformers (ViTs).

The MobileViTv3 module, proposed by WADEKAR ShaktiN et al. [32], addresses scalability issues and simplifies learning tasks by enhancing the fusion module. As shown in Figure 6, this module seamlessly integrates features from both local and global representation blocks within the fusion block, utilizing 1×1 convolution. This approach not only simplifies the learning task but also prevents significant increases in parameters and FLOPs during scaling. Within the local representation block, 3×3 depthwise separable convolution is employed to further reduce parameters. Before the module's final output, input features are combined with the output of the fusion block, leveraging residual connections to enhance accuracy.

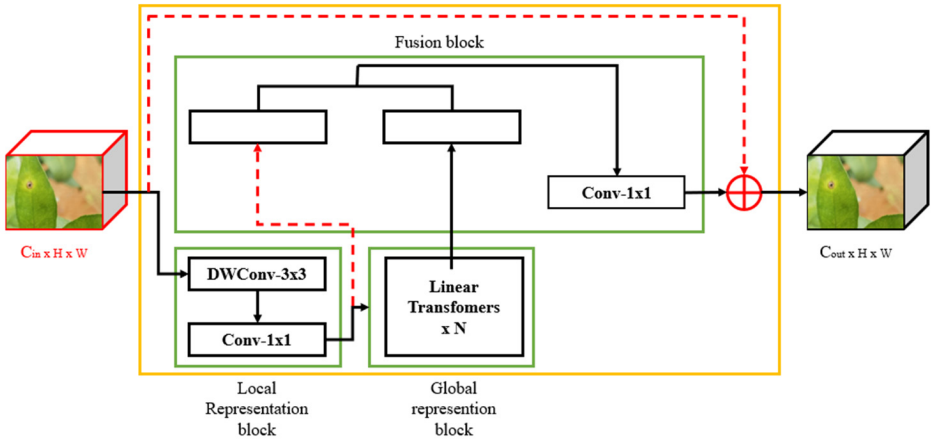

Figure 6. Structure of the MobileViTv3 module.

The incorporation of the MobileViTv3 module significantly contributes to the reduction in model parameters and the enhancement of model accuracy.

(2) Adjustments to the YOLOv7 Network Structure.

We have refined the YOLOv7 network structure with the aim of further minimizing computational complexity, parameter count, and floating-point operations, all while boosting the model's recognition accuracy and generalization capability.

As depicted in Figure 7, we initially removed two ELAN modules from the backbone section and introduced C3 modules to compensate for the network's feature extraction capabilities. Subsequently, to further elevate model performance, we embedded MobileViTv3 modules between the remaining two ELAN modules. In the head section, we substituted all ELAN-W modules with C3 modules, culminating in the new APD-YOLOv7 network structure.

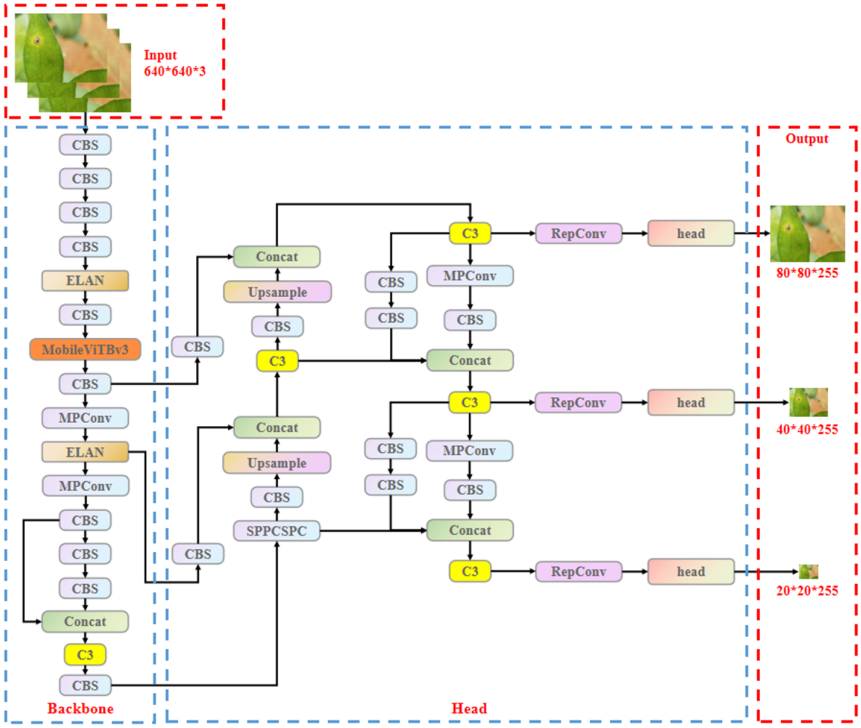

Figure 7. APD-YOLOv7 network structure.

3. Experiment and Result Analysis

3.1. Experimental Setup

In this study, we employed the open-source PyTorch deep learning framework for model development and enhancement. Additionally, the training and evaluation of these models were conducted on a system running Ubuntu 20. The hardware configuration comprised an Intel (R) Core (TM) i9-11900K processor and an NVIDIA GeForce RTX 3090 graphics card. The GPU was equipped with 24 GB of video memory and used CUDA version 11.3. During the experiments, consistent training parameters were applied as follows:

- The input image size was consistently adjusted to 640 × 640 pixels.
- The batch size parameter was set to 16.
- The epochs parameter was set to 300.
- The initial training weights were set to empty.
- All other parameters were left at their default settings.

3.2. Experimental Evaluation Metrics

When identifying targets of agricultural pests and diseases in complex environments, it is crucial to consider both the accuracy and real-time performance of the model. To evaluate the model's performance in identifying agricultural pests and diseases, this paper uses mean average precision as the primary metric, with single-image inference time and model parameter size serving as secondary evaluation criteria. It is worth noting that mean average precision is related to both precision and recall, and its detailed calculation formulas are provided below:

1. P represents precision, indicating the proportion of actually positive samples among those predicted as positive by the model.

$$P = \frac{TP}{TP + FP} \tag{10}$$

2. R represents recall, indicating the proportion of positive samples correctly predicted as positive by the model among all actual positive samples.

$$R = \frac{TP}{TP + FN} \tag{11}$$

3. AP represents average precision for a single category.

$$AP = \int_0^1 P(R)dR \tag{12}$$

4. mAP represents the mean of average precisions for all categories.

$$mAP = \frac{1}{N}\sum_{i=1}^{N} AP_i \tag{13}$$

In this context, TP denotes the count of samples accurately classified as positive, FP signifies the count of samples erroneously classified as positive, and FN indicates the count of samples mistakenly identified as negative.

Hence, AP75 refers to the AP computed with an IoU threshold of 0.75, while mAP75 signifies the mAP calculated at the same IoU threshold. For the purpose of this study, mAP75 serves as the primary evaluation metric.

In the subsequent experiments, we utilized the FIP6Set dataset, developed in this study, and implemented a 5-fold cross-validation approach across all experimental trials.

3.3. Experimental Analysis of Focal-DDRIoU Loss Function

We conducted experiments utilizing the YOLOv7 model, establishing baseline results by training the model with the default CIoU loss function. To assess the influence of various

IoU loss functions, we trained YOLOv7 using GIoU, DIoU, CIoU, EIoU, DDRIoU, and Focal-DDRIoU. To achieve this, we substituted the BBR IoU loss with L_{GIoU}, L_{DIoU}, L_{CIoU}, L_{EIoU}, L_{DDRIoU}, and $L_{Focal-DDRIoU}$, respectively.

As shown in Table 1, under the FIP6Set dataset, using a 5-fold cross-validation method and averaging the results of five experiments, DDRIoU loss and Focal-DDRIoU loss demonstrated superior performance compared to GIoU loss, DIoU loss, CIoU loss, and EIoU loss. Specifically, DDRIoU loss enhanced the mAP75 value by 0.76% over the default CIoU loss used in YOLOv7, while Focal-DDRIoU loss achieved a notable improvement of 1.04%.

Table 1. Performance comparison of YOLOv7 model trained with default loss (L_{CIoU}) and L_{GIoU}, L_{DIoU}, L_{EIoU}, L_{DDRIoU}, and $L_{Focal-DDRIoU}$ using 5-fold cross-validation method.

mAP75 / Loss	Rounds					5-Fold Average	5-Fold Variance
	1	2	3	4	5		
L_{GIoU}	0.561	0.548	0.536	0.555	0.588	0.5576 (−0.08%)	0.000300
L_{DIoU}	0.571	0.536	0.549	0.560	0.566	0.5564 (−0.20%)	0.000158
L_{CIoU}	0.544	0.542	0.560	0.560	0.586	0.5584 (0)	0.000249
L_{EIoU}	0.546	0.531	0.573	0.546	0.573	0.5538 (−0.46%)	0.000276
L_{DDRIoU}	0.566	0.556	0.555	0.571	0.582	0.5660 (+0.76%)	0.000100
$L_{Focal-DDRIoU}$	0.560	0.561	0.561	0.569	0.593	0.5688 (+1.04%)	0.000157

The improved performance can be primarily attributed to the introduction of the diagonal difference ratio in DDRIoU loss. This innovation addresses a limitation of other IoU loss functions, which produce the same loss value when the predicted and ground truth bounding boxes have the same aspect ratio but differ in size. DDRIoU loss more accurately captures these differences, guiding the model towards more precise optimization and enhancing detection accuracy.

Beyond improving accuracy, the diagonal difference ratio in DDRIoU loss also contributes to a more nuanced reflection of disparities between predicted and ground truth bounding boxes. This refinement aids in enhancing the stability and consistency of prediction results, thereby reducing outcome variance.

Compared to DDRIoU loss, Focal-DDRIoU loss exhibits a slight increase in variance, yet it maintains a relatively low level. This suggests that incorporating the focal loss concept allows the model to focus more effectively on challenging samples during training. As a result, it boosts performance without significantly compromising result stability. This design fosters a more robust and reliable model, especially in complex scenarios.

3.4. Experimental Analysis of Network Structure

In this section, a comparative experiment was conducted between the APD-YOLOv7 network structure devised in this study (hereinafter referred to as the APD network structure) and the original YOLOv7 network structure. The experimental steps are outlined below:

1. Initially, we modified only the IoU loss function. Specifically, we trained the YOLOv7 model using GIoU loss, DIoU loss, CIoU loss, EIoU loss, and Focal-DDRIoU loss, respectively, to evaluate the performance of the original YOLOv7 network structure.
2. Next, we substituted the original YOLOv7 network structure with the APD network structure.
3. Finally, we replicated the first step, adjusting the loss function to assess the performance of the APD network structure.

As demonstrated in Table 2, the performance of the APD network structure exceeds that of the YOLOv7 network structure when evaluated under equivalent IoU loss conditions. Following the substitution of the YOLOv7 network structure with the APD network structure, we conducted experiments employing GIoU loss, DIoU loss, CIoU loss, EIoU loss, and Focal-DDRIoU loss, respectively. The results showed an improvement in mAP75 by 0.68%, 0.68%, 0.78%, 0.6%, and 0.56%, respectively. In summary, the YOLOv7 model, when trained with Focal-DDRIoU loss and incorporating the APD-YOLOv7 network structure,

achieved a 2.06% improvement in mAP75 compared to the model trained with EIoU loss and the standard YOLOv7 architecture under identical conditions.

Table 2. Performance comparison between the YOLOv7 network structure and the APD network structure trained using L_{GIoU}, L_{DIoU}, L_{CIoU}, L_{EIoU}, and $L_{Focal-DDRIoU}$, respectively, under the YOLOv7 model.

Loss \ Evaluation	YOLOv7 5-Fold Average mAP75	APD-YOLOv7 5-Fold Average mAP75
L_{GIoU}	0.5576	0.5644 (+0.68%)
L_{DIoU}	0.5564	0.5632 (+0.68%)
L_{CIoU}	0.5584	0.5662 (+0.78%)
L_{EIoU}	0.5538	0.5598 (+0.60%)
$L_{Focal-DDRIoU}$	0.5688	0.5744 (+0.56%)

Table 3 illustrates that, in comparison to the YOLOv7 network structure, the APD network structure achieves a 3.9MB parameter reduction and a 9.74% decrease in giga floating-point operations per second (GFLOPs), effectively lowering the model's computational demands and parameter count.

Table 3. Comparison of parameters between YOLOv7 network structure and APD network structure (batch = 1).

Model	Parameters (MB)	GFLOPs	Inference Time (ms)
YOLOv7	36.9	104.7	7.2
APD-YOLOv7	33.0	94.5	10.8

This significant improvement is primarily attributed to modifications made to the YOLOv7 network structure. We adjusted the YOLOv7 architecture by removing redundant modules, appropriately supplementing with C3 modules, and embedding MobileViTv3 modules. These modifications successfully reduced the model's parameter count and computational complexity while further enhancing the model's accuracy.

Although the inference time for a single image has increased relatively, in agricultural pest and disease identification, strict real-time performance is not typically required. As long as the identification results can be provided within a few seconds, it is sufficient for most application scenarios, thus still meeting the need for rapid identification of agricultural pests and diseases.

3.5. Experimental Analysis of the APD-YOLOv7 Model

We conducted a performance evaluation of other widely used object detection models on the FIP6Set dataset.

Referring to Table 4, the APD-YOLOv7 model presented in this study demonstrates significant superiority over the YOLOv7 and YOLOv5-L models in terms of detection accuracy. Specifically, the APD-YOLOv7 model achieves an mAP75 metric that is 1.86% and 1.6% higher than the YOLOv5-L and YOLOv7 models, respectively. Additionally, it reduces the model parameter count by 13.5 MB and 3.9 MB compared to YOLOv5-L and YOLOv7, respectively. In terms of model detection speed, the inference time for a single image using the APD-YOLOv7 model is 10.8 ms. Although the inference time of the APD-YOLOv7 model is slightly longer compared to the YOLOv7 and YOLOv5-L models, it still satisfies the requirements for swift identification of agricultural pests and diseases.

Table 4. Performance comparison between APD-YOLOv7 and other object detection models.

Model	5-Fold Average mAP75	Parameters (MB)	Inference Time (ms)
YOLOv5-L	0.5558	46.5	9.2
YOLOv7	0.5584	36.9	7.2
APD-YOLOv7	0.5744	33.0	10.8

4. Conclusions and Future Directions

In this study, we present a model, termed APD-YOLOv7, for the detection of agricultural pests and diseases. Firstly, we design a novel loss function, termed Focal-DDRIoU loss, to enhance the efficiency and accuracy of BBR. Secondly, we introduce C3 and MobileViTv3 modules to modify the YOLOv7 network structure. Specifically, the MobileViTv3 module is integrated into the 17th layer of the backbone, thereby reducing the computational complexity, parameter count, and floating-point operations of the model, while simultaneously enhancing its recognition accuracy.

The performance evaluation is conducted on the FIP6Set dataset, employing mAP75 as the evaluation metric. Experimental results indicate that the APD-YOLOv7 model surpasses mainstream models, such as YOLOv7 and YOLOv5-L, by achieving improvements of 1.6% and 1.86%, respectively. Additionally, our proposed Focal-DDRIoU loss outperforms popular loss functions, including GIoU, DIoU, CIoU, and EIoU, yielding improvements of 1.12%, 1.24%, 1.04%, and 1.5%, respectively.

Although the inference time for a single image is increased in the APD-YOLOv7 model, it still fulfills the criteria for swift identification of agricultural pests and diseases.

While the APD-YOLOv7 model strikes a good balance between recognition accuracy and speed, its recognition diversity remains to be further explored. Future research efforts should focus on training the APD-YOLOv7 model with additional pest and disease samples to ascertain its robustness. Moreover, as the investigation of Focal-DDRIoU loss is currently confined to the FIP6Set dataset, future studies should seek to apply it across diverse datasets to assess its generalizability.

Author Contributions: Conceptualization, J.L., S.L., and C.L.; methodology, J.L., S.L., C.L., and S.Z.; software, J.L. and S.Z.; validation, J.L. and D.C.; formal analysis, J.L. and D.C.; investigation, J.L., S.L., and S.Z.; resources, J.L. and C.L.; data curation, J.L. and S.L.; writing—original draft preparation, J.L. and D.C.; writing—review and editing, J.L. and S.L.; visualization, J.L., S.L., and S.Z.; supervision, J.L., S.L., D.C., S.Z., and C.L.; project administration, J.L., S.L., S.Z., and C.L.; funding acquisition, S.L., S.Z., and C.L. All authors have read and agreed to the published version of the manuscript.

Funding: This work was funded in part by The Basic Ability Enhancement Program for Young and Middle-aged Teachers of Guangxi (2022KY0381), Guangxi Key Research and Development Project (Application Number: 2023AB19026, Research and demonstration of rapid monitoring technology and early warning system for major citrus diseases and pests based on deep learning).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The public IP102 dataset can be downloaded from https://github.com/xpwu95/IP102 (accessed on 15 May 2023). Publicly available web data were used in this study and can be accessed without restrictions. The FIP6Set dataset, which is a combination of the IP102 dataset, publicly available web data, and private data, used and/or analyzed during the current study is available from the corresponding author upon reasonable request.

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

References

1. Manavalan, R. Automatic identification of diseases in grains crops through computational approaches: A review. *Comput. Electron. Agric.* **2020**, *178*, 105802. [CrossRef]
2. Kong, J.; Wang, H.; Wang, X.; Jin, X.; Fang, X.; Lin, S. Multi-stream hybrid architecture based on cross-level fusion strategy for fine-grained crop species recognition in precision agriculture. *Comput. Electron. Agric.* **2021**, *185*, 106134. [CrossRef]

3. Zheng, Y.Y.; Kong, J.L.; Jin, X.B.; Wang, X.Y.; Su, T.L.; Zuo, M. CropDeep: The crop vision dataset for deep-learning-based classification and detection in precision agriculture. *Sensors* **2019**, *19*, 1058. [CrossRef] [PubMed]
4. Li, Y.; Nie, J.; Chao, X. Do we really need deep CNN for plant diseases identification? *Comput. Electron. Agric.* **2020**, *178*, 105803. [CrossRef]
5. Shruthi, U.; Nagaveni, V.; Raghavendra, B.K. A review on machine learning classification techniques for plant disease detection. In Proceedings of the 2019 5th International Conference on Advanced Computing & Communication Systems (ICACCS), Coimbatore, India, 15–16 March 2019; IEEE: Piscataway, NJ, USA, 2019; pp. 281–284.
6. Shrivastava, V.K.; Pradhan, M.K. Rice plant disease classification using color features: A machine learning paradigm. *J. Plant Pathol.* **2021**, *103*, 17–26. [CrossRef]
7. Kim, K.G. Book review: Deep learning. *Healthc. Inform. Res.* **2016**, *22*, 351. [CrossRef]
8. Li, L.; Zhang, S.; Wang, B. Plant disease detection and classification by deep learning—A review. *IEEE Access* **2021**, *9*, 56683–56698. [CrossRef]
9. Oppenheim, D.; Shani, G.; Erlich, O.; Tsror, L. Using deep learning for image-based potato tuber disease detection. *Phytopathology* **2019**, *109*, 1083–1087. [CrossRef] [PubMed]
10. Chen, Q.; Wang, Y.; Yang, T.; Zhang, X.; Cheng, J.; Sun, J. You only look one-level feature. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 19–25 June 2021; IEEE: Piscataway, NJ, USA, 2021; pp. 13039–13048.
11. Redmon, J.; Farhadi, A. YOLO9000: Better, faster, stronger. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; IEEE: Piscataway, NJ, USA, 2017; pp. 7263–7271.
12. Redmon, J.; Farhadi, A. Yolov3: An incremental improvement. *arXiv* **2018**, arXiv:1804.02767.
13. Bochkovskiy, A.; Wang, C.Y.; Liao, H.Y.M. Yolov4: Optimal speed and accuracy of object detection. *arXiv* **2020**, arXiv:2004.10934.
14. YOLOv5: An Open-Source Object Detection Algorithm. Available online: https://github.com/ultralytics/yolov5 (accessed on 15 May 2023).
15. Wang, C.Y.; Bochkovskiy, A.; Liao, H.Y.M. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 18–22 June 2023; IEEE: Piscataway, NJ, USA, 2023; pp. 7464–7475.
16. Wang, C.Y.; Yeh, I.H.; Liao, H.Y.M. YOLOv9: Learning What You Want to Learn Using Programmable Gradient Information. *arXiv* **2024**, arXiv:2402.13616.
17. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich feature hierarchies for accurate object detection and semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; IEEE: Piscataway, NJ, USA; pp. 580–587.
18. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards real-time object detection with region proposal networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2016**, *39*, 1137–1149. [CrossRef] [PubMed]
19. He, K.; Gkioxari, G.; Dollár, P.; Girshick, R. Mask r-cnn. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; IEEE: Piscataway, NJ, USA, 2017; pp. 2961–2969.
20. Zhang, H.; Chang, H.; Ma, B.; Wang, N.; Chen, X. Dynamic R-CNN: Towards high quality object detection via dynamic training. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020; Springer: Berlin, Germany, 2020; pp. 260–275.
21. Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; Zagoruyko, S. End-to-end object detection with transformers. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020; Springer: Berlin, Germany, 2020; pp. 213–229.
22. Yu, J.; Jiang, Y.; Wang, Z.; Cao, Z.; Huang, T. Unitbox: An advanced object detection network. In Proceedings of the 24th ACM international conference on Multimedia, Amsterdam, The Netherlands, 15–19 October 2016; ACM: New York, NY, USA, 2016; pp. 516–520.
23. Tychsen-Smith, L.; Petersson, L. Improving object localization with fitness nms and bounded iou loss. In Proceedings of the IEEE conference on computer vision and pattern recognition., Salt Lake City, UT, USA, 18–22 June 2018; IEEE: Piscataway, NJ, USA, 2018; pp. 6877–6885.
24. Rezatofighi, H.; Tsoi, N.; Gwak, J.; Sadeghian, A.; Reid, I.; Savarese, S. Generalized intersection over union: A metric and a loss for bounding box regres-sion. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; IEEE: Piscataway, NJ, USA, 2019; pp. 658–666.
25. Zheng, Z.; Wang, P.; Liu, W.; Li, J.; Ye, R.; Ren, D. Distance-IoU loss: Faster and better learning for bounding box regression. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; AAAI: Menlo Park, CA, USA, 2020; Volume 34, pp. 12993–13000.
26. Zheng, Z.; Wang, P.; Ren, D.; Liu, W.; Ye, R.; Hu, Q.; Zuo, W. Enhancing geometric factors in model learning and inference for object detection and instance segmentation. *IEEE Trans. Cybern.* **2021**, *52*, 8574–8586. [CrossRef] [PubMed]
27. Zhang, Y.-F.; Ren, W.; Zhang, Z.; Jia, Z.; Wang, L.; Tan, T. Focal and efficient IOU loss for accurate bounding box regression. *arXiv* **2021**, arXiv:2101.08158. [CrossRef]

28. Wu, X.; Zhan, C.; Lai, Y.K.; Cheng, M.M.; Yang, J. Ip102: A large-scale benchmark dataset for insect pest recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; IEEE: Piscataway, NJ, USA, 2019; pp. 8787–8796.
29. Shahinfar, S.; Meek, P.; Falzon, G. "How many images do I need?" Understanding how sample size per class affects deep learning model performance metrics for balanced designs in autonomous wildlife monitoring. *Ecol. Inform.* **2020**, *57*, 101085. [CrossRef]
30. Wang, C.Y.; Yeh, I.H.; Liao, H.Y.M. You only learn one representation: Unified network for multiple tasks. *arXiv* **2021**, arXiv:2105.04206.
31. Ge, Z.; Liu, S.; Wang, F.; Li, Z.; Sun, J. Yolox: Exceeding yolo series in 2021. *arXiv* **2021**, arXiv:2107.08430.
32. Wadekar, S.N.; Chaurasia, A. Mobilevitv3: Mobile-friendly vision transformer with simple and effective fusion of local, global and input features. *arXiv* **2022**, arXiv:2209.15159.