agronomy

Article

Recognition Model for Tea Grading and Counting Based on the Improved YOLOv8n

Yuxin Xia [1,2], Zejun Wang [2], Zhiyong Cao [2], Yaping Chen [2], Limei Li [2], Lijiao Chen [2], Shihao Zhang [1,2], Chun Wang [1,2], Hongxu Li [2] and Baijuan Wang [2,*]

[1] College of Mechanical and Electrical Engineering, Yunnan Agricultural University, Kunming 650201, China; 15140964046@163.com (Y.X.); 18637905872@163.com (S.Z.); chunwangkk@163.com (C.W.)
[2] Yunnan Organic Tea Industry Intelligent Engineering Research Center, Kunming 650201, China; wangzejun0529741x@163.com (Z.W.); czy@ynau.edu.cn (Z.C.); cyp83@ynau.edu.cn (Y.C.); 18214178910@163.com (L.L.); 2015056@ynau.edu.cn (L.C.); 13085315910@163.com (H.L.)
* Correspondence: wangbaijuan2023@163.com

Abstract: Grading tea leaves efficiently in a natural environment is a crucial technological foundation for the automation of tea-picking robots. In this study, to solve the problems of dense distribution, limited feature-extraction ability, and false detection in the field of tea grading recognition, an improved YOLOv8n model for tea grading and counting recognition was proposed. Firstly, the SPD-Conv module was embedded into the backbone of the network model to enhance the deep feature-extraction ability of the target. Secondly, the Super-Token Vision Transformer was integrated to reduce the model's attention to redundant information, thus improving its perception ability for tea. Subsequently, the loss function was improved to MPDIoU, which accelerated the convergence speed and optimized the performance. Finally, a classification-positioning counting function was added to achieve the purpose of classification counting. The experimental results showed that, compared to the original model, the precision, recall and average precision improved by 17.6%, 19.3%, and 18.7%, respectively. The average precision of single bud, one bud with one leaf, and one bud with two leaves were 88.5%, 89.5% and 89.1%. In this study, the improved model demonstrated strong robustness and proved suitable for tea grading and edge-picking equipment, laying a solid foundation for the mechanization of the tea industry.

Keywords: tea grading; counting statistics; Super-Token Vision Transformer; MPDIoU

Citation: Xia, Y.; Wang, Z.; Cao, Z.; Chen, Y.; Li, L.; Chen, L.; Zhang, S.; Wang, C.; Li, H.; Wang, B. Recognition Model for Tea Grading and Counting Based on the Improved YOLOv8n. *Agronomy* **2024**, *14*, 1251. https://doi.org/10.3390/agronomy14061251

Academic Editor: Mario Cunha

Received: 11 May 2024
Revised: 28 May 2024
Accepted: 6 June 2024
Published: 10 June 2024

1. Introduction

Modern agriculture has evolved towards automation and mechanization [1], leading to an increased demand for intelligent equipment [2]. China is the world's largest producer and consumer of tea (https://www.thepaper.cn/newsDetail_forward_26661265, accessed on 13 March 2024). It has established a modern tea industry system with coordinated development of a tea culture, tea industry, tea technology, and tea ecology. Yunnan Province, with its rich national tea culture resources and famous mountain ancient tree resources [3], has maintained a steady and increasing trend in planting area, making tea its "golden signboard" [4]. Currently, tea harvesting in Yunnan Province relies on manual labor, resulting in low efficiency and high labor cost [5]. At present, most of the existing tea-picking machines are hand-held, suitable for small-area tea gardens. However, the freshly picked are often mixed, making subsequent processing difficult and compromising the production and processing of high-quality tea. This compromises the integrity and availability of tea, causing economic losses, to a certain extent [6]. Therefore, developing a tea grading algorithm that considers both detection accuracy and speed for edge-device target detection is particularly important.

A traditional target-detection algorithm [7] could only extract shallow features based on color, texture, etc., and its generalization ability and robustness were poor, limiting their

application [8]. With the rise of artificial intelligence, object-detection algorithms based on deep learning [9] have gradually entered the field. These algorithms can achieve automatic selection and extraction of features, effectively improving the quality and efficiency of features. At present, the YOLO (You Only Look Once) series with better performance has been widely used in agriculture. Camacho et al. [10] compared the R-CNN network and the YOLOv8 model to identify and detect tomatoes with different maturity in the same data set, and the results showed that the YOLOv8 model had better performance; Shuang et al. [11] combined deformable convolution, attention mechanism and improved spatial pyramid pooling into the YOLOv8 model to detect the tea bud target, which enhanced the ability of the model to learn the invariance of complex targets. The improved model's mAP reached 88.27%; Trinh et al. [12] proposed an improved YOLOv8 to improve the performance of the rice leaf disease detection system, and its accuracy rate was as high as 89.9%; Li et al. [13] integrated the designed Transformer architecture into the YOLOv7 model, and the improved average recognition precision value for one bud with one leaf of tea reached 90%.

In summary, in order to effectively solve the problems of dense distribution, limited feature extraction ability, missed detection, and false detection in the field of tea grading recognition, based on the YOLOv8n model, this study successfully realized a high precision and strong robustness for tea grading and the counting detection model. (1) Firstly, the SPD-Conv module [14] was used to replace the convolution module of the original model's backbone network, so as to improve detection performance of the model when the image resolution was low or the detection object was small. (2) Secondly, the Super-Token Vision Transformer [15] was introduced to strengthen the model's ability to extract key information in the feature map, thereby further improving its precision. (3) Subsequently, the CIoU in the loss function was replaced with a more suitable MPDIoU [16] to improve the regression effect of the detection bounding box and accelerate the model's convergence speed. (4) Finally, the improved YOLOv8n added an automatic counting function to enable the quality grading statistics of tea, providing a new approach for estimating tea yield. This study proposed a high-precision, lightweight YOLOv8n algorithm, laying a solid foundation for the further deployment of tea-picking robots.

2. Materials and Methods

2.1. Data Collection

In this study, tea grading was divided into three categories: single bud, one bud with one leaf, and one bud with two leaves. The image data were collected from Hekai Base of Yuecheng Technology Co., Ltd., located in Menghai County, Xishuangbanna Prefecture, Yunnan Province. The specific location is shown in Figure 1.

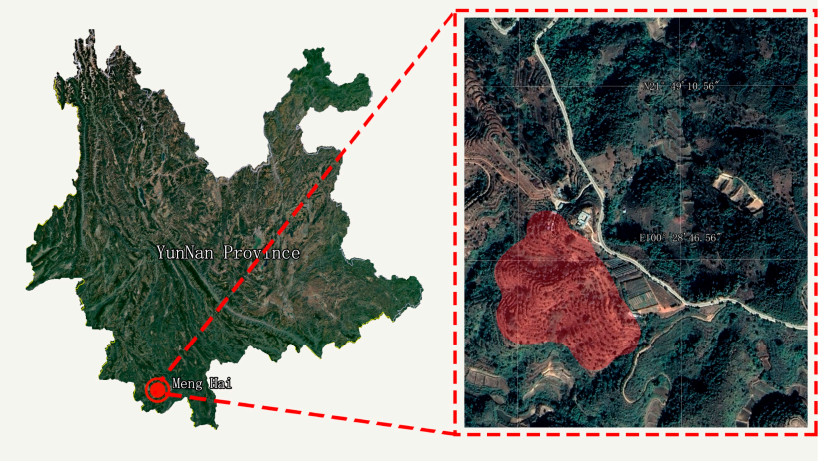

Figure 1. Display of data acquisition site.

During the image capture process, the photographers simulated the positioning relationship between the tea-picking robot and the tea tree. The distance between them was set to be 55–85 cm, with the camera positioned 20–30 cm from the tender shoots of the tea tree, and approximately 120 cm above the ground. The SONY-α6400 camera (Origin, Saitama, Japan) was used to collect data from multiple angles. During the image acquisition stage, data were collected under different lighting conditions, at various times of the day, and in different weather conditions. This ensured the richness and comprehensiveness of the dataset, provided a more realistic tea garden environment for model training, and contributed to improving the robustness of the model. In this study, a total of 1915 images of single bud, one bud with one leaf, and one bud with two leaves of Yunnan large-leaf tea trees were collected. In order to improve the quality of the data set, 826 images were selected as the original data set after strict screening. As shown in Figure 2, to address the issue of model performance degradation due to sample imbalance, this study applied data augmentation methods such as flipping, contrast enhancement, and rotation to expand the fresh tea-leaf images. Finally, a total of 3612 datasets were obtained, effectively preventing the phenomenon of overfitting during training and enhancing the generalization ability of the training model.

Figure 2. Data enhancement comparison.

Labeling was used for manual annotation [17]. The tag box was positioned as centrally as possible in the image, and the tag file were created in both txt and xml formats. The dataset was randomly divided into training set (2168 images), validation set (722 images) and test set (722 images) in a 6:2:2 ratio. The specific data information for the training set is shown in Figure 3, including the number of classifications, the size and quantity distribution of the bounding box, the position of the center points relative to the whole image, and the aspect ratio of the target relative to the whole image. In the figure, A represents a single bud, B represents one bud with one leaf, and C represents one bud with two leaves.

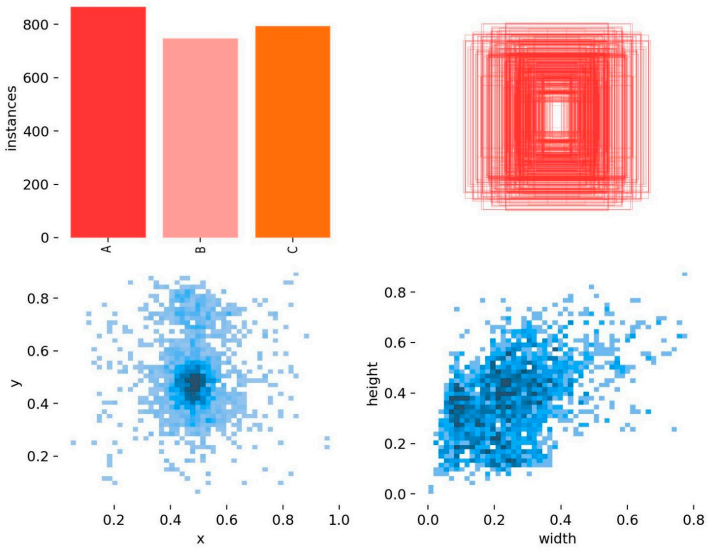

Figure 3. Data information of training set.

2.2. Model Improvements

In this study, an improved YOLOv8n model was proposed to address the problems of dense distribution, limited feature-extraction ability, missed detection and false detection in tea grading recognition. The specific improvements are as follows: (1) The SPD-Conv module replaced the convolution layer of the backbone network part, enhancing the model's feature extraction ability for tea. (2) The Super-Token Vision Transformer was integrated to improve the model's perception of small targets, effectively enhancing model performance. (3) The original loss function was optimized to MPDIoU, reducing loss and accelerating model convergence. (4) An automatic counting function was added, enabling the model to automatically calculate the number of single buds, one bud with one leaf, and one bud with two leaves in the recognition image. The specific network structure is shown in Figure 4.

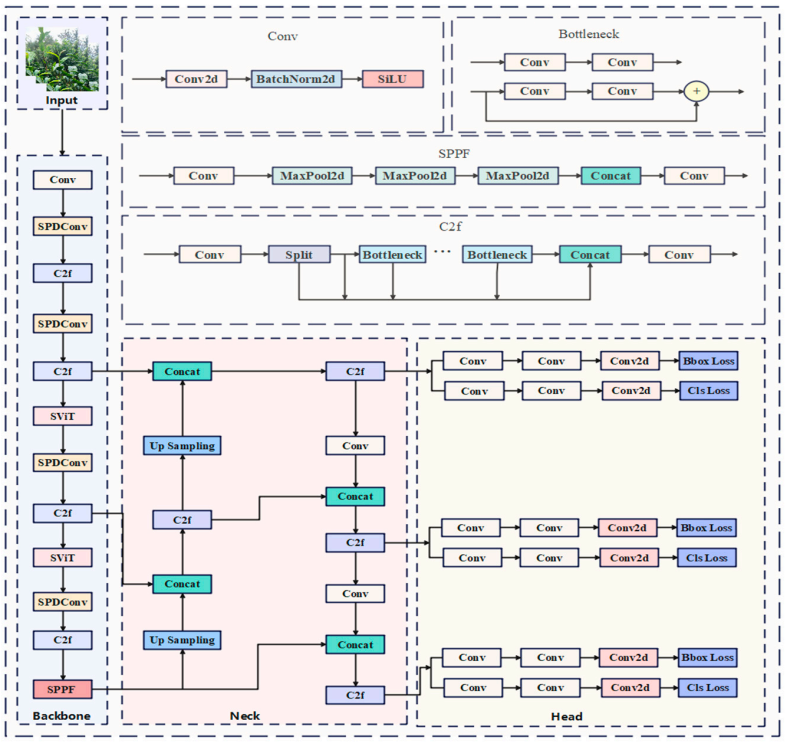

Figure 4. Improved YOLOv8 network structure diagram.

2.2.1. Introduction of YOLOv8

YOLO is a one-stage target-detection algorithm. Its core idea is to divide the image by region and to predict [18]. Compared with other models, its significant feature is fast and efficient training [19]. With the continuous evolution of the YOLO model, the performance is also improving [20]. The input feature map size of the YOLOv8 model is 640 × 640, and its backbone feature-extraction network is composed of the CBS module, C2f module and SPPF (Serial Parallel Pooling Fusion) module [21]. The C2f module integrates the design concepts of the C3 module and the Efficient Lightweight Attention Network (ELAN), consisting of Conv, Split and BottleNeck components [22]. This design enables the module to capture rich gradient-flow information while maintaining the model's light weight, significantly enhancing feature fusion capabilities, accelerating inference speed, and achieving model lightweightness. In order to further enhance model performance, the SPPF module is introduced, which is designed based on the concept of SPP (Spatial Pyramid Pooling) [23]. Serving as a spatial pyramid pooling module, the SPPF module boasts lower parameter count and computational load. When incorporated into the C2f module, it facilitates feature map pooling on various scales, effectively expanding the model's receptive field. This enables extraction of richer feature information without adding computational burden, thereby enhancing model recognition precision. The Head layer adopts a decoupled head design, replacing the traditional coupled head. Each scale is equipped with an independent detector, comprised of a set of convolutional and fully connected layers. This configuration enables the capturing of target information on different scales [24]. It effectively improves target detection precision and mitigates issues such as inaccurate positioning and classification errors in complex scenes.

The research and application scenarios of traditional visual recognition models are too limited to cope with the detection requirements in complex and diverse backgrounds, while the YOLOv8 model exhibits poor performance in detection precision. To enhance model recognition precision and meet real-time detection requirements, this study made improvements based on YOLOv8n, incorporating the SPD-Conv module, the Super-Token Vision Transformer, changing the loss function to MPDIoU, and the addition of classification recognition and counting functions.

2.2.2. SPD-Conv Module

The traditional neural network model exhibits poor generalization ability when extracting low-resolution and small targets [25]. This is primarily due to the continuous downsampling of the feature map as the model deepens through the network hierarchy, resulting from calculations in the convolutional and pooling layers. As a consequence, fine-grained information becomes challenging to capture effectively during feature extraction, leading to diminished detection accuracy and robustness. In YOLOv8, the feature-extraction module Conv employs a step-size convolutional layer, which poses challenges in tasks involving low image resolution or small detection objects, often resulting in detection performance degradation. In order to overcome this challenge, this study introduced a combination of space-to-depth layer and non-strided convolution layer, replacing the step convolution layer in the original model's backbone network. This improvement enhanced the model's performance in processing low-resolution images and complex small targets, allowing for more effective expression of image features, particularly in tea leaf recognition, and thereby improving the model's robustness.

The space-to-depth layer in the SPD-Conv module [26] partitions the input feature map into subgraph sets in a parameter-scale factor-controlled manner. The subsequent non-straddle convolution layer is designed to maximize the preservation of these subgraph sets, effectively reducing information loss and improving the detection precision of small target areas.

The space-to-depth layer transforms the original feature into an intermediate feature map containing feature identification information. Considering the intermediate feature mapping of any size $S \times S \times C_1$, the sub-feature maps are obtained by sampling in the x and

y dimensions, according to the set step-size scale. The sequence of sub-feature mappings is then sliced, and the principle formula for the calculation process is as follows (1):

$$\begin{cases} f_{0,0} = X[0:M:S, 0:M:S], \ldots f_{0,s-1} = X[0:M:S, S-1:M:S] \\ f_{0,1} = X[1:M:S, 0:M:S], \ldots f_{1,s-1} = X[1:M:S, S-1:M:S] \\ f_{1,0} = X[1:M:S, 0:M:S], \ldots f_{1,s-1} = X[1:M:S, S-1:M:S] \\ f_{s-1,0} = X[S-1:M:S, 0:M:S], \ldots f_{s-1,s-1} = X[S-1:M:S, S-1:M:S] \end{cases} \quad (1)$$

When the scale factor is set to 2, the width and height of the output feature map will be halved compared to the input, indicating that the sub-images are obtained by downsampling the input feature map twice. The four sub-images are denoted as $f_{0,0}$, $f_{0,1}$, $f_{1,0}$, and $f_{1,1}$. The principle formula for the calculation process is as follows (2):

$$\begin{cases} f_{0,0} = f[0:S:scale, 0:S:scale] \\ f_{0,1} = f[0:S:scale, 1:S:scale] \\ f_{1,0} = f[1:S:scale, 0:S:scale] \\ f_{1,1} = f[1:S:scale, 1:S:scale] \end{cases} \quad (2)$$

As illustrated in Figure 5, following the SPD feature conversion layer, a non-strided convolution is added to preserve all discriminant feature information to the fullest extent. The four sub-feature maps are then concatenated to form a feature size of $\frac{S}{2} \times \frac{S}{2} \times C_2$, where each sub-image has the same number of channels as the input feature map. The resultant sub-feature images are combined via standard convolution [27] to ensure efficient information extraction during the downsampling process, thus achieving the objective of enhancing detection precision.

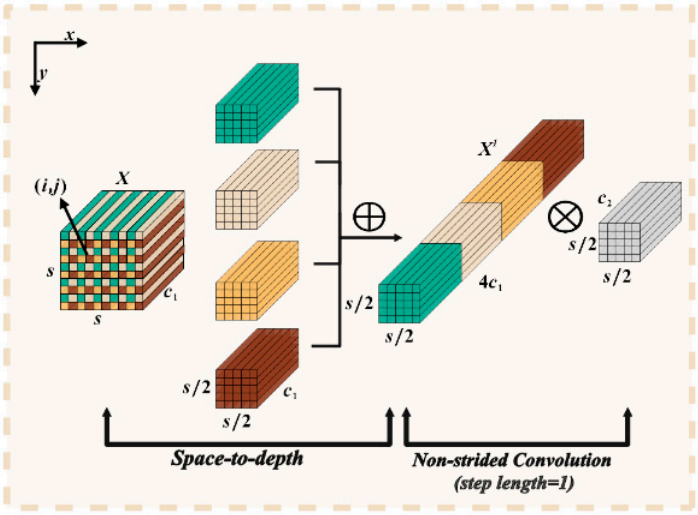

Figure 5. SPD-Conv structure diagram.

2.2.3. The Super-Token Vision Transformer (SViT)

Attention mechanism [28] is a resource allocation scheme. In the case of limited computing power, it is preferentially assigned to more important tasks to effectively solve the problem of information overload. In neural network learning, the attention mechanism is introduced to make the model focus on the more critical information of the current task among many types of input information, reduce the attention to other information, and even filter out redundant information, so as to solve the problem of information overload and improve the efficiency and accuracy of task processing. The attention mechanism is widely used in deep learning algorithms. In order to bolster the YOLOv8 model's hierarchical detection capabilities for tea targets and prioritize key targets amidst a complex tea-garden background, the scalability and robustness of the model are improved. This study introduced a Super-Token Vision Transformer, whose structure is shown in Figure 6.

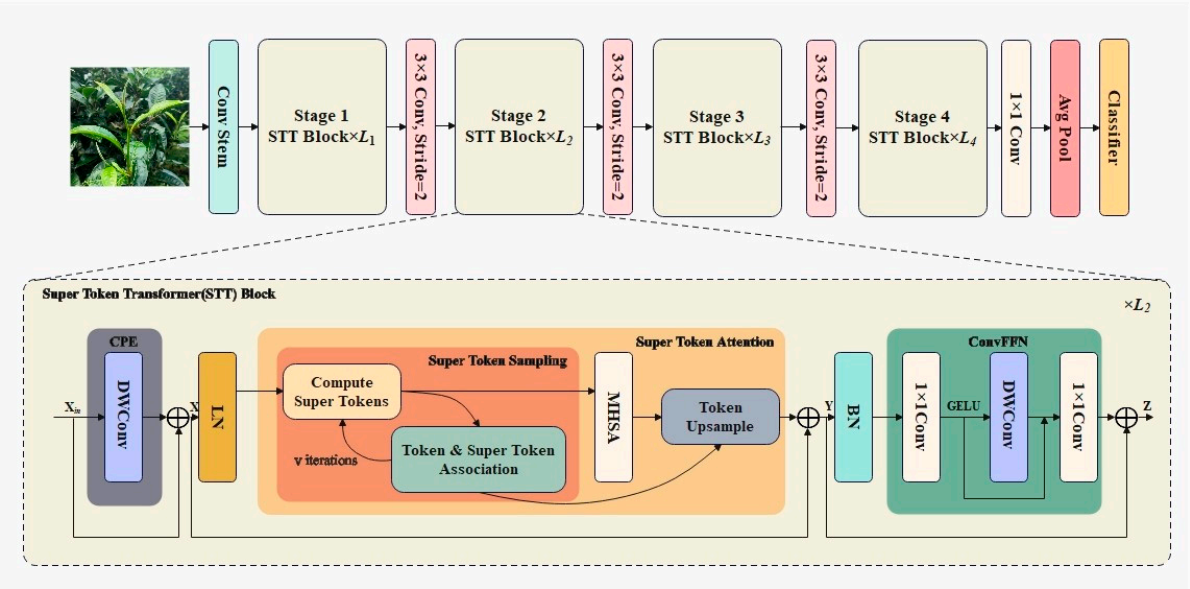

Figure 6. The architecture of the Super-Token Vision Transformer (SViT).

In the study, a visual hierarchical hybrid structure incorporating convolutional layers is devised to capture the constraints of comprehensive features through convolutional layer compensation. The Super-Token Vision Transformer comprises a convolution module, an STT Block module, an average pooling layer and classifier, facilitating efficient learning, enhancing the model's feature-extraction capabilities and then improving the detection accuracy of the model, thereby boosting detection precision. The Super-Token Vision Transformer's hierarchical visual converter comprises three key modules: convolutional position embedding (CPE), Super-Token Sampling (STA) and the convolutional feedforward network (Conv FFN). The formulations for each component are as follows:

$$X = CPE(X_{in}) + X_{in} \tag{3}$$

$$Y = STA(LN(X)) + X \tag{4}$$

$$Z = ConvFFN(BN(Y)) + Y \tag{5}$$

As shown in Figure 6 above, Depth-wise Separable Convolution plays a crucial role in the CPE module and is particularly suited for lightweight detection models. The CPE structure is shown in Figure 7. It comprises Depth-wise Conv [29] and Pointwise Conv [30], with its structure illustrated in Figure 6, below. Unlike traditional convolution, Depth-wise Conv operates uniquely by independently convoluting each channel of the input layer, with each convolution kernel responsible for a single channel. Upon completion, the number of feature maps remains identical to the number of channels in the input layer, with the convolution operation of each channel executed in parallel. In the subsequent Pointwise Conv stage, a 1×1 convolution kernel is utilized to convolve the output of deep convolution, integrating and interacting with the feature-map information generated independently by each channel. This process yields a new feature map as output, aimed at enhancing the model's learning ability concerning the correlation between different channels.

The STA module is adept at learning global information, with its specific structure depicted in Figure 6 above. A fast sampling method was used to anticipate the Super-Token Vision Transformer by discerning the sparse relationship between labels and super labels. Subsequently, multi-headed self-attention operations are performed within the Super-Token Vision Transformer space to capture the dependencies among Super-Token Vision Transformer for modeling purposes. Ultimately, the information from the Super-Token Vision Transformer is re-mapped back to the original label space through an up-

sampling algorithm, effectively minimizing redundant features. The formulas are presented in (6) and (7), as follows:

$$\Omega(STA) = \Omega(STS) + \Omega(MHSA) + \Omega(TU) \tag{6}$$

$$STA(X) = Q(A(S)(\hat{Q}^T)W_v) = \widetilde{A}(X)V(X) \tag{7}$$

where $\widetilde{A}(X) = QA(S)\hat{Q}^T$ is the corresponding attention map of the input label.

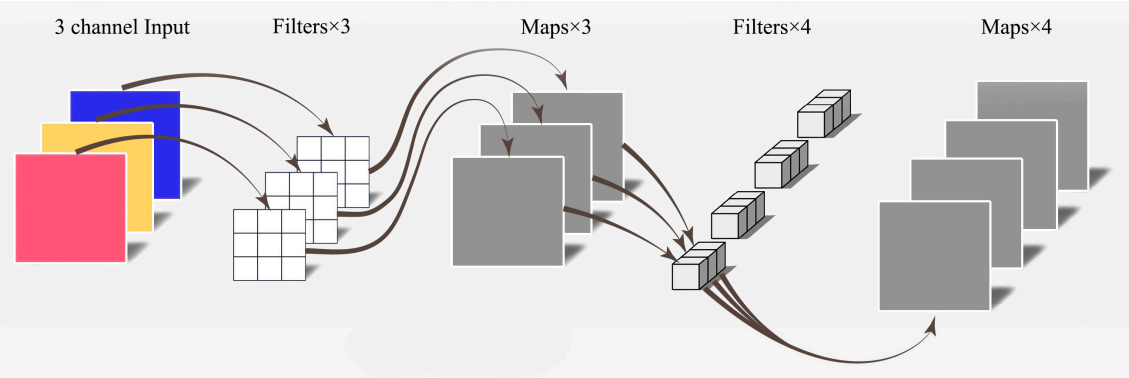

Figure 7. CPE module structure diagram.

As a common and critical deep learning model, the structure of Conv FFN (Convolutional Feedforward Neural Network) is shown in Figure 6 above, and the expression is given in Equation (8):

$$FFN(x) = f\left(x \cdot W_1^T\right) \cdot W_2 \tag{8}$$

It has excellent representation learning capabilities and has been widely used in various fields [31]. First, each input channel is weighted and summed by the convolution kernel to adjust the feature dimensions. Second, in order to enhance the representativeness of local features, independent convolution operations are performed on each input channel. Finally, a 1×1 convolution operation is performed again to further adjust the dimensions and combination of the features.

2.2.4. Improvement of Loss Function

The loss function of YOLOv8 is composed of location loss and classification loss [32]. Unlike previous models that include confidence loss calculation, YOLOv8 uses the Binary Cross-Entropy (BCE) loss from YOLOv5 for classification loss. The location loss is composed of DFL (Distribution Focal Loss) and CIoU (Complete IoU) [33], with the expression shown in Equation (9). Among them, IoU (Intersection over Union) is a straightforward function for calculating location loss by evaluating the overlap of two bounding boxes. In the context of actual tea garden picking, the complex background often leads to occlusion issues among fresh tea leaves. To address the problems of missed detection and false detection in such settings, this enhancement accelerates the model's convergence speed, optimizes performance, and improves detection precision.

$$LOSS_{CIoU} = 1 - IoU + \frac{\rho^2\left(b, b^{gt}\right)}{c^2} + \alpha v \tag{9}$$

Here, b and b^{gt} denote the prediction box and the true box, respectively; $\rho^2\left(b, b^{gt}\right)$ represents the Euclidean distance between them, and is the diagonal distance of the minimum enclosing region that can contain both the prediction box and the true box. V and α are

the evaluation parameters and balance factors of the length–width ratio, respectively. The formulas are shown in Formulas (10) and (11):

$$v = \frac{4}{x^2}\left(arctan\frac{w^{gt}}{h^{gt}} - arctan\frac{w}{h}\right)^2 \tag{10}$$

$$\alpha = \frac{v}{1 - IoU + v} \tag{11}$$

CIoU only considers the intersection area of the bounding box, the distance from the center point, and the aspect ratio of the bounding box [34]. In this study, the MPDIoU loss function is used for the regression of overlapping and non-overlapping bounding boxes [35]. It takes into account the distance from the center point and the deviation of the width and height. MPDIoU uses the bounding-box similarity measure based on the minimum point distance. The expressions are shown in Equations (12)–(15), and the structure of the improved loss function is illustrated in Figure 8.

$$\mathcal{L}_{MPDIoU} = 1 - MPDIoU \tag{12}$$

$$MPDIoU = \frac{A \cap B}{A \cup B} - \frac{d_1^2}{w^2 + h^2} - \frac{d_2^2}{w^2 + h^2} \tag{13}$$

$$d_1^2 = \left(x_1^B - x_1^A\right)^2 + \left(y_1^B - y_1^A\right)^2 \tag{14}$$

$$d_2^2 = \left(x_2^B - x_2^A\right)^2 + \left(y_2^B - y_2^A\right)^2 \tag{15}$$

Here, A and B represent the prediction box and the real box, respectively. (x_1^A, y_1^A) and (x_2^A, y_2^A) represent the upper-left and lower-right corner coordinates of bounding box A, respectively. Similarly, (x_1^B, y_1^B) and (x_2^B, y_2^B) represent the upper-left and lower-right corner coordinates of bounding box B, respectively.

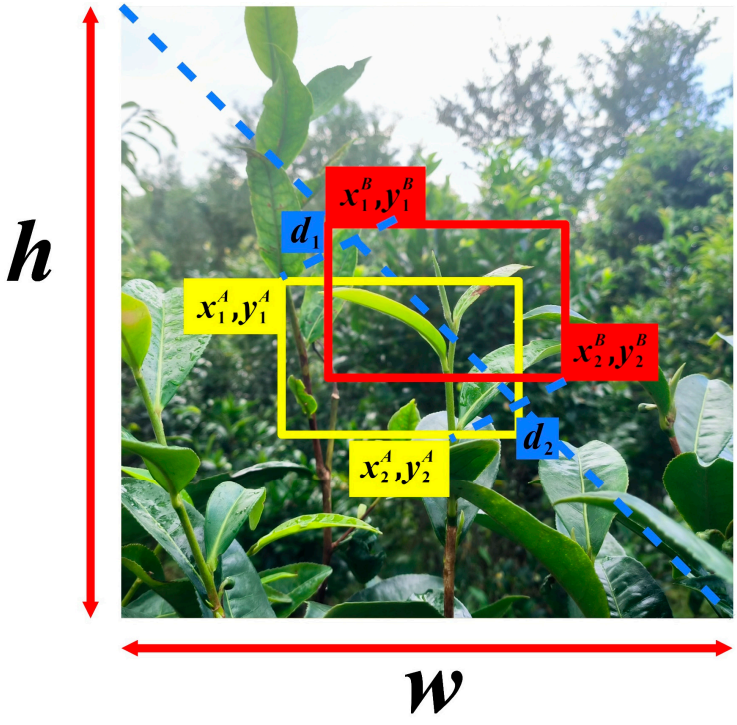

Figure 8. MPDIoU structure diagram.

2.3. Evaluation Index

In this study, precision, recall and mean average precision were selected as the evaluation indexes for model performance [36]. Here, AP (Average Precision) represents the area under the curve surrounded by precision and recall, and mAP is the mean of AP for each target category. The specific formula is expressed as follows:

$$P = \frac{TP}{TP + FP} \tag{16}$$

$$R = \frac{TP}{TP + FN} \tag{17}$$

$$mAP = \int_0^1 P * (R) dR \tag{18}$$

TP (True Positives) represents the number of correctly identified samples; FP (False Positives) represents the number of incorrectly identified samples; and FN (False Negatives) represents the number of missed samples. The evaluation-index score ranges from 0 to 1, with higher values indicating better model performance. For an in-depth discussion of edge-computing performance comparisons across different models, the computational load (measured in GFLOPs) and model size (measured in MB) are introduced as metrics.

2.4. Experimental Environment

The training model in this study operated on Windows 11, with an Intel (R) Core (TM) i9-12900H processor running at 2.50 GHz, 16 GB of memory, and an NVIDIA GeForce RTX 3060 graphics card. PyTorch served as the deep learning framework within a virtual environment of CUDA 11.1 version, while Python 3.8 was the programming language employed.

3. Results

3.1. Comparison of Models before and after Improvement

The AP value represents the area enclosed by the PR curve and the coordinate axis. A larger area indicates a higher AP value, indicating greater detection precision of the model for such objects. The mAP is a commonly used evaluation index in the target detection task. It measures the average AP value of the model across multiple categories and provides a comprehensive assessment of positioning precision and prediction precision. As shown in Figure 9 above, the model's performance significantly improved after enhancements.

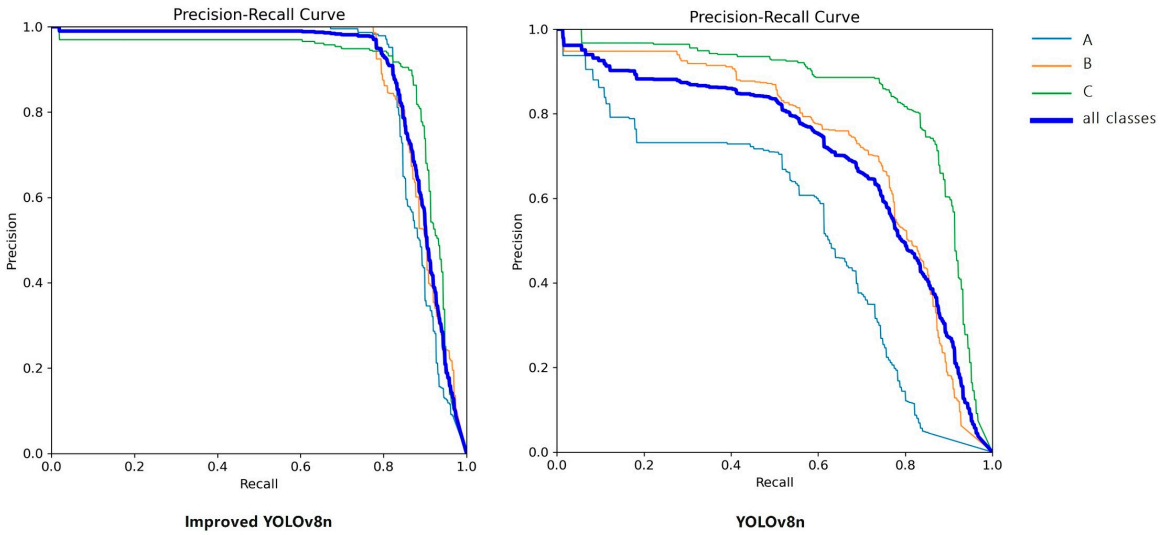

Figure 9. Comparison of P-R curves.

As depicted in Table 1, the model training results revealed that the overall mAP of the improved YOLOv8 model reached 89.1%, a notable increase from 70.4% prior to the enhancements, marking an improvement of 18.7 percentage points. Moreover, the mAP50 values for single bud, one bud with one leaf, and one bud with two leaves witnessed increases of 34.9%, 16.5%, and 4.7%, respectively. The improved model demonstrated significant advantages in average detection precision. Additionally, the new model incorporated a classification-counting function, and its visual detection effect is illustrated in Figure 10.

Table 1. mAP comparison before and after improvement.

Model	Category	mAP
Improved YOLOv8n	single bud	88.5%
	one bud and one leaf	89.5%
	one bud and two leaves	89.1%
YOLOv8n	single bud	53.6%
	one bud and one leaf	73%
	one bud and two leaves	84.4%

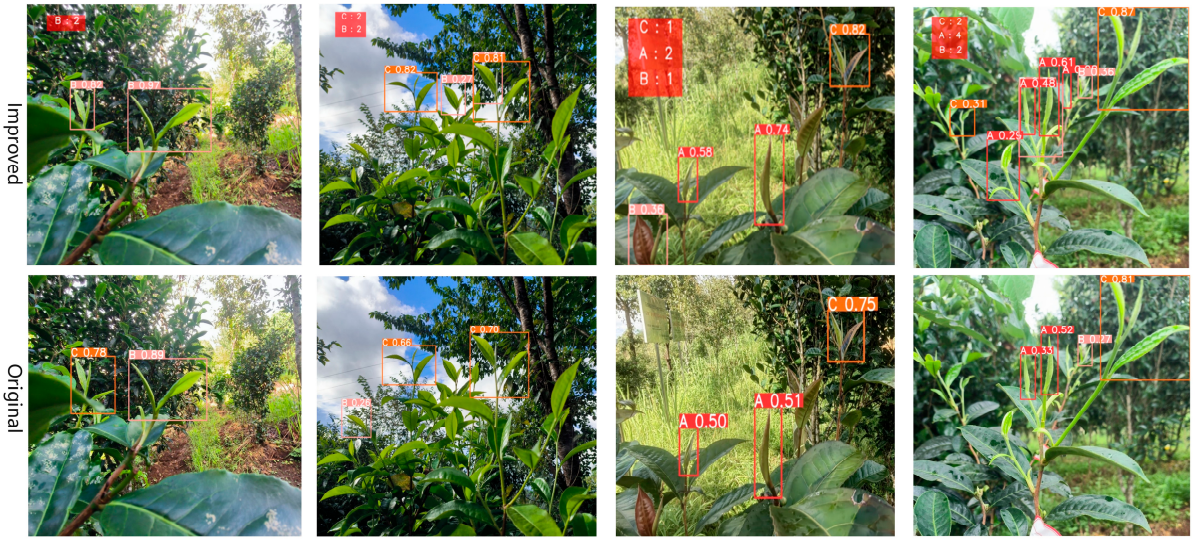

Figure 10. Visual comparison of detection effect.

3.2. Ablation Experiment

In this study, improvements were made to the YOLOv8n model, and the experimental results demonstrate the effectiveness of the enhanced method. As illustrated in Table 2, the introduction of the SPD-Conv module into the main feature extraction process resulted in a double-downsampling feature map with enriched feature information by expanding the number of channels, thereby enhancing the receptive field. This algorithm significantly boosted the detection precision of the model. Upon integrating the SViT into the backbone structure, the recall and mAP50 metrics showed varying degrees of improvement, signifying the method's effectiveness. MPDIoU was employed as the bounding-box regression function, allowing the model to converge to higher detection precision without escalating model complexity, thereby meeting real-time and accuracy requirements. Following the model enhancement, the P, R, and mAP50 saw increments of 17.6%, 19.3%, and 18.7%, respectively. While there was a slight increase in the number of parameters and a 1.2 MB augmentation in model size, the GFLOPs rose by 3.6. Nonetheless, the model retained its lightweight feature. These results highlight the model's capability to improve detection accuracy accurately while maintaining low computational costs. Ablation experiments further underscored the model's advantages and practical value.

Table 2. Comparison of ablation experimental results.

Algorithm	SPD-Conv	SViT	MPDIoU	P (%)	R (%)	mAP50 (%)	Parameters	GFLOPs	Model Size
1	×	×	×	69.3%	66.2%	70.4%	3,011,433	8.2	6.3
2	√	×	×	72.8%	75.5%	77.3%	3,418,320	12.4	6.8
3	√	√	×	86.1%	85.2%	88.6%	3,617,551	12.0	7.5
4	√	√	√	86.9%	85.5%	89.1%	3,617,746	11.8	7.5

Note: √ using this algorithm; × not using this algorithm.

3.3. Comparison of Different Models

To further validate the model's performance and provide a more intuitive comparison of its advantages, a comparative experiment was conducted using the same dataset and training parameters. The comparison results of the loss values for the training and validation sets of the four models—improved YOLOv8n, YOLOv8n, Faster R-CNN, and SSD—are depicted in Figure 11, above. The x-axis in the figure represents the number of training epochs, with 400 epochs set for this study. The loss function indicates the degree of disparity between the model's predicted value and the true value. Smaller loss values signify better model robustness and adaptability to data changes. It can be observed from the graph that the loss-rate curves for the training and validation sets of the four models gradually decline as the number of training epochs increases. The steepest decline occurs in the initial 50 epochs. A smoother convergence curve, indicating gradual stabilization of the model, is characterized by smaller gradients. Among the four models, the improved YOLOv8n model exhibits lower loss, stable curve fluctuations, and superior detection performance.

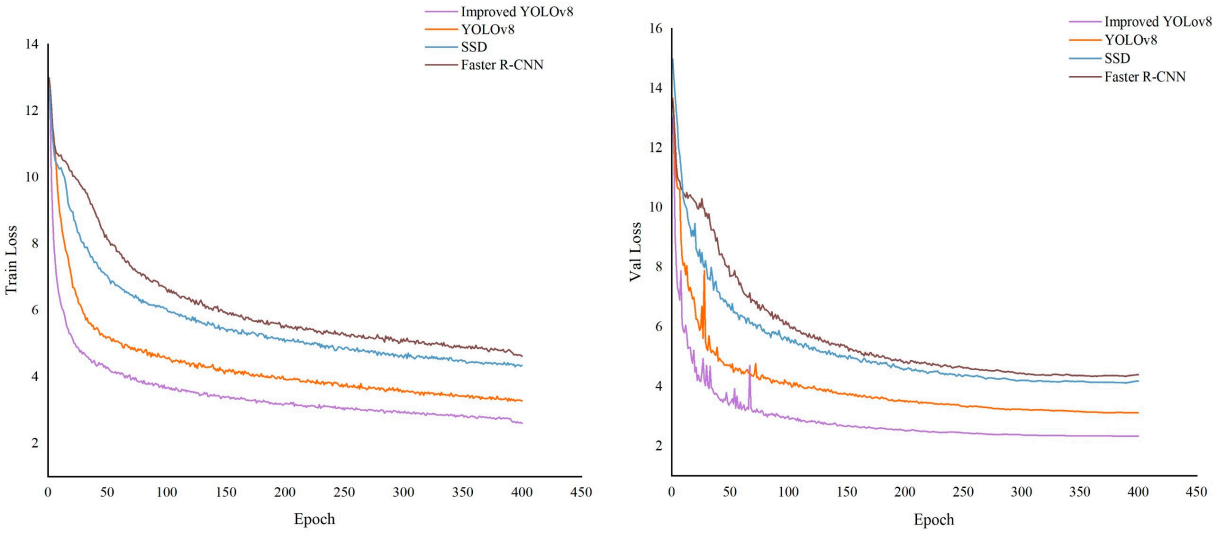

Figure 11. Comparison of loss values before and after improvement of different models.

The comparison of model parameters, computational load, and model size across different algorithms is presented in Table 3. The improved YOLOv8n model demonstrates a notable enhancement in detection precision compared to other models. Although there is a slight reduction in detection speed, the overall performance of the model improves, resulting in better detection capabilities. The model parameters are 7.5 M, and the computational load is 11.8. Despite having small parameters, the model exhibits superior detection performance for complex targets while maintaining fast detection speeds. The precision, recall, and mAP50 of the model for tea grading reach 86.9%, 85.5%, and 89.1%, respectively. Specifically, the mAP50 for the three categories—single bud, one bud with one leaf, and one bud with two leaves—are 88.5%, 89.5%, and 89.1%, respectively, indicating highly accurate results. In comparison with the original YOLOv8n model, there are notable improvements in precision, recall, and mAP50, with increases of 17.6%, 19.3%, and 18.7%, respectively.

Table 3. Performance comparison of different models.

Model	P (%)	R (%)	mAP50 (%)	Model Size (M)	GFLOPs	Parameters
improved YOLOv8n	86.9%	85.5%	89.1%	7.5	11.8	3,617,746
YOLOv8	69.3%	66.2%	70.4%	6.3	8.2	3,011,433
SSD	68.3%	73.3%	75.1%	515.2	62.75	26,285,486
Faster R-CNN	77.1%	79.3%	81.7%	2097.87	370.21	137,098,724

Compared with the one-stage target-detection algorithm SSD, the parameter quantity was only 13.76%, and the complexity was only 18.8%. The precision and recall mAP were improved by 18.6%, 12.2%, and 14%. When contrasted with the classical two-stage target-detection algorithm Faster R-CNN model, the improved detection algorithm exhibited greater accuracy in target detection and positioning, significantly reducing computational costs.

In summary, it achieved a commendable balance between detection speed and precision. It enhanced global information acquisition capability, heightened sensitivity to small targets, and rendered the model more adaptable and robust. The three indexes of P, R and mAP50 increased progressively with training times, eventually stabilizing, indicating the training data's stability and normalcy. The model's addition of the automatic counting function for tea classification not only improved detection accuracy but also addressed issues such as missed detection, false detection, and small targets. Furthermore, it provided a novel approach for estimating tea yield. In this dataset, samples featuring easily detectable targets comprised only a small proportion, with the majority exhibiting complex backgrounds. This setup better verifies the model's efficacy and offers valuable insights for the practical implementation of tea grading recognition.

4. Discussion

The purpose of this study is to enhance the grade-recognition performance of the YOLOv8n model for tea in a natural tea-garden environment. To achieve this, a new model for the grading and counting recognition of tea is proposed. Through the introduction of an SPD-Conv module, the SViT attention mechanism, an improved loss function MPDIoU, and a classification counting function, the model's robustness was markedly enhanced. The total mAP reached 89.1%, compared with the original model growth of 18.7%. Experimental results demonstrated significant improvements in recognition precision for single buds, one bud with one leaf, and one bud with two leaves.

The aforementioned enhancements not only greatly improve the accuracy and convergence of model loss values but also integrate a grading and counting function. Most existing studies on tea-leaf grading recognition models are primarily based on YOLOv5 and YOLOv7, with relatively uniform data backgrounds and easy identifiability. This study's model, based on the more stable and lightweight YOLOv8n architecture and featuring counting functionality, provides a fresh foundation for extended application in tea-picking robots and yield estimation.

While this study yielded positive outcomes, several limitations and avenues for future enhancement remain. The dataset used here originated from tea gardens in specific regions, potentially not fully encapsulating the diverse characteristics of tea across different regions and planting conditions. Hence, future research might expand data collection to include tea samples from various regions and climates, thereby bolstering the model's generalization capabilities. Additionally, despite the improved model's promising performance in experiments, its lightweight nature, considering computing resource constraints, presents a crucial research direction. By further optimizing its structure and employing knowledge distillation techniques, the model's parameter count and computations can be reduced, rendering it more suitable for deployment on resource-limited edge devices. Lastly, future investigations could explore deploying and applying the model in field settings, along with

enhancing its stability and reliability through real-time feedback and adaptive learning mechanisms. These endeavors aim to provide more robust and practical technical support for the intelligence and automation of the tea industry.

5. Conclusions

This study aims to address issues such as dense distribution, limited feature extraction, missed detection, and false alarms in tea-leaf grading recognition, with the goal of enhancing accuracy in tea gardens. Initially, we integrated the SPD-Conv module into the backbone network of the model, enhancing its feature extraction capability and subsequently improving accuracy. Secondly, the integration of SViT aims to enhance the model's focus on tea leaf targets, reduce background interference, and enhance detection performance. Additionally, optimizing the loss function with MPDIoU accelerates model convergence and enhances overall performance. Finally, we introduced and implemented a tea leaf-classification and counting-recognition function.

Experimental results demonstrate that, despite slight increases in parameters and volume compared to the original model, the model still maintains its lightweight nature. Precision, recall, and average precision reached 86.95%, 85.5%, and 89.1%, respectively, marking improvements of 17.6%, 19.3%, and 18.7% compared to the original model. Particularly, the mAP values for single buds, one bud with one leaf, and one bud with two leaves significantly increased, to 88.5%, 89.5%, and 89.1%, respectively, representing growth of 34.9%, 16.5%, and 4.7%. The performance of the proposed model has been significantly enhanced. Compared with the one-stage object-detection algorithm SSD, the parameter count is only 13.76%, and complexity is only 18.8%, but mAP improvement is 14%. Compared with the classic two-stage object-detection algorithm Faster R-CNN, the improved detection algorithm exhibits higher precision in object detection and localization, with an mAP increase of 7.4% and significantly reduced computational costs.

In summary, the improved model demonstrates significant advantages compared to other object-detection models, holding crucial application value in future tea-garden management and production. Its lightweight yet accurate design meets the demands of real-time precision, making it highly suitable for deployment on small-scale devices. This reaffirms the significant advantages of the improved YOLOv8n model in enhancing tea-leaf grading accuracy and implementing counting functionality.

Author Contributions: Conception of the overall framework of the paper and writing—original draft preparation, Y.X.; establishment and optimization of overall model and data curation, Z.W.; data acquisition and processing and data curation, Z.C.; model selection and optimization theory and investigation Y.C.; data set construction and methodology, L.L. and L.C.; model optimization and validation, S.Z.; software and validation, C.W.; visualization and software, H.L.; writing—review and editing, B.W.; funding acquisition, B.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by study on the Yunnan Menghai County Smart Tea Industry Science Technology Mission (202304BI090013), study on the Screening Mechanism of Phenotypic Plasticity Characteristics of Yunnan Big-leaf Tea Plant Driven by AI Based on Data Fusion (202201AS070325), Yunnan Tea Industry Artificial Intelligence and Big Data Application Innovation Team (202405AS350025), Special scientific and technological mission to modern border well-off villages in Xuelin Wa Township and Nuofu Township, Lancang County, Yunnan Province (202204BI090079), and the Development and demonstration of intelligent agricultural data sensing technology and equipment in plateau mountainous areas (202302AE09002001).

Data Availability Statement: The datasets analyzed during the current study are available from the corresponding author on reasonable request.

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

References

1. Subeesh, A.; Mehta, C.R. Automation and digitization of agriculture using artificial intelligence and internet of things. *Artif. Intell. Agric.* **2021**, *5*, 278–291. [CrossRef]
2. Shi, L.; Shi, G.; Qiu, H. General review of intelligent agriculture development in China. *China Agric. Econ. Rev.* **2019**, *11*, 39–51. [CrossRef]
3. Benn, J.A. *Tea in China: A Religious and Cultural History*; University of Hawaii Press: Honolulu, Hawaii USA, 2015; ISBN 0824853989.
4. Zhi, X.; Xianjin, H.; Zheng, Z.; Yang, H. Spatio-temporal variation and the driving forces of tea production in China over the last 30 years. *J. Geogr. Sci.* **2018**, *28*, 275–290.
5. Meng, J.; Wang, Y.; Zhang, J.; Tong, S.; Chen, C.; Zhang, C.; An, Y.; Kang, F. Tea Bud and Picking Point Detection Based on Deep Learning. *Forests* **2023**, *14*, 1188. [CrossRef]
6. Wijeratne, M.A. Pros and cons of mechinical harvesting: A review of experience on tea harvesters tested by the Tea Research Institute of Sri Lanka. *Tea Bull.* **2012**, *21*, 1–9.
7. Chunlin, C.; Jinzhu, L.; Mingchuan, Z.; Yi, J.; Liao, M.; Gao, Z. A YOLOv3-based computer vision system for identification of tea buds and the picking point. *Comput. Electron. Agric.* **2022**, *198*, 107116.
8. Saputro, A.K.; Wibisono, K.A.; Pratiwi, F.P. Identification of Disease Types on Tea–Plant Varieties Based Image Processing with K-Nearest Neighbor Method. *J. Phys. Conf. Ser.* **2020**, *1569*, 32078. [CrossRef]
9. Deng, J.; Jun, D.; Xiaojing, X.; Li, Z.; Yao, H.; Wang, Z. A review of research on object detection based on deep learning. *J. Phys. Conf. Ser.* **2020**, *1684*, 12028. [CrossRef]
10. Camacho, J.C.; Morocho-Cayamcela, M.E. In Mask R-CNN and YOLOv8 Comparison to Perform Tomato Maturity Recognition Task. In *Conference on Information and Communication Technologies of Ecuador*; Springer: Cham, Switzerland, 2023; pp. 382–396.
11. Shuang, X.; Hongwei, S. Tea-YOLOv8s: A Tea Bud Detection Model Based on Deep Learning and Computer Vision. *Sensors* **2023**, *23*, 6576.
12. Trinh, D.C.; Mac, A.T.; Dang, K.G.; Nguyen, H.T.; Nguyen, H.T.; Bui, T.D. Alpha-EIOU-YOLOv8: An Improved Algorithm for Rice Leaf Disease Detection. *AgriEngineering* **2024**, *6*, 302–317. [CrossRef]
13. Li, Y.; Fan, Q.; Huang, H.; Han, Z.; Gu, Q. A Modified YOLOv8 Detection Network for UAV Aerial Image Recognition. *Drones* **2023**, *7*, 304. [CrossRef]
14. Jiwei, L.; Yanchao, L.; Tao, N.; Song, J.; Duan, X. Football player identification based on YOLOv5 backbone and SPD-Conv. In Proceedings of the Eighth International Conference on Electronic Technology and Information Science (ICETIS 2023), Dalian, China, 24–26 March 2023.
15. Wei, W.; Ran, J.; Ning, C.; Li, Q.; Yuan, F.; Xiao, Z. Semi-supervised vision transformer with adaptive token sampling for breast cancer classification. *Front. Pharmacol.* **2022**, *13*, 929755.
16. Siliang, M.; Yong, X. MPDIoU: A Loss for Efficient and Accurate Bounding Box Regression. *arXiv* **2023**, arXiv:2307.07662.
17. Lou, H.; Duan, X.; Guo, J.; Liu, H.; Gu, J.; Bi, L.; Chen, H. DC-YOLOv8: Small-Size Object Detection Algorithm Based on Camera Sensor. *Electronics* **2023**, *12*, 2323. [CrossRef]
18. Terven, J.; Cordova-Esparza, D. A Comprehensive Review of YOLO: From YOLOv1 to YOLOv8 and Beyond. *arXiv* **2023**, arXiv:2304.00501.
19. Slimani, H.; Mhamdi, J.E.; Jilbab, A. Artificial Intelligence-based Detection of Fava Bean Rust Disease in Agricultural Settings: An Innovative Approach. *Int. J. Adv. Comput. Sci. Appl.* **2023**, *14*, 119–128. [CrossRef]
20. Hussain, M. YOLO-v1 to YOLO-v8, the Rise of YOLO and Its Complementary Nature toward Digital Manufacturing and Industrial Defect Detection. *Machines* **2023**, *11*, 677. [CrossRef]
21. Casas, G.G.; Ismail, Z.H.; Limeira, M.M.C.; da Silva, A.A.L.; Leite, H.G. Automatic Detection and Counting of Stacked Eucalypt Timber Using the YOLOv8 Model. *Forests* **2023**, *14*, 2369. [CrossRef]
22. Terven, J.; Córdova-Esparza, D.M.; Romero-González, J.A. A Comprehensive Review of YOLO Architectures in Computer Vision: From YOLOv1 to YOLOv8 and YOLO-NAS. *Mach. Learn. Knowl. Extr.* **2023**, *5*, 1680–1716. [CrossRef]
23. Zhenyu, L.; Haoyuan, L. YOLO_Bolt: A lightweight network model for bolt detection. *Sci. Rep.* **2024**, *14*, 656.
24. Shen, L.; Lang, B.; Song, Z. DS-YOLOv8-Based Object Detection Method for Remote Sensing Images. *IEEE Access* **2023**, *11*, 125122–125137. [CrossRef]
25. Sunkara, R.; Luo, T. No More Strided Convolutions or Pooling: A New CNN Building Block for Low-Resolution Images and Small Objects. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*; Springer: Cham, Switzerland, 2022.
26. Hsu, P.H.; Lee, P.J.; Bui, T.A.; Chou, Y.S. YOLO-SPD: Tiny objects localization on remote sensing based on You Only Look Once and Space-to-Depth Convolution. In Proceedings of the 2024 IEEE International Conference on Consumer Electronics (ICCE), Las Vegas, NV, USA, 6–8 January 2024.
27. Yu, Z.; Lei, Y.; Shen, F.; Zhou, S.; Yuan, Y. Research on Identification and Detection of Transmission Line Insulator Defects Based on a Lightweight YOLOv5 Network. *Remote Sens.* **2023**, *15*, 4552. [CrossRef]
28. Wang, G.; Zhao, Y.; Tang, C.; Luo, C.; Zeng, W. When Shift Operation Meets Vision Transformer: An Extremely Simple Alternative to Attention Mechanism. *Proc. AAAI Conf. Artif. Intell.* **2002**, *36*, 2423–2430. [CrossRef]
29. Li, Y.; Li, S.; Du, H.; Chen, L.; Zhang, D.; Li, Y. YOLO-ACN: Focusing on Small Target and Occluded Object Detection. *IEEE Access* **2020**, *8*, 227288–227303. [CrossRef]

30. Gulati, A.; Qin, J.; Chiu, C.C.; Parmar, N.; Zhang, Y.; Yu, J.; Han, W.; Wang, S.; Zhang, Z.; Wu, Y.; et al. Conformer: Convolution-augmented Transformer for Speech Recognition. *arXiv* **2020**, arXiv:2005.08100.
31. Ren, J.; Yang, J.; Zhang, W.; Cai, K. RBS-YOLO: A vehicle detection algorithm based on multi-scale feature extraction. *Signal Image Video Process.* **2024**, *18*, 3421–3430. [CrossRef]
32. Li, H.; Wang, C.; Liu, Y. YOLO-FDD: Efficient defect detection network of aircraft skin fastener. *Signal Image Video Process.* **2024**, *18*, 3197–3211. [CrossRef]
33. Du, B.; Zhang, L. Target detection based on a dynamic subspace. *Pattern Recogn.* **2014**, *47*, 344–358. [CrossRef]
34. Jinwei, C.; Jie, Y.; Xiaoning, H.; Xie, B.; Zhang, M. Lightweight model of remote sensing ship classification based on YOLOv7-tiny improvement. *J. Phys. Conf. Ser.* **2023**, *2666*, 012023.
35. Wang, Z.; Luo, X.; Li, F.; Zhu, X. Lightweight Pig Face Detection Method Based on Improved YOLOv8. In Proceedings of the 2023 13th International Conference on Information Science and Technology (ICIST), Cairo, Egypt, 8–14 December 2023.
36. Olorunshola, O.E.; Irhebhude, M.E.; Evwiekpaefe, A.E. A Comparative Study of YOLOv5 and YOLOv7 Object Detection Algorithms. *J. Comput. Social. Inform.* **2023**, *2*, 1–12. [CrossRef]