agronomy

MDPI

Article

YOLOv11-Pose-BEH: An Enhanced Multi-Scale Attention Network for Tea Bud Detection and Two-Dimensional Picking Point Localization

Weihao Liu [1,2,†], Junjie He [2,3,†], Chun Wang [2,4], Miao Zhou [1,2], Chunyan Zhao [1,2], Tianyu Wu [1,2], Zhiyong Cao [2], Xinya Chen [1,2], Man Zou [1,2], Kai Peng [1,2], Shasha Feng [1] and Baijuan Wang [1,2,*]

[1] College of Tea Science, Yunnan Agricultural University, Kunming 650201, China; liuweihao2210@163.com (W.L.); 18724979985@163.com (M.Z.); zcy19940726@126.com (C.Z.); wty15508851042@163.com (T.W.); chenxinya128@163.com (X.C.); 15770383235@163.com (M.Z.); 18287115819@163.com (K.P.); 13688749927@163.com (S.F.)
[2] Yunnan Organic Tea Industry Intelligent Engineering Research Center, Kunming 650201, China; 18087827443@163.com (J.H.); chunwangkk@163.com (C.W.); czy@ynau.edu.cn (Z.C.)
[3] College of Agronomy and Biotechnology, Yunnan Agricultural University, Kunming 650201, China
[4] College of Mechanical and Electrical Engineering, Yunnan Agricultural University, Kunming 650201, China
* Correspondence: wangbaijuan2023@163.com
† These authors contributed equally to this work.

Abstract

The terrain of tea gardens in Yunnan is complex, and the branches and leaves of tea trees are densely interlaced. Traditional manual picking methods are labor-intensive and inefficient, and can no longer meet the needs of modern tea industry development. To achieve automatic recognition of tea buds and leaves and accurate two-dimensional localization of picking points in complex natural environments, this study constructs a lightweight and high-precision tea bud and leaf detection and keypoint localization model, YOLOv11-pose-BEH, based on the YOLOv11-pose model. Based on the original network, this model introduces the BiFPN feature fusion module to achieve efficient bidirectional transmission of multi-scale information. The EMA (Efficient Multi-scale Attention) mechanism is integrated to form the C2PSA_EMA module, improving the effect of multi-scale feature extraction. HetConv is introduced to form the C3K2_HetConv module, enhancing the extraction of local texture and edge features. Compared with the baseline network, the improved YOLOv11-pose-BEH achieved significant improvements in both tea bud object detection and picking point localization. For the tea bud object detection task, the precision, recall, mAP0.5, and F1-score increased by 7.41, 6.39, 5.28, and 6.88 percentage points, respectively. For the picking point localization task, the precision, recall, mAP0.5, and F1-score increased by 5.86, 5.83, 4.83, and 5.85 percentage points, respectively. These results indicate that the proposed model can achieve more accurate and stable tea bud and leaf detection and two-dimensional picking point localization under complex backgrounds, providing efficient and reliable technical support for the visual perception of intelligent tea-picking robots in tea gardens.

Keywords: tea bud and leaf detection; 2D picking point localization; YOLOv11-pose-BEH; BiFPN; EMA mechanism; HetConv convolution

Academic Editor: Baohua Zhang and Yongliang Qiao

Received: 12 May 2026
Revised: 28 June 2026
Accepted: 30 June 2026
Published: 2 July 2026

1. Introduction

Tea is one of the three major non-alcoholic beverages in the world and has a wide consumer base globally [1,2]. As one of China's main tea-producing regions, Yunnan

possesses unique plateau climates and diverse ecological environments, which endow its tea trees with a long growth cycle and excellent bud and leaf quality [3]. However, the distinctive mountainous terrain of the plateau also restricts the automation and intelligence development of tea picking. At present, tea harvesting is still mainly performed manually. This traditional picking method is inefficient, requires intensive labor input, and faces increasingly severe labor shortages, which have become major constraints on the sustainable development of the tea industry in Yunnan [4]. Therefore, how to utilize computer vision technology to achieve accurate and rapid recognition of tea buds and leaves, as well as precise localization of picking points, has become a key issue in realizing automated and intelligent tea picking [5].

In automated tea-picking systems, visual perception constitutes one of the core components. The detection of tea buds and the precise identification and localisation of picking points directly influence the operational effectiveness of harvesting machinery [6]. Compared with conventional crop object detection tasks, tea bud and leaf recognition and localization in the natural environment of Yunnan tea gardens face more complex challenges. First, tea buds and leaves are small in size and similar in morphology, making them easily affected by branch and leaf overlap and background interference. Second, the illumination in tea gardens changes significantly with time, slope direction, and canopy occlusion, resulting in unstable texture and color features of the targets. Third, intelligent picking requires not only the bounding box locations of tea buds and leaves, but also the further determination of two-dimensional picking point positions that meet the picking standards. Therefore, tea bud and leaf recognition and localization in the natural environment of Yunnan tea gardens exhibit typical characteristics of small-object detection, while also imposing higher requirements for the precise localization of picking points in complex scenarios [7,8]. The task of detecting small objects places exceptionally high demands on the precision and robustness of detection algorithms [9]. This challenge is further amplified in complex natural settings.

To address this issue, early methods mainly relied on manually designed color, shape, and texture features. Zhang et al. achieved tea bud segmentation using G–B component enhancement and the watershed algorithm, obtaining a segmentation accuracy of 94.26% on 100 samples. However, this method was highly sensitive to illumination changes and was difficult to apply in natural tea garden environments. With the widespread application of the YOLO series models in agricultural vision tasks, researchers began exploring their feasibility in tea bud detection [10]. Yang et al. constructed the TBD dataset, including four tea varieties, and systematically compared seven mainstream algorithms such as YOLOv4, YOLOv5, and YOLOX. The results showed that YOLOX achieved a mAP of 95.47%, but the accuracy dropped by about 14% in the classification of Anji Baicha and Nong Kangzao due to their similar color [11]. Zhang et al. proposed TS-YOLO, replacing the YOLOv4 backbone with MobileNetV3 and introducing deformable convolution and coordinate attention mechanisms. While maintaining an AP of 82.12%, the model size was reduced to 11.78 MB, an 81.7% decrease compared with YOLOv4. However, its recall rate was only 78.42% under bud-leaf overlap and occlusion scenarios [12]. In addition, Zhang et al. removed the Focus layer and pruned the Neck layer in ShuffleNetv2-YOLOv5-Lite-E, reducing the model to 27% of YOLOv5's size. It achieved 8.6 fps real-time detection on Raspberry Pi and 97.43% mAP on PC, but was limited to single "one-bud-two-leaves" morphology, restricting its applicability in multi-grade tea classification [13]. Li et al. targeting the severe overlap of "one-bud-one-leaf/two-leaves" in large-leaf tea varieties, proposed Tea-YOLO, integrating the GhostNet lightweight backbone, CBAM attention, and SIoU loss function. Under complex lighting and weather conditions, it achieved a mAP of 85.15%, an improvement of 1.08% over YOLOv4, while reducing parameters by 82.36%,

achieving a good balance between robustness and lightweight performance [14]. Li et al. proposed the TeaBudNet model to address the challenges of small tea bud targets, complex backgrounds, and limited computational resources of edge devices in outdoor tea garden environments. By introducing Weight-FPN, a P2 detection layer, and a Group–Taylor pruning strategy, their method improved tea bud detection accuracy while also achieving a better balance between model lightweighting and practical deployment capability [15].

This line of research mainly focuses on the task of small tea bud detection itself, with particular emphasis on improving detection performance and deployment efficiency, while paying relatively limited attention to the precise localisation of picking points required for intelligent harvesting operations. Therefore, relying solely on tea bud detection is still insufficient to meet the practical demands of integrated visual perception in intelligent tea-picking systems, where both target recognition and picking-point determination are required.

Tea shoot detection is only the first step toward automated tea harvesting; to truly replace manual picking with machines, it is also essential to accurately localise the picking points on the tea shoots. Wang et al. refined the stem–leaf mask based on Mask R-CNN, using the centroid of the stem as the picking point, achieving 93.95% positioning accuracy on 100 complex-scene images. However, this method required extensive semantic annotations and was limited by its inability to simultaneously locate multi-level picking points and its sensitivity to occlusion [16]. To improve robustness, Pan et al. broke away from the CNN paradigm and proposed the SORC-SFT framework, using Swin-Transformer as the backbone. By combining rotated box detection with semantic segmentation, they transformed the picking point localization into a mask centroid calculation problem. The method achieved an average success rate of 92.97% across four viewpoint categories, realizing the first "view–variety–pose" three-dimensional generalization [17]. However, this approach followed a two-stage pipeline, resulting in inference latency about three times higher than traditional methods, which limited its deployment on edge devices. Shuai et al. introduced a six-point keypoint regression strategy based on YOLOv5, proposing the YOLO-Tea model. The picking point was defined as the midpoint between the bud base and the stem tip, achieving a mAP of 76.53%. Although this model improved the mAP by 5.26% over the baseline and supported dual-level picking point localization ("single bud"/"one bud and one leaf"), its localization precision still did not meet the requirements of actual picking operations [18].

In summary, although existing studies have made some progress in tea shoot detection and picking-point localisation, significant challenges still remain in complex tea plantation environments. Traditional detection frameworks often lack robustness to illumination changes, leaf overlap, and background interference in natural scenes, while two-stage detection and segmentation methods usually require relatively high computational cost and are therefore difficult to deploy in real-time tea-picking systems. In addition, existing single-stage models still have limited accuracy in the joint task of one-bud-two-leaves detection and picking-point localisation. Based on this, this study explores a one-stage-model-based method for detecting tea buds and leaves and two-dimensional picking points in complex environments. The main contributions of this paper are as follows:

(1) To address the problems of obvious scale variation in tea bud targets and insufficient feature representation under different growth morphologies, the BiFPN structure is introduced into the feature fusion stage of YOLOv11-pose to enhance information interaction among multi-scale features, thereby improving the model's representation ability for tea bud targets of different scales and their local structural features.
(2) To reduce the adverse effects of branch and leaf overlap and complex background interference on target recognition and keypoint localization in natural tea gardens,

the EMA mechanism is integrated into the feature extraction process of the model to strengthen the model's attention to key tea bud regions and spatial location information, improving the specificity and stability of feature extraction under complex background conditions.

(3) The HetConv heterogeneous convolution structure is introduced to enhance feature representation ability while controlling the number of model parameters and computational complexity, enabling the model to better balance recognition accuracy and computational efficiency.

2. Materials and Methods

2.1. Dataset Construction

2.1.1. Image Acquisition

In this study, tea buds of Camellia sinensis (Yunnan large-leaf tea) were selected as the research objects. The image data of tea buds and leaves were collected at the Yuecheng Base, Menghai County, Xishuangbanna Prefecture, Yunnan Province, China, from April to September 2025. To ensure the representativeness of the experimental images, a Canon EOS R5 camera (Canon Inc., Tokyo, Japan) was used as the data acquisition device. Image acquisition was conducted between 9:00 and 17:00 across different field plots and individual tea plants, so that the dataset included tea bud and leaf samples under varying illumination intensities, shooting angles, and natural field conditions, thereby improving the diversity of the original image dataset.

A total of 1249 tea bud and leaf images were collected. To ensure dataset quality, 1000 images were finally selected as the original dataset after strict screening. To avoid data leakage caused by augmented versions of the same original images appearing simultaneously in the training set, validation set, or test set, the 1000 original images were first divided into the training set, validation set, and test set at a ratio of 8:1:1 before data augmentation. The original images covered different field plots and individual tea plants, and the test set was composed only of original, non-augmented images that were not used for model training or validation. On this basis, data augmentation operations were performed only on the training set images, including random rotation, exposure adjustment, and brightness adjustment [19], expanding the training set to 3200 images. The validation set and test set remained unchanged and consisted exclusively of original, non-augmented images. Therefore, they were used to objectively evaluate the detection and picking point localization performance of the model on unseen original samples.

2.1.2. Image Annotation

This study used the labelme image annotation tool to annotate the target regions and keypoints of tea buds and leaves. Figure 1 shows an example of data annotation. When annotating the target region, this study used the whole one-bud-two-leaf structure as the detection object and annotated its enclosing region with a rectangular bounding box. Considering that different picking standards, such as single bud, one bud with one leaf, and one bud with two leaves, exist in the actual tea picking process, only annotating the target detection box is difficult to meet the requirement of picking point localization. Therefore, this study further annotated six keypoints on each tea bud and leaf target to describe the two-dimensional picking positions under different picking grades of tea buds and leaves.

Specifically, points 1, 2, and 3 correspond to the bud tip, the first leaf tip, and the second leaf tip, respectively. They are mainly used to represent the morphological structure of tea buds and leaves, the leaf unfolding direction, and the local spatial relationship, and serve as auxiliary structural keypoints in model training. Points 4, 5, and 6 correspond

to the picking points under the picking standards of single bud, one bud with one leaf, and one bud with two leaves, respectively, and are used to represent the two-dimensional picking positions under different picking requirements. Among them, the picking points are determined according to the actual picking standards of different picking grades of tea buds and leaves, and are specifically annotated near the connection area between the base of the corresponding bud-leaf combination and the tender stem. This position refers to the requirements for bud and leaf integrity, retained tender stem length, and picking grade consistency during manual picking, and is used to represent the two-dimensional suitable cutting positions under the picking standards of single bud, one bud with one leaf, and one bud with two leaves. To ensure the reliability of the annotation results, the keypoint annotation rules were jointly determined by researchers in tea science based on manual picking experience. After annotation, the samples were reviewed, and samples with obvious deviations or disputes in keypoint positions were re-annotated or removed.

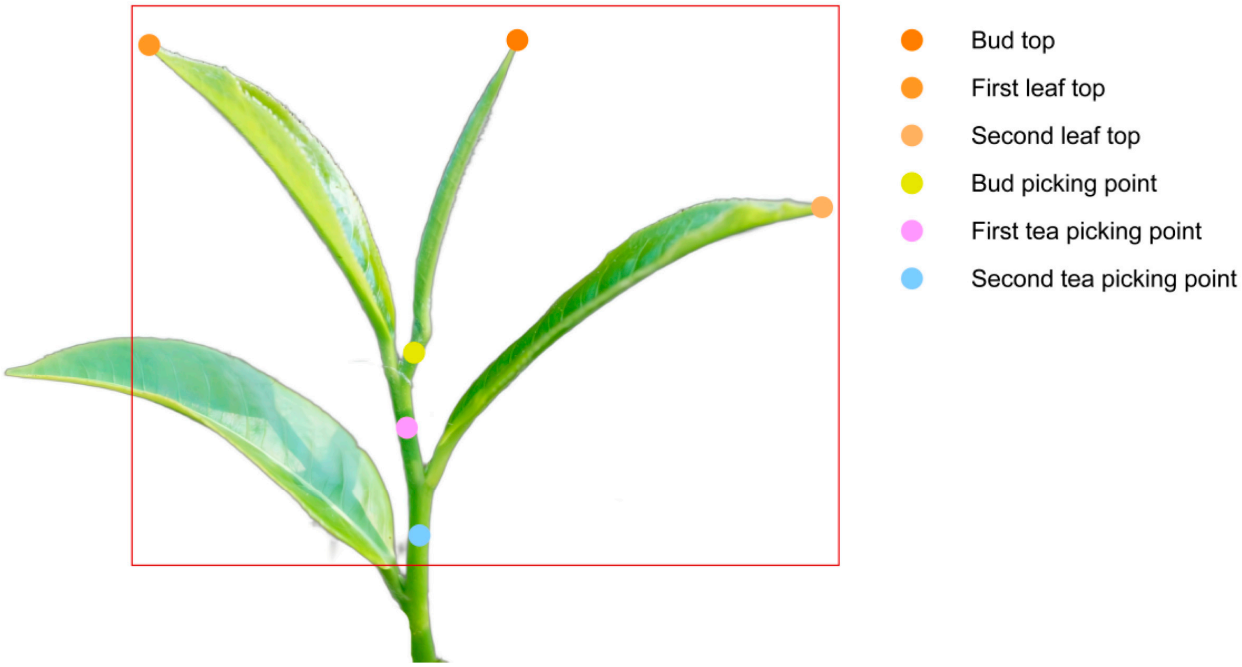

Figure 1. Schematic Diagram of Image Annotation for the Dataset.

During keypoint training, this paper follows the default pose estimation training method of the YOLOv11-pose framework, and jointly optimizes the bounding boxes, classes, and keypoints. Related studies such as YOLO-Pose introduce Object Keypoint Similarity (OKS) into the design of keypoint loss, and measure the keypoint localization error through the spatial distance between the predicted keypoints and the ground-truth keypoints [20]. The keypoint loss can be formulated as shown in Equation (1):

$$L_{kpts} = 1 - \frac{\sum_{n=1}^{N_{kpts}} \exp\left(\frac{-d_n^2}{2s^2k_n^2}\right)\delta(v_n > 0)}{\sum_{n=1}^{N_{kpts}} \delta(v_n > 0)} \quad (1)$$

where d_n denotes the Euclidean distance between the predicted position and the ground-truth position of the n-th keypoint, s denotes the object scale, k_n denotes the keypoint-related weight, and $\delta(v_n > 0)$ indicates whether the keypoint is visible or valid for calculation.

In this study, six keypoints were annotated for each tea bud target. The 1st, 2nd, and 3rd keypoints correspond to the bud tip, the first leaf tip, and the second leaf tip, respectively, and mainly serve as auxiliary structural keypoints. The 4th, 5th, and 6th

keypoints correspond to the two-dimensional picking points under the picking standards of single bud, one bud with one leaf, and one bud with two leaves, respectively. During the training stage, all six keypoints participated in the training of the YOLOv11-pose model. This study did not set additional manual weights for different keypoints, nor did it separately construct an explicit structural constraint loss. The 1st, 2nd, and 3rd auxiliary structural keypoints mainly help the model learn the overall morphological structure of tea buds and leaves, the leaf unfolding direction, and the local spatial relationship through joint regression with the 4th, 5th, and 6th picking points, thereby providing indirect spatial constraints for picking point localization.

2.2. YOLOv11 Algorithm Improvement

2.2.1. YOLOv11 Network

YOLOv11 is an efficient object detection algorithm proposed by the Ultralytics team on 30 September 2024 [21]. Compared with YOLOv8, YOLOv11 replaces the original C2f module with the C3K2 module. The C3K2 module optimizes the computational performance of the Cross-Stage Partial (CSP) mechanism by using two smaller convolutional layers instead of one large convolution, thereby improving the processing speed while maintaining network balance [22]. In the Backbone, a new C2PSA module is introduced while retaining the original SPPF module. This enhances spatial attention within the feature maps, enabling the network to efficiently focus on important regions. In the Head, depthwise separable convolutions are adopted to reduce the number of parameters and computational cost, while extending support for multiple tasks such as object detection, instance segmentation, and pose estimation, thus demonstrating stronger generalization capability. Most importantly, the newly added keypoint detection function in YOLOv11 enables precise prediction of key structural points of targets, providing model support for the two-dimensional localization of tea bud picking points. This feature grants YOLOv11 higher research potential and application value in agricultural vision tasks [23,24]. Overall, YOLOv11 achieves higher mAP0.5 on the COCO dataset while significantly reducing latency, effectively balancing speed and accuracy. It performs excellently in real-time applications and offers a reliable network foundation for subsequent studies on tea bud recognition and picking point localization.

To enhance the performance of YOLOv11 in tea bud small-object detection and picking point localization tasks, this study improves the network architecture from three aspects: (1) The BiFPN module is introduced into the Neck part to achieve efficient multi-scale fusion of shallow spatial details and deep semantic information, optimizing the problem of insufficient small-target feature extraction. (2) In the Backbone part, the Efficient Multi-Scale Attention (EMA) mechanism is integrated into the C2PSA module to construct the C2PSA_EMA structure, which improves the precision of feature selection and the capability of contextual information extraction, thereby enhancing the model's recognition performance of tea buds and picking point regions under complex tea garden backgrounds. (3) The heterogeneous convolution (HetConv) is introduced into the C3K2 module to form the C3K2_HetConv module, which strengthens the feature representation of local textures and edge details while improving the efficiency of convolution computation, thereby helping to improve the detection accuracy of small-scale targets.

The synergistic effect of these three components enables the improved YOLOv11 model to maintain real-time performance while achieving stronger small-object recognition capability and higher keypoint localization precision. This provides effective technical support for intelligent tea-picking applications. The architecture of the improved YOLOv11 network is shown in Figure 2.

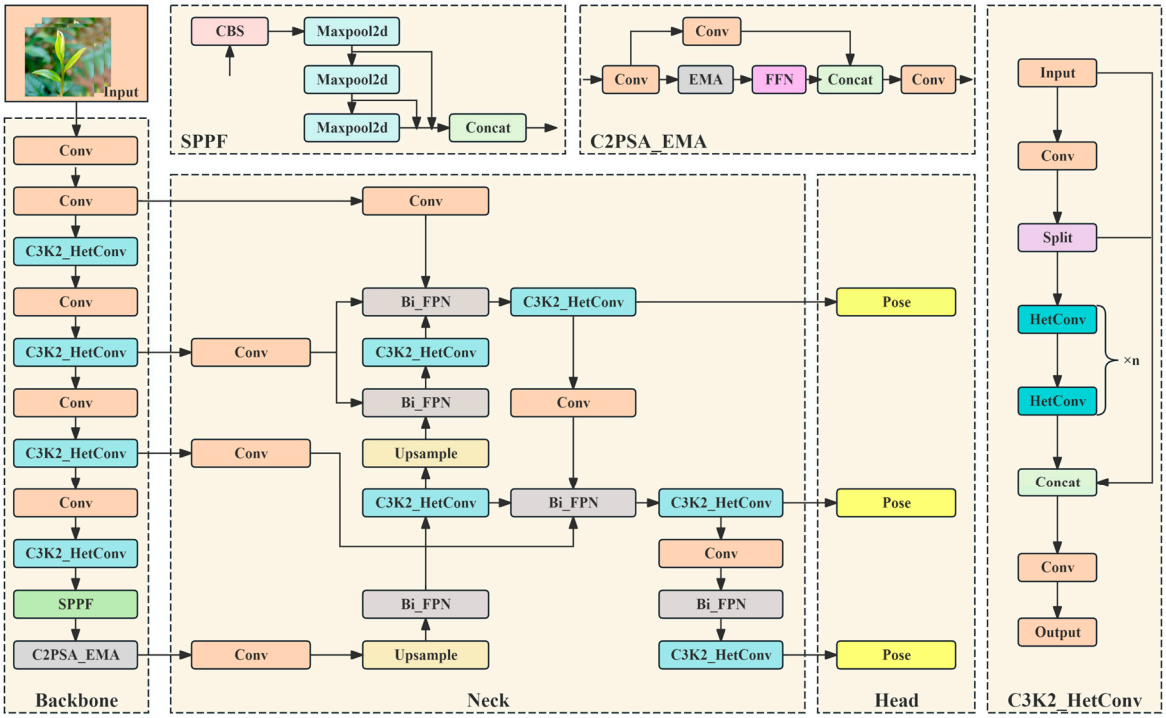

Figure 2. Schematic Diagram of the Improved YOLOv11-pose-BEH Model Network Architecture.

2.2.2. Improvement of Multi-Scale Feature Fusion Based on BiFPN

As a lightweight keypoint detection network, YOLOv11 primarily employs the FPN + PAN structure in its Neck for multi-scale feature fusion. However, this structure has two main limitations: first, the direction of feature information transmission is relatively single, leading to insufficient cross-layer interaction; and second, all input features are equally weighted during fusion, making it impossible to adaptively adjust their contributions according to the importance of features with different resolutions [25]. These limitations are particularly prominent in tea bud and leaf recognition tasks, where they can easily result in unstable detection, keypoint offset, or missed detections.

To address the aforementioned issues, this study introduces the BiFPN module into the Neck part of YOLOv11. The BiFPN module, proposed by Mingxing Tan et al., is an efficient multi-scale feature fusion network that enhances information flow and fusion capability among feature layers [26]. Built upon the traditional Feature Pyramid Network (FPN), BiFPN further optimizes cross-layer connections and feature fusion strategies. Through bidirectional feature fusion, it allows features to flow and merge both from top to bottom and from bottom to top across the network layers, thereby avoiding the information loss caused by a single-path structure and enhancing the model's perception ability for multi-scale objects [27–29].

The network architecture comparison between BiFPN and three other feature pyramid fusion modules is shown in Figure 3.

For feature maps of different resolutions, BiFPN adopts a weighted feature fusion mechanism to optimize the effectiveness of feature integration. In the traditional FPN, feature fusion is performed simply by summing all input features without considering the varying influence of different-resolution features on the output [30]. In contrast, BiFPN assigns distinct learnable weights to features of different resolutions, achieving more flexible and efficient feature fusion. This mechanism not only enhances the model's ability to detect multi-scale objects but also effectively reduces redundant information interference,

thereby further improving the overall network performance [31]. The output of the fused feature map can be expressed as Equation (2):

$$O = \sum_i \frac{\omega_i}{\in + \sum_j \omega_j} \cdot I_i \quad (2)$$

where O denotes the output fused feature map, I_i represents the i-th input feature map, ω denotes the learnable weight parameter, and $\in$ is a stability constant.

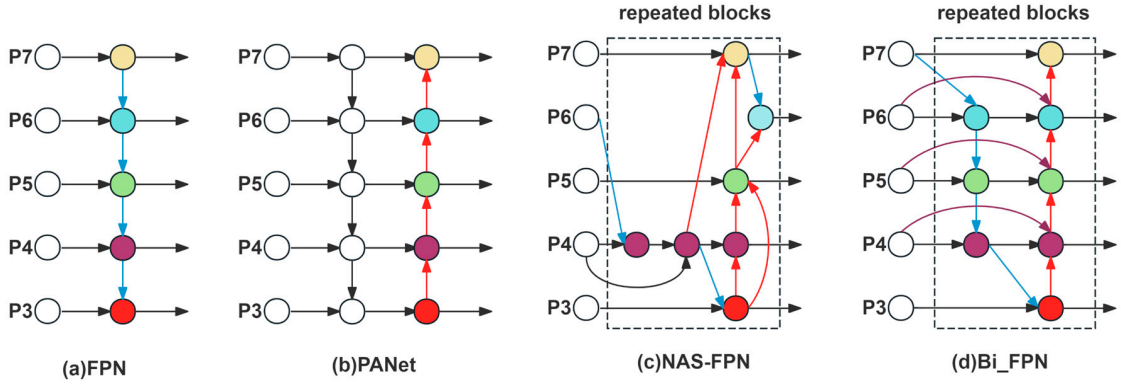

Figure 3. Comparison of BiFPN with three other multi-scale feature fusion architectures: (**a**) Feature Pyramid Network (FPN); (**b**) Path Aggregation Network (PANet); (**c**) Neural Architecture Search Feature Pyramid Network (NAS-FPN); (**d**) Bidirectional Feature Pyramid Network (BiFPN).

2.2.3. C2PSA_EMA Module

In the task of small tea bud and leaf target recognition and picking point detection, the complex illumination conditions and backgrounds of tea plantations often lead to low target contrast, posing significant challenges to the model's feature extraction capability. Although the original C2PSA module provides certain advantages in enhancing local feature representation, its performance still shows limitations under complex conditions. To address this issue, this study introduces the Efficient Multi-Scale Attention (EMA) module and integrates it with the C2PSA module to form the C2PSA_EMA module, thereby improving the model's detection accuracy and stability in this task.

The EMA module is a novel efficient multi-scale attention module proposed by Daliang Ouyang et al. in 2023 [32]. This module can enhance feature representation ability, improve multi-scale feature extraction performance, and thus improve object detection performance and optimize computational efficiency, enabling the model to better adapt to complex scenes and small-object detection tasks while maintaining high efficiency.

As shown in Figure 4, the first part of the EMA module divides the input feature map into G sub-features along the channel dimension, where each sub-feature group contains C/G channels. The grouped feature maps can be represented as $\{X_0, X_1, ..., X_{G-1}\}$, with each group denoted as $X_g \in R^{C/G \times H \times W}$, and each group can independently learn different semantic information. This multi-group feature processing method effectively avoids the information loss caused by channel dimensionality reduction in the original network and forms differentiated feature representations among different groups, which helps the network better capture differences in morphology and texture structures of tea buds and leaves [33]. The second part employs a parallel substructure consisting of two 1×1 branches and one 3×3 branch to process the feature maps output from the first part. The processed feature maps from the 1×1 and 3×3 branches are denoted as $F_{1\times1}$ and $F_{3\times3}$, respectively. Subsequently, a Softmax normalization operation is applied to generate a two-dimensional Gaussian-like spatial weight map. Finally, by performing a matrix dot

product operation, the attention features from the 1 × 1 and 3 × 3 branches are interactively fused to obtain a new cross-spatial attention map.

As shown in the network architecture diagram of the C2PSA_EMA module in Figure 2, the EMA mechanism is integrated into the original C2PSA module. This fusion not only preserves the local feature representation ability of the C2PSA module, but also suppresses irrelevant background noise, outputting more discriminative and robust high-level semantic features, which enables the model to achieve more accurate keypoint localization and small-object detection under complex backgrounds.

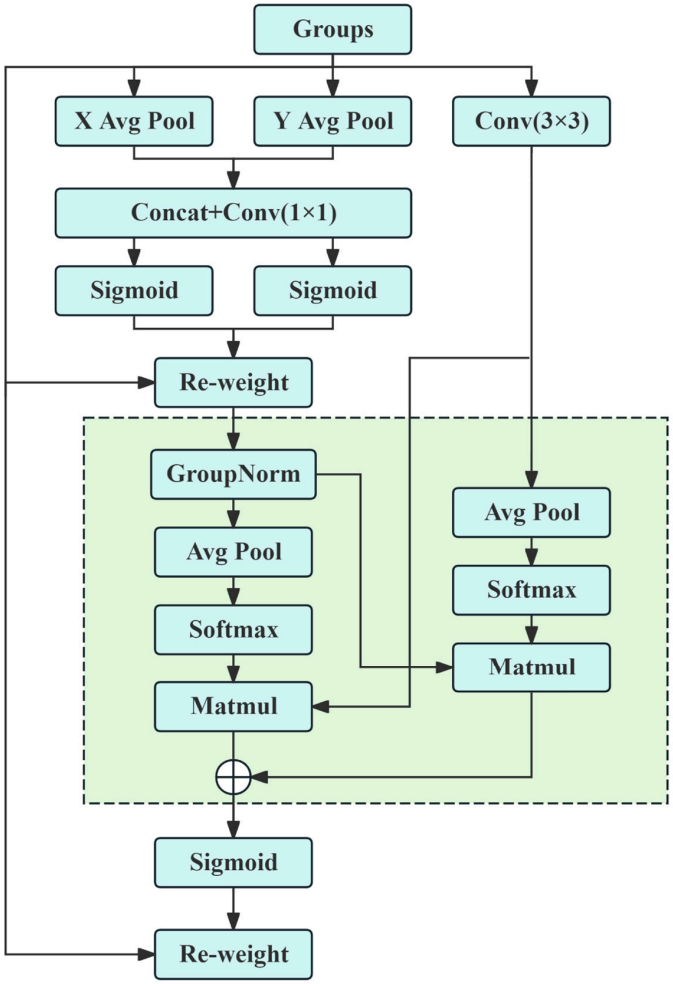

Figure 4. Schematic diagram of the Efficient Multi-Scale Attention (EMA) module.

2.2.4. C3K2_HetConv Module

In the YOLOv11 architecture, the C3K2 module operates by splitting the input features into two branches. One branch preserves the original information through a residual connection, whereas the other performs feature extraction through a bottleneck structure composed of two 3 × 3 convolutions. The outputs of the two branches are then concatenated and fused. However, because the C3K2 module contains a relatively large number of 3 × 3 convolution operations, its parameter count and computational cost remain comparatively high. In small-object detection tasks such as tea bud detection, an excessive use of homogeneous 3 × 3 convolutions not only increases the computational burden of the model, but also makes it more difficult for the network to achieve a better balance between lightweight design and detection accuracy.

To address this issue, HetConv is introduced into the C3K2 module in this study to construct the C3K2_HetConv module. HetConv (Heterogeneous Kernel-Based Convolu-

tion), proposed by Singh et al., is a convolutional operation based on heterogeneous kernels, whose core idea is to mix kernels of different sizes within the same convolutional filter [34]. Specifically, a proportion of the channels retains conventional $K \times K$ kernels for spatial feature extraction, while the remaining channels use 1×1 kernels to reduce the number of parameters and FLOPs. When $K = 3$, some of the channels still employ 3×3 convolutions to preserve spatial correlation modelling, whereas the remaining channels use 1×1 convolutions to perform lightweight feature transformation.

The key advantage of HetConv lies in retaining the convolutional layer's ability to model local spatial information as much as possible while reducing computational cost. If all channels were replaced by 1×1 kernels, the convolutional filter would lose the necessary capacity to model spatial correlations, which would in turn lead to a decline in model accuracy. Therefore, HetConv does not simply "lightweight" standard convolution into a purely 1×1 operation. Instead, it preserves 3×3 kernels on a subset of channels so that the network can still effectively capture local structures and spatial relationships, whilst using 1×1 kernels on the remaining channels to reduce redundant computation. In this way, HetConv achieves a more favourable balance between representational capacity and computational efficiency. The computational complexity of HetConv relative to standard convolution can be formulated as shown in Equation (3):

$$R_{HetConv} = \frac{1}{P} + \frac{1 - 1/P}{K^2} \tag{3}$$

Here, P denotes the heterogeneity ratio control parameter, and K represents the kernel size. When $K = 3$ and $P = 4$, the computational cost of a single layer can be reduced by approximately 66.7%. The comparison between standard convolution filters and heterogeneous convolution filters is shown in Figure 5.

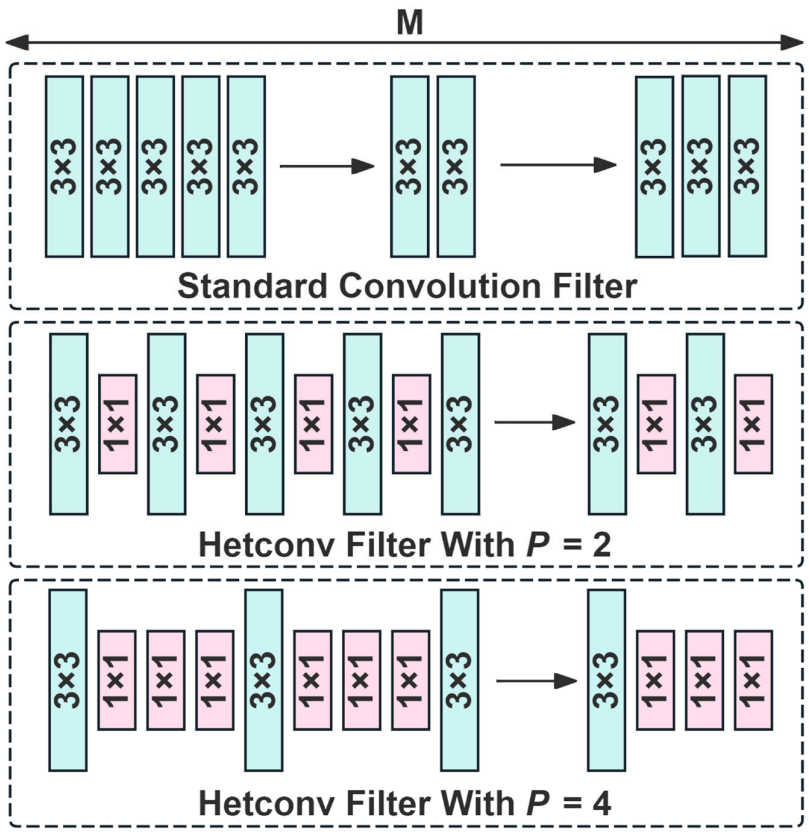

Figure 5. Comparison between standard convolution filters and heterogeneous convolution filters.

Compared with the original C3K2 module, the proposed C3K2_HetConv module preserves the residual connection and cross-stage feature concatenation structure, whilst replacing part of the homogeneous 3×3 convolutions with a heterogeneous kernel design. This enables the network to reduce parameter redundancy and computational cost while maintaining high sensitivity to fine-grained local features of tea buds and picking points, such as texture, contour, and boundary information. As a result, the proposed module achieves an effective balance between the requirements of lightweight deployment and improved detection accuracy.

3. Results and Analysis

3.1. Experimental Environment

The experimental environment for this study is configured as shown in Table 1:

Table 1. Configuration of the experimental environment.

Configuration Item	Configuration Parameter
Operating system	Windows 11
CPU	Intel(R) Core (TM) i7-14650HX
Memory	16 GB
GPU	NVIDIA GeForce RTX 4060
Compiled language	Python 3.8.19
Software framework	PyCharm 2024
CUDA	CUDA Version 11.8

3.2. Evaluation Metrics

To comprehensively evaluate the performance of the proposed model in tea shoot detection and picking-point localisation, Precision, Recall, mAP0.5, and F1-score were selected as the main evaluation metrics [35]. Because the model jointly outputs bounding-box detection results and keypoint prediction results, the metrics were reported separately for two subtasks. Here, B refers to the one-bud-two-leaves detection task, and P refers to the picking-point localisation task. Precision indicates the proportion of correct predictions among all predicted results, Recall indicates the proportion of correctly identified targets among all annotated targets, mAP0.5 reflects the overall performance of the model at a threshold of 0.5, and F1-score is the harmonic mean of Precision and Recall. The specific formulas are shown in Equations (4)–(7), where TP, FP, and FN denote true positives, false positives, and false negatives, respectively. For the B and P tasks, these metrics were computed independently to evaluate the model performance in object detection and keypoint localisation from different perspectives.

$$precision = \frac{TP}{TP + FP} \tag{4}$$

$$recall = \frac{TP}{TP + FN} \tag{5}$$

$$\mathrm{m}AP0.5 = \frac{1}{N}\sum_{\mathrm{k}=1}^{N} AP_{\mathrm{k}}(0.5) \tag{6}$$

$$F1 = \frac{2 \cdot precision \cdot recall}{precision + recall} \tag{7}$$

3.3. Tea Bud and Picking Point Detection Experiments

3.3.1. Parametric Analysis of the C3K2_HetConv Module

To further evaluate the impact of the hyperparameter P within the C3K2_HetConv module on model performance, a series of comparative experiments were conducted. Based on the improved YOLOv11-pose-BEH framework, the value of P was configured to 2, 4, 8, and 16, respectively. The corresponding experimental results are summarized in Table 2.

Table 2. Performance and computational overhead of the C3K2_HetConv module across varying values of *P*.

Model	P(B)/%	R(B)/%	mAP0.5(B)/%	F1(B)/%	P(P)/%	R(P)/%	mAP0.5(P)/%	F1(P)/%	Parameters /M	GFLOPs
P = 2	86.05	79.23	87.88	82.50	84.83	78.11	87.10	81.33	2.73	7.4
P = 4	89.25	84.36	90.08	86.74	88.16	84.36	89.64	86.22	2.66	7.0
P = 8	86.47	82.09	87.80	84.22	86.47	82.09	87.73	84.22	2.63	6.8
P = 16	82.05	79.33	87.38	80.67	81.63	71.99	83.77	76.51	2.62	6.7

Note: P, R, and F1 denote Precision, Recall, and F1-score, respectively. B denotes the object detection task, and P denotes the picking point localization task.

The experimental results indicate that, as the parameter P gradually increases, the overall detection performance of the model first improves and then declines, while the degree of model lightweighting progressively increases.

When P = 2, a relatively large proportion of channels employ 3 × 3 convolutions for spatial feature extraction. Although this design can enhance the local spatial receptive field, an excessive number of 3 × 3 convolutions introduces considerable computational redundancy and may lead to an overemphasis on local texture responses, thereby increasing the influence of background noise. In the tea bud detection task, the plantation environment typically contains complex backgrounds, such as overlapping branches and leaves as well as varying illumination conditions. Under such circumstances, overly strong local spatial feature extraction may instead weaken the model's ability to discriminate the target regions. Moreover, within the cross-stage feature fusion structure of C3K2, an excessively high proportion of 3 × 3 convolutions may reduce the effectiveness of 1 × 1 convolutions in channel-wise information fusion, thereby disturbing the balance of feature representation and ultimately resulting in a decline in overall model performance.

When P = 4, the model achieved a relatively good overall performance in both the object detection task and the picking point detection task. Taking the object detection task (B) as an example, its Precision, Recall, mAP0.5, and F1-score reached 89.25%, 84.36%, 90.08%, and 86.74%, respectively. In the keypoint detection task P, the Precision, Recall, mAP0.5, and F1-score reached 88.16%, 84.36%, 89.64%, and 86.22%, respectively. Compared with P = 2, P = 4 improved Precision, Recall, mAP0.5, and F1-score in the object detection task by 3.20, 5.13, 2.20, and 4.24 percentage points, respectively. In the keypoint detection task, Precision, Recall, mAP0.5, and F1-score increased by 3.33, 6.25, 2.54, and 4.89 percentage points, respectively. Meanwhile, the number of model parameters decreased from 2.73 M to 2.66 M, and the computational cost decreased from 7.4 GFLOPs to 7.0 GFLOPs.

When P further increased to 8 and 16, the number of model parameters and the computational cost continued to decrease, but the detection performance and picking point localization performance decreased to varying degrees. This may be because, when the P value is too large, the proportion of 3 × 3 convolution decreases, and only a small number of channels are used for spatial feature extraction, resulting in a weakened ability of the model to model the local spatial structures near tea bud edges, tender stems, and picking points. Especially in small-object detection and fine-grained keypoint localization tasks, local spatial information plays an important role in identifying target edges, tender stem directions, and picking point regions. Therefore, excessively reducing the proportion

of 3 × 3 convolution may weaken the model's fine-grained feature representation ability, thereby affecting the detection and localization performance.

Considering the object detection performance, picking point detection performance, number of parameters, and GFLOPs, P = 4 achieved a good balance between accuracy and computational cost under the current experimental settings of this study. Therefore, P = 4 was selected as the parameter configuration of the C3K2_HetConv module and was used in the subsequent ablation experiments and model comparison experiments.

3.3.2. Ablation Study

To comprehensively evaluate the effectiveness of the proposed enhancements, an ablation study was conducted based on YOLOv11. By combining different core modules, the contributions of the BiFPN, EMA, and HetConv modules to the performance of the tea bud and keypoint detection model were analyzed in depth. Table 3 summarizes the results of the ablation experiments, while Figure 6 presents the corresponding comparative visualization.

Table 3. Results of the ablation experiments.

Model	P(B)/%	R(B)/%	mAP0.5(B)/%	P(P)/%	R(P)/%	mAP0.5(P)/%	Parameters/M	GFLOPs
YOLOv11-pose	81.84	77.97	84.80	82.30	78.53	84.81	2.67	6.7
YOLOv11-pose-B	82.28	80.45	86.58	81.47	79.89	86.28	2.76	7.4
YOLOv11-pose-E	85.27	79.89	86.62	84.88	79.33	84.54	2.62	6.7
YOLOv11-pose-H	82.04	79.09	86.01	80.26	79.89	85.20	2.64	6.5
YOLOv11-pose-BH	84.70	81.01	86.87	84.12	81.01	86.13	2.71	7.0
YOLOv11-pose-EH	86.23	81.56	88.98	85.63	81.01	88.73	2.59	6.5
YOLOv11-pose-BE	88.20	83.53	88.96	87.61	82.99	88.07	2.71	7.4
YOLOv11-pose-BEH	89.25	84.36	90.08	88.16	84.36	89.64	2.66	7.0

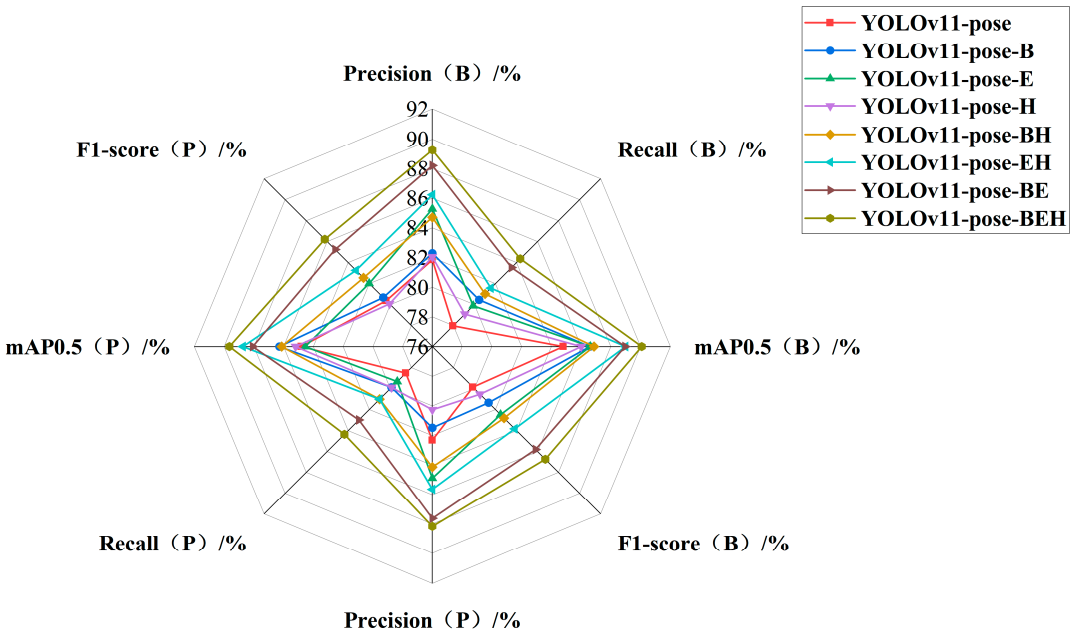

Figure 6. Ablation study results. Note: Metrics with the suffix (B) correspond to the one-bud-two-leaves detection task, while metrics with the suffix (P) correspond to the tea leaf picking point detection task.

As shown by the ablation experiment results, different improved modules played different roles in improving the performance of the YOLOv11-pose model. Compared with the original YOLOv11-pose model, after only introducing the BiFPN module, most indicators of the model in the object recognition task and picking point localization task were improved, indicating that the bidirectional weighted feature fusion mechanism of

BiFPN can enhance the information interaction among features of different scales, thereby improving the model's representation ability for tea bud targets and picking point regions.

After introducing the EMA module, the Precision, mAP0.5, and F1-score of the model in the object recognition task were further improved, and the Precision and F1-score in the picking point localization task were also improved, indicating that the EMA module can enhance the model's attention to key target regions and reduce the interference of complex backgrounds and illumination changes on feature extraction. After only introducing the HetConv module, the model showed a certain improvement in the object recognition task, but some indicators in the picking point localization task fluctuated, indicating that the enhancement effect of HetConv on fine-grained keypoint features is relatively limited when used alone.

Compared with the improvement using a single module, the model performance was improved more significantly after combining multiple modules. Among them, YOLOv11-pose-BE showed good overall performance in both the object recognition task and the picking point localization task. After further introducing HetConv, the final model, YOLOv11-pose-BEH, achieved the best results. In the object recognition task, its Precision, Recall, mAP0.5, and F1-score reached 89.25%, 84.36%, 90.08%, and 86.74%, respectively. In the picking point localization task, the four indicators reached 88.16%, 84.36%, 89.64%, and 86.22%, respectively. Compared with the baseline model, the final model achieved obvious improvements in all indicators of both tasks.

In summary, the BiFPN, EMA, and HetConv modules have good complementarity. BiFPN mainly enhances the multi-scale feature fusion ability, EMA improves the model's ability to select features from effective regions, and HetConv enhances the local structure modeling ability while reducing redundant feature representation. The synergistic effect of the three modules enabled the final model to achieve the best overall performance in both the object recognition task and the picking point localization task, verifying the effectiveness of the improved strategy proposed in this paper.

3.3.3. Comparative Experiments

To verify the performance advantages of the YOLOv11-pose-BEH model in small-object recognition and keypoint detection tasks, this study conducted a systematic comparison with three mainstream keypoint detection models—YOLOv11-pose, YOLOv8-pose-p6, and YOLOv8-pose—under the same dataset and training parameters. The results are shown in Table 4.

Table 4. Results of the comparative experiments.

Model	P(B)/%	R(B)/%	mAP0.5(B)/%	F1(B)/%	P(P)/%	R(P)/%	mAP0.5(P)/%	F1(P)/%	Parameters/M	GFLOPs
YOLOv11-pose-BEH	89.25	84.36	90.08	86.74	88.16	84.36	89.64	86.22	2.66	7.0
YOLOv11-pose	81.84	77.97	84.80	79.86	82.30	78.53	84.81	80.37	2.67	6.7
YOLOv8-pose-p6	74.20	74.01	78.45	74.10	73.88	74.03	78.46	73.95	4.90	8.5
YOLOv8-pose	74.37	71.75	78.65	73.04	74.81	70.62	76.19	72.65	3.09	8.5

Note: P, R, and F1 denote Precision, Recall, and F1-score, respectively. B denotes the object detection task, and P denotes the picking point localization task.

Table 4 shows that the YOLOv11-pose-BEH model proposed in this paper achieved the best performance in both the object recognition task (B) and the picking point localization task (P).

In the object recognition task, YOLOv11-pose-BEH obtained Precision, Recall, mAP0.5, and F1-score values of 89.25%, 84.36%, 90.08%, and 86.74%, respectively, on the test set. Compared with the baseline model YOLOv11-pose, these four metrics increased by 7.41, 6.39, 5.28, and 6.88 percentage points, respectively. Under the same experimental conditions, this result indicates that the proposed improvement strategy improved the object

recognition performance of the original YOLOv11-pose model. Compared with YOLOv8-pose-p6, the Precision, Recall, mAP0.5, and F1-score of YOLOv11-pose-BEH increased by 15.05, 10.35, 11.63, and 12.64 percentage points, respectively. Compared with YOLOv8-pose, the corresponding improvements were 14.88, 12.61, 11.43, and 13.70 percentage points, respectively. These results suggest that YOLOv11-pose-BEH has better ability to extract multi-scale features of tea bud targets in complex tea garden backgrounds, thereby improving object recognition performance.

In the picking point localization task, YOLOv11-pose-BEH obtained Precision, Recall, mAP0.5, and F1-score values of 88.16%, 84.36%, 89.64%, and 86.22%, respectively, on the test set. Compared with the baseline model YOLOv11-pose, these four metrics increased by 5.86, 5.83, 4.83, and 5.85 percentage points, respectively. This result indicates that the proposed model improved the keypoint localization performance of the original YOLOv11-pose model under the same experimental conditions. Compared with YOLOv8-pose-p6, the Precision, Recall, mAP0.5, and F1-score of YOLOv11-pose-BEH increased by 14.28, 10.33, 11.18, and 12.27 percentage points, respectively. Compared with YOLOv8-pose, the corresponding improvements were 13.35, 13.74, 13.45, and 13.57 percentage points, respectively. These results indicate that YOLOv11-pose-BEH achieved better overall performance in both tea bud object recognition and picking point localization tasks under the same experimental configuration.

From the perspective of model complexity, YOLOv11-pose-BEH has 2.66 M parameters, which is lower than the 2.67 M, 4.90 M, and 3.09 M of YOLOv11-pose, YOLOv8-pose-p6, and YOLOv8-pose, respectively. Its computational cost is 7.0 GFLOPs, which is slightly higher than the 6.7 GFLOPs of YOLOv11-pose, but lower than the 8.5 GFLOPs of YOLOv8-pose-p6 and YOLOv8-pose. Overall, YOLOv11-pose-BEH significantly improved the object recognition and picking point localization performance while maintaining a relatively low parameter scale and a moderate computational cost, indicating that the proposed improved model achieved a good balance between accuracy and computational cost.

To further quantify the spatial localization accuracy of the model for tea bud keypoints, especially in the picking point localization task, this study introduces Mean Pixel Error in pixels (MPE) as a supplementary evaluation metric [36]. For each valid keypoint, the pixel error was defined as the Euclidean distance between the predicted keypoint and the manually annotated ground-truth keypoint in the image coordinate system. The MPE value was then obtained by averaging the pixel errors over all valid keypoints involved in the evaluation. In this study, the overall MPE of the six keypoints and the MPE values of the picking points were calculated separately, as shown in Table 5.

Table 5. Comparison of the Mean Pixel Error of Picking Point Localization among Different Models.

Model	All Keypoints MPE/px	Picking Points MPE/px	Points 1 MPE/px	Points 2 MPE/px	Points 3 MPE/px	Points 4 MPE/px	Points 5 MPE/px	Points 6 MPE/px
YOLOv11-pose-BEH	91.27	80.70	94.11	91.41	120.02	75.64	62.23	104.23
YOLOv11-pose	125.68	111.03	73.24	131.77	215.96	85.98	122.21	124.91
YOLOv8-pose-p6	173.12	119.29	164.14	231.78	284.95	105.56	102.13	150.17
YOLOv8-pose	191.53	134.10	196.59	239.30	310.97	122.62	121.10	158.58

As shown in Table 5, the YOLOv11-pose-BEH model achieved the lowest mean pixel error in both overall keypoint localization and picking point localization. Specifically, the All-keypoints MPE and Picking points MPE of YOLOv11-pose-BEH were 91.27 px and 80.70 px, respectively, which were significantly lower than those of the other comparison models. Compared with the original YOLOv11-pose model, YOLOv11-pose-BEH reduced the overall keypoint mean error by 34.41 px and the picking point mean error by 30.33 px. Compared with YOLOv8-pose-p6 and YOLOv8-pose, its overall keypoint errors were

reduced by 81.85 px and 100.26 px, respectively, and its picking point errors were reduced by 38.59 px and 53.40 px, respectively. These results indicate that the improved model can not only maintain high detection accuracy, but also effectively reduce the spatial deviation of keypoint prediction.

From the perspective of single-point errors, the MPE values of YOLOv11-pose-BEH at Point 4, Point 5, and Point 6, which are the three picking points, were 75.64 px, 62.23 px, and 104.23 px, respectively, all of which were the lowest among all models. This indicates that the model has more stable localization ability for picking points corresponding to different picking grades. In contrast, the localization error of Point 3 was relatively large in all models, which may be related to the maturity of the leaf where this point is located and its similarity to the background. Since the leaf color, texture, and edge features of the region corresponding to Point 3 are relatively similar to those of surrounding mature leaves and background branches and leaves, more obvious spatial deviation may occur in keypoint prediction. Overall, the introduction of structures such as BiFPN, EMA, and HetConv enhanced the model's representation ability for local tea bud structures and multi-scale features, thereby improving the accuracy and stability of picking point localization and providing a more reliable visual localization basis for subsequent precise picking by robotic arms.

3.3.4. Visualization and Analysis of Results

To further verify the performance advantages of the improved YOLOv11-pose-BEH model in tea bud and key picking point detection, this study conducted a qualitative comparison with mainstream keypoint detection models, including YOLOv8-pose, YOLOv8-pose-p6, and YOLOv11-pose. Figure 7 presents the visualization results of different models on the test set.

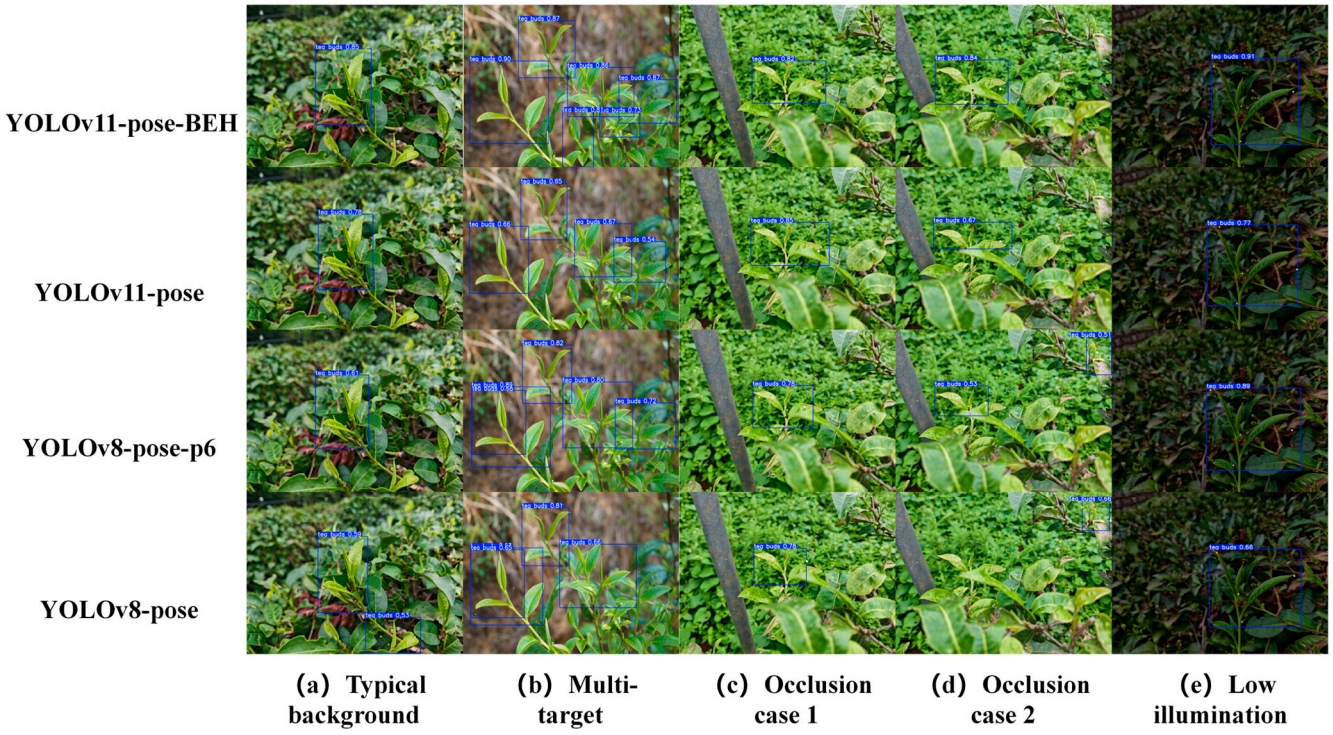

Figure 7. Detection visualization results of different models under representative challenging tea garden scenes: (**a**) typical complex tea garden background; (**b**) multi-target scene; (**c**,**d**) two representative occlusion cases; and (**e**) low-illumination scene.

As shown in Figure 7, under complex tea garden backgrounds, multi-object, occlusion, and low-illumination scenes, all four models showed missed detections, false detections, and keypoint localization deviations to varying degrees, indicating that complex natural environments still cause certain interference with tea bud target recognition and picking point localization. In comparison, the YOLOv11-pose-BEH model showed better overall performance, with a higher degree of fit between the detection boxes and tea bud targets and relatively smaller keypoint localization deviations. It could still maintain stable detection performance under normal complex backgrounds, mild occlusion, and low-illumination scenes. However, in the multi-object recognition scene, due to mutual overlap among tea buds, similar background textures, and large differences in target scales, the YOLOv11-pose-BEH model still showed some keypoint deviations, indicating that there is still room for further optimization in dense target and fine-grained keypoint localization.

In contrast, the YOLOv11-pose model could complete object detection in some scenes, but missed detections occurred in the multi-object recognition scene, and the keypoint localization deviations were relatively obvious under severe occlusion and low-illumination conditions. Compared with YOLOv8-pose, YOLOv8-pose-p6 showed certain improvement and could detect most tea bud targets in the five complex tea garden scenes. However, its detection box fitting degree and keypoint localization accuracy were still lower than those of the YOLOv11-pose-BEH model, especially in the severe occlusion scene, where false detections and smaller detection boxes occurred. The YOLOv8-pose model showed relatively weaker overall performance, with more obvious missed detections, false detections, and picking point deviations in the multi-object recognition and severe occlusion scenes, indicating that it has insufficient adaptability to the joint tea bud detection and keypoint localization task under conditions of similar backgrounds, dense targets, and occlusion.

To further analyze the differences in model perception of key regions, Figure 8 illustrates the heatmap distributions generated by different models on the same samples. In the heatmaps, the red areas represent regions with high response values, while the blue areas indicate weak responses [37]. The results show that the feature attention regions of different models were obviously different under the five scenes, including normal complex background, multi-object recognition, mild occlusion, severe occlusion, and low-illumination environments. The heatmap response of the YOLOv8-pose model was generally scattered, and some high-response regions appeared on background leaves, branches, or non-target regions, indicating that its ability to focus on the main features of tea buds was insufficient under complex backgrounds and occlusion conditions. Compared with YOLOv8-pose, YOLOv8-pose-p6 showed some improvement and could focus on tea buds and their surrounding regions in most scenes. However, its heatmap distribution was still relatively broad, and it was easily affected by adjacent leaves and background textures in multi-object and severe occlusion scenes. The heatmap response of the YOLOv11-pose model was further concentrated toward the main tea bud regions, indicating that its feature extraction ability for target regions was enhanced. However, in some complex scenes, there were still problems such as an excessively large attention range and insufficiently precise response regions. In contrast, the YOLOv11-pose-BEH model showed more concentrated heatmap responses in the five scenes, mainly distributed in the tea bud body, tender stem, and regions near the picking points, with relatively fewer responses to background interference. This indicates that the improved model proposed in this paper can more effectively extract key features related to tea bud recognition and picking point localization, and has stronger feature discrimination and target focusing abilities in complex tea garden environments.

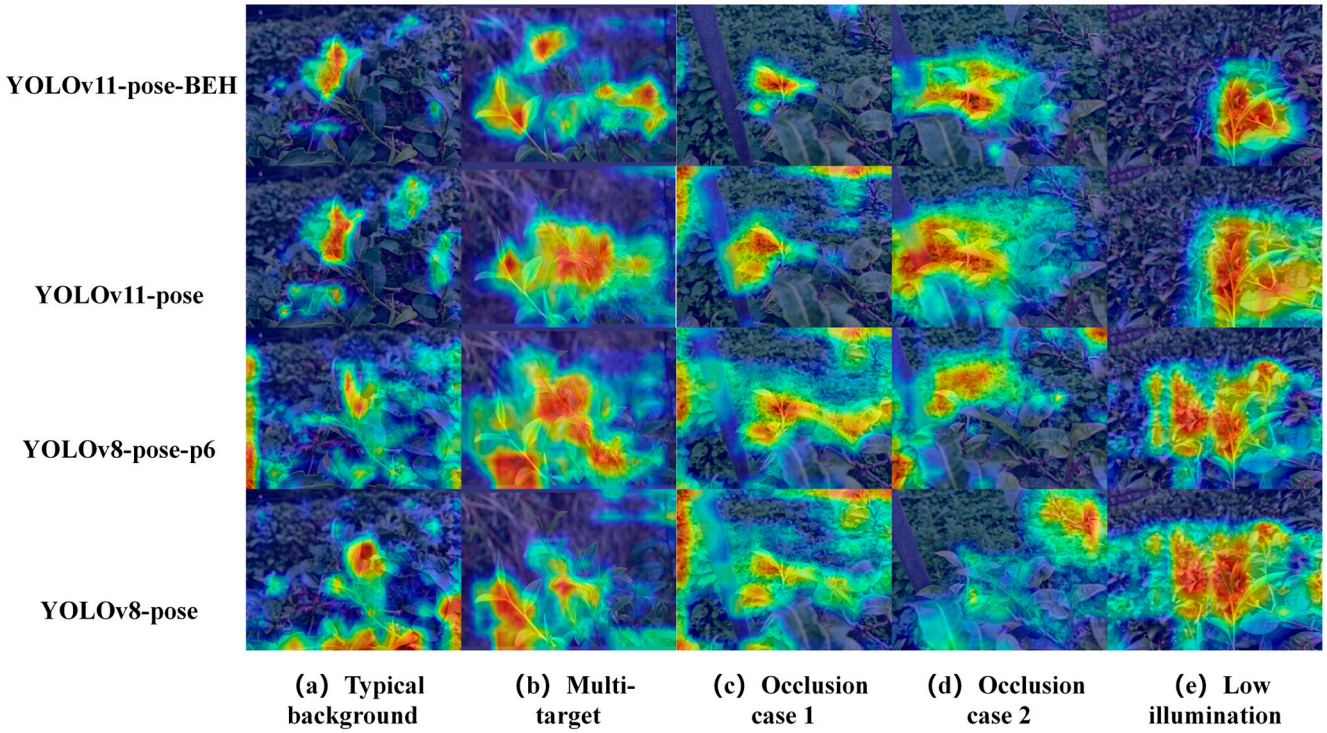

Figure 8. Visualization of heatmap results from different models under representative challenging tea garden scenes: (**a**) typical complex tea garden background; (**b**) multi-target scene; (**c**,**d**) two representative occlusion cases; and (**e**) low-illumination scene.

In summary, the improved YOLOv11-pose-BEH model achieved significant improvements not only in quantitative metrics but also in qualitative visualization results, demonstrating superior feature focusing capability and target recognition accuracy. The comparative experiments confirm that the proposed model exhibits remarkable advantages in detecting small targets under complex backgrounds. These findings indicate that YOLOv11-pose-BEH can effectively enhance the accuracy of tea bud identification and picking-point localization, providing theoretical and technical support for high-precision perception and decision-making in intelligent tea-picking robot vision systems.

4. Discussion

As shown by the visualization results of complex scenes, although YOLOv11-pose-BEH achieved better overall detection performance than the original YOLOv11-pose, relatively obvious keypoint localization deviations still occurred in the multi-object recognition scene. The possible reasons for this problem are mainly as follows. (1) In multi-object scenes, the distances between tea buds are relatively small, and some buds and leaves overlap or cross in the image, causing spatial interference among the leaves, tender stems, and picking point regions of different targets. When adjacent tea buds have similar morphological features, the model may easily associate the local features of one target with another target, resulting in keypoint matching deviations. (2) The picking point of tea buds is not an independent visual target with obvious boundaries, but a spatial position determined by the morphology of buds and leaves, the petiole connection relationship, and the extension direction of tender stems. In dense multi-object scenes, tender stem regions are easily occluded by leaves or mixed with background branches and leaves, making it difficult for the model to accurately distinguish the picking point positions corresponding to different targets. (3) Although the proposed model enhances multi-scale feature fusion through BiFPN and improves the feature representation of key regions through EMA, the

local response regions may simultaneously cover multiple tea bud targets in dense target scenes, leading to inter-target feature competition during keypoint prediction. Especially for fine-grained positions such as picking points, when the detection box contains multiple similar structures or edge information from adjacent targets, the model may produce spatial localization deviations.

Compared with the original YOLOv11-pose model, the proposed YOLOv11-pose-BEH model achieved varying degrees of improvement in Precision, Recall, mAP0.5, and F1-score, and also showed better performance in terms of keypoint spatial localization error. The experimental results show that the All-keypoints MPE and Picking points MPE of YOLOv11-pose-BEH were 91.27 px and 80.70 px, respectively, which were 34.41 px and 30.33 px lower than those of the original YOLOv11-pose model. This indicates that the improved model not only enhanced the tea bud detection performance, but also effectively reduced the spatial offset between the predicted picking point positions and the manually annotated ground-truth positions. This is mainly attributed to the BiFPN structure, which enhances the information fusion ability among features of different scales; the EMA mechanism, which improves the model's attention to the tea bud body, tender stem, and picking point-related regions; and the HetConv module, which enhances local feature extraction ability while maintaining the lightweight characteristics of the model. Therefore, YOLOv11-pose-BEH can maintain good detection stability and keypoint localization ability under complex tea garden backgrounds.

In addition, compared with existing tea bud detection and picking point localization methods, the proposed method has certain differences in task objectives and model output forms. Methods such as TS-YOLO, Tea-YOLO, and TeaBudNet mainly focus on small-object detection tasks for tea buds or tea shoots, with emphasis on improving detection accuracy, lightweight deployment capability, and model robustness in complex tea garden environments. However, most methods do not directly output the picking point positions under different picking standards. The YOLO-Tea method proposed by Shuai et al. is based on the YOLOv5 framework, introduces modules such as CARAFE, CBAM, and Bottleneck Transformer, and adds six keypoint regression heads to achieve tea shoot detection and picking point localization. This method further combines the geometric relationships between keypoints and image post-processing methods to determine the picking point positions under different picking grades, providing an important reference for target recognition and picking point localization in intelligent tea picking. Compared with this method, the present study differs in the definition of picking points and the form of model output. The proposed YOLOv11-pose-BEH annotates and predicts the two-dimensional picking points under different picking standards as keypoints, and establishes corresponding picking point localization outputs for single bud, one bud with one leaf, and one bud with two leaves. This design reduces, to some extent, the dependence on post-processing steps such as midpoint calculation of keypoint connections or skeleton extraction during the inference stage, enabling the model to directly obtain the two-dimensional coordinates of multiple types of picking points while outputting tea bud target boxes.

The picking point localization accuracy evaluated in this paper is mainly based on two-dimensional pixel coordinates in the image plane. The two-dimensional localization error of the model is measured by the pixel distance between the predicted keypoints and the manually annotated keypoints. Therefore, the reduction in MPE indicates that the predicted keypoint positions of the model in RGB images are closer to the ground-truth annotated positions, but it does not mean that the three-dimensional spatial coordinates required for field picking execution are directly obtained. In practical robotic picking applications, the end effector also needs to combine camera calibration, depth information, or binocular vision results to further convert the two-dimensional picking points into three-dimensional

spatial coordinates, so as to achieve precise control of the tea bud picking positions. In addition, recent machine learning-based agricultural decision-support studies have shown that soil nutrients, soil pH, and climatic variables can be integrated for crop-related prediction and recommendation tasks [38]. This suggests that future intelligent tea-picking systems may further combine visual perception results with environmental and agronomic information, such as illumination conditions, tea bud growth status, and field environment variables, to improve their adaptability under natural tea garden conditions. Therefore, the proposed model is more suitable as a front-end detection and two-dimensional localization basis in the visual perception stage of intelligent tea-picking robots, providing key position information in the image plane for subsequent three-dimensional localization, path planning, and end-effector control.

5. Conclusions

In this study, aiming at the problem that the detection accuracy of tea buds and leaves is easily affected by light variation, leaf overlap, and small target scale in complex tea garden backgrounds, an improved YOLOv11-based method for tea bud and leaf recognition and keypoint localisation is proposed. Based on YOLOv11-pose, three modules—BiFPN, EMA, and HetConv—were introduced to optimize the network's feature fusion, attention modeling, and convolution operation. The BiFPN module enhances the interaction and transmission capability of multi-scale features, effectively improving the robustness of the model for bud and leaf targets of different scales; the EMA module improves the weight distribution between global and local features, making the model better highlight key regions in complex backgrounds; the HetConv module reduces redundant computation while strengthening fine-grained feature extraction, which helps improve the detection accuracy of small targets.

The experimental results show that the improved YOLOv11-pose-BEH model was generally superior to the original YOLOv11-pose model and the YOLOv8 series comparison models in core indicators such as Precision, Recall, mAP0.5, and F1-score, while maintaining a relatively low parameter scale and computational cost. Further keypoint error analysis showed that YOLOv11-pose-BEH could effectively reduce the two-dimensional pixel deviation between the predicted keypoints and the ground-truth annotated keypoints. The All-keypoints MPE and Picking points MPE were both lower than those of the original YOLOv11-pose model, indicating that this model has better overall performance in tea bud target recognition and two-dimensional picking point localization.

In summary, YOLOv11-pose-BEH can effectively perform tea bud and leaf target detection and two-dimensional picking point coordinate prediction in complex tea garden environments, providing an effective front-end recognition and localization method for the visual perception stage of intelligent tea-picking robots. However, this study is still mainly based on RGB images for two-dimensional picking point localization, and has not yet further achieved three-dimensional spatial localization for actual picking execution. Future research will focus on exploring cross-modal fusion methods between RGB images and 3D point cloud information on the basis of the current study, so as to achieve more accurate tea bud and leaf recognition and three-dimensional picking point localization. Meanwhile, further research on model lightweighting and edge deployment will be carried out to improve the running efficiency and practical application capability of the model on resource-constrained terminal devices.

Author Contributions: Conceptualization, Z.C.; methodology, M.Z. (Miao Zhou), T.W. and Z.C.; software, W.L., C.W., K.P. and S.F.; validation, W.L., T.W. and M.Z. (Man Zou); formal analysis, M.Z. (Miao Zhou); investigation, M.Z. (Man Zou) and S.F.; data curation, J.H. and Z.C.; resources, B.W.; writing—original draft preparation, W.L. and X.C.; writing—review and editing, W.L., J.H., C.W. and

B.W.; visualization, X.C. and K.P.; supervision, C.Z.; project administration, C.Z. and B.W.; funding acquisition, B.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Yunnan Provincial Innovation Team for Artificial Intelligence and Big Data Application in the Tea Industry, grant number 202405AS350025; the Yunnan International Joint Laboratory for Smart Tea Industry, grant number 202403AP140022; the Yunnan Tea Industry Technology Innovation Center, grant number 202505AK340010; and the Research on Intelligent Identification and Localisation of Tea Buds and Leaves Based on 3D Point Clouds and Deep Learning Technology, grant number 2025Y0579.

Data Availability Statement: The data presented in this study are available upon reasonable request with the approval of the corresponding author.

Acknowledgments: The authors would like to acknowledge the College of Tea Science, Yunnan Agricultural University, and the Yunnan Organic Tea Industry Intelligent Engineering Research Center for their support during this study.

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

References

1. Pan, S.-Y.; Nie, Q.; Tai, H.-C.; Song, X.-L.; Tong, Y.-F.; Zhang, L.-J.; Wu, X.-W.; Lin, Z.-H.; Zhang, Y.-Y.; Ye, D.-Y.; et al. Tea and tea drinking: China's outstanding contributions to the mankind. *Chin. Med.* **2022**, *17*, 27. [CrossRef] [PubMed]
2. Wang, F.; Yang, X.; Li, W.; Ning, P.; Huang, X. Metabolomics and transcriptomics reveal the accumulation of key metabolites and flavor formation in the fresh leaves of wild and cultivated tea plants. *Ind. Crop. Prod.* **2025**, *235*, 121780. [CrossRef]
3. Chen, M.-M.; Liao, Q.-H.; Qian, L.-L.; Zou, H.-D.; Li, Y.-L.; Song, Y.; Xia, Y.; Liu, Y.; Liu, H.-Y.; Liu, Z.-L. Effects of geographical origin and tree age on the stable isotopes and multi-elements of Pu-erh tea. *Foods* **2024**, *13*, 473. [CrossRef] [PubMed]
4. Wang, J.; Li, X.; Yang, G.; Wang, F.; Men, S.; Xu, B.; Xu, Z.; Yang, H.; Yan, L. Research on tea trees germination density detection based on improved YOLOv5. *Forests* **2022**, *13*, 2091. [CrossRef]
5. Wei, Y.; Wen, Y.; Huang, X.; Ma, P.; Wang, L.; Pan, Y.; Lv, Y.; Wang, H.; Zhang, L.; Wang, K.; et al. The dawn of intelligent technologies in tea industry. *Trends Food Sci. Technol.* **2024**, *144*, 104337. [CrossRef]
6. Shi, M.; Zheng, D.; Wu, T.; Zhang, W.; Fu, R.; Huang, K. Small object detection algorithm incorporating Swin Transformer for tea buds. *PLoS ONE* **2024**, *19*, e0299902. [CrossRef] [PubMed]
7. Li, H.; Kong, M.; Shi, Y. Tea bud detection model in a real picking environment based on an improved YOLOv5. *Biomimetics* **2024**, *9*, 692. [CrossRef] [PubMed]
8. Li, Y.; He, L.; Jia, J.; Lv, J.; Chen, J.; Qiao, X.; Wu, C. In-field tea shoot detection and 3D localization using an RGB-D camera. *Comput. Electron. Agric.* **2021**, *185*, 106149. [CrossRef]
9. Wang, Z.; Zhang, S.; Chen, Y.; Xia, Y.; Wang, H.; Jin, R.; Wang, C.; Fan, Z.; Wang, Y.; Wang, B. Detection of small foreign objects in Pu-erh sun-dried green tea: An enhanced YOLOv8 neural network model based on deep learning. *Food Control* **2025**, *168*, 110890. [CrossRef]
10. Zhang, L.; Zou, L.; Wu, C.; Jia, J.; Chen, J. Method of famous tea sprout identification and segmentation based on improved watershed algorithm. *Comput. Electron. Agric.* **2021**, *184*, 106108. [CrossRef]
11. Yang, M.; Yuan, W.; Xu, G. YOLOX target detection model can identify and classify several types of tea buds with similar characteristics. *Sci. Rep.* **2024**, *14*, 2855. [CrossRef] [PubMed]
12. Zhang, Z.; Lu, Y.; Zhao, Y.; Pan, Q.; Jin, K.; Xu, G.; Hu, Y. TS-YOLO: An all-day and lightweight tea canopy shoots detection model. *Agronomy* **2023**, *13*, 1411. [CrossRef]
13. Zhang, S.; Yang, H.; Yang, C.; Yuan, W.; Li, X.; Wang, X.; Zhang, Y.; Cai, X.; Sheng, Y.; Deng, X.; et al. Edge device detection of tea leaves with one bud and two leaves based on ShuffleNetv2-YOLOv5-Lite-E. *Agronomy* **2023**, *13*, 577. [CrossRef]
14. Li, J.; Li, J.; Zhao, X.; Su, X.; Wu, W. Lightweight detection networks for tea bud on complex agricultural environment via improved YOLOv4. *Comput. Electron. Agric.* **2023**, *211*, 107955. [CrossRef]
15. Li, Y.; Zhang, Z.; Zhang, J.; Shi, J.; Zhu, X.; Chen, B.; Lan, Y.; Jiang, Y.; Cai, W.; Tan, X.; et al. TeaBudNet: A lightweight framework for robust small tea bud detection in outdoor environments via Weight-FPN and adaptive pruning. *Agronomy* **2025**, *15*, 1990. [CrossRef]
16. Wang, T.; Zhang, K.; Zhang, W.; Wang, R.; Wan, S.; Rao, Y.; Jiang, Z.; Gu, L. Tea picking point detection and location based on Mask-RCNN. *Inf. Process. Agric.* **2023**, *10*, 267–275. [CrossRef]
17. Pan, Z.; Gu, J.; Wang, W.; Fang, X.; Xia, Z.; Wang, Q.; Wang, M. Picking point identification and localization method based on Swin-Transformer for high-quality tea. *J. King Saud Univ. Comput. Inf. Sci.* **2024**, *36*, 102262. [CrossRef]

18. Shuai, L.; Mu, J.; Jiang, X.; Chen, P.; Zhang, B.; Li, H.; Wang, Y.; Li, Z. An improved YOLOv5-based method for multi-species tea shoot detection and picking point location in complex backgrounds. *Biosyst. Eng.* **2023**, *231*, 117–132. [CrossRef]
19. Zhao, B.; Gong, X.; Wang, J.; Zhao, L. Low-light image enhancement based on normal-light image degradation. *Signal Image Video Process.* **2022**, *16*, 1409–1416. [CrossRef]
20. Maji, D.; Nagori, S.; Mathew, M.; Poddar, D. YOLO-Pose: Enhancing YOLO for multi person pose estimation using object keypoint similarity loss. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, New Orleans, LA, USA, 19–20 June 2022; pp. 2636–2645. [CrossRef]
21. Khanam, R.; Hussain, M. YOLOv11: An overview of the key architectural enhancements. *arXiv* **2024**, arXiv:2410.17725. [CrossRef]
22. Afifah, V.; Erniwati, S. YOLOv8 for object detection: A comprehensive review of advances, techniques, and applications. *Int. J. Adv. Comput. Inform.* **2025**, *2*, 53–61. [CrossRef]
23. Wu, C.; Zhang, S.; Wang, W.; Wu, Z.; Yang, S.; Chen, W. Computation and analysis of phenotypic parameters of Scylla paramamosain based on YOLOv11-DYPF keypoint detection. *Aquac. Eng.* **2025**, *111*, 102571. [CrossRef]
24. Zhang, Y.; Liu, Z.; Guo, X.; Li, C.; Teng, G. Wheat head detection in field environments based on an improved YOLOv11 model. *Agriculture* **2025**, *15*, 1765. [CrossRef]
25. Cao, X.; Li, Y.; Zhang, Y.; Zhong, Z.; Bai, R.; Yang, P.; Pan, F.; Fu, X. Full-time sequence assessment of okra seedling vigor under salt stress based on leaf area and leaf growth rate estimation using the YOLOv11-HSECal instance segmentation model. *Front. Plant Sci.* **2025**, *16*, 1625154. [CrossRef] [PubMed]
26. Tan, M.; Pang, R.; Le, Q.V. EfficientDet: Scalable and efficient object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 10781–10790. [CrossRef]
27. Wu, Z.; Zhang, Y.; Zhang, Z.; Shen, F.; Li, L.; He, X.; Zhong, H.; Zhou, Y. An improved YOLOv8n-based method for detecting rice shelling rate and brown rice breakage rate. *Agriculture* **2025**, *15*, 1595. [CrossRef]
28. Lin, T.-Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature Pyramid Networks for object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 936–944. [CrossRef]
29. Liu, S.; Qi, L.; Qin, H.; Shi, J.; Jia, J. Path Aggregation Network for instance segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 8759–8768. [CrossRef]
30. Xu, W.; Yang, R.; Karthikeyan, R.; Shi, Y.; Su, Q. GBiDC-PEST: A novel lightweight model for real-time multiclass tiny pest detection and mobile platform deployment. *J. Integr. Agric.* **2025**, *24*, 2749–2769. [CrossRef]
31. Li, H.; Mo, Y.; Chen, J.; Chen, J.; Li, J. Accurate Orah fruit detection method using lightweight improved YOLOv8n model verified by optimized deployment on edge device. *Artif. Intell. Agric.* **2025**, *15*, 707–723. [CrossRef]
32. Ouyang, D.; He, S.; Zhang, G.; Luo, M.; Guo, H.; Zhan, J.; Huang, Z. Efficient multi-scale attention module with cross-spatial learning. In Proceedings of the ICASSP 2023—2023 IEEE International Conference on Acoustics, Speech and Signal Processing, Rhodes Island, Greece, 4–10 June 2023; pp. 1–5. [CrossRef]
33. Wang, Q.; Yan, N.; Qin, Y.; Zhang, X.; Li, X. BED-YOLO: An enhanced YOLOv10n-based tomato leaf disease detection algorithm. *Sensors* **2025**, *25*, 2882. [CrossRef] [PubMed]
34. Singh, P.; Verma, V.K.; Rai, P.; Namboodiri, V.P. HetConv: Heterogeneous kernel-based convolutions for deep CNNs. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 4830–4839. [CrossRef]
35. Chen, Y.; Xu, H.; Chang, P.; Huang, Y.; Zhong, F.; Jia, Q.; Chen, L.; Zhong, H.; Liu, S. CES-YOLOv8: Strawberry maturity detection based on the improved YOLOv8. *Agronomy* **2024**, *14*, 1353. [CrossRef]
36. Du, W.; Yi, G.; Omisore, O.M.; Duan, W.; Chen, X.; Akinyemi, T.; Liu, J.; Lee, B.-G.; Wang, L. Guidewire endpoint detection based on pixel-adjacent relation during robot-assisted intravascular catheterization: In vivo mammalian models. *Adv. Intell. Syst.* **2024**, *6*, 2300687. [CrossRef]
37. Selvaraju, R.R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; Batra, D. Grad-CAM: Visual explanations from deep networks via gradient-based localization. *Int. J. Comput. Vis.* **2020**, *128*, 336–359. [CrossRef]
38. Dey, B.; Ferdous, J.; Ahmed, R. Machine learning based recommendation of agricultural and horticultural crop farming in India under the regime of NPK, soil pH and three climatic variables. *Heliyon* **2024**, *10*, e25112. [CrossRef] [PubMed]