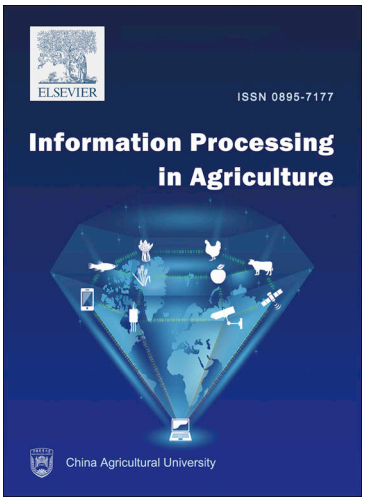
Journal Pre-proofs



YOLO-MEST: a re-parameterized multi-scale fusion model with enhanced detection head for high-accuracy tea bud detection

Chuanyang Yu, Yi Xue, Liuyang Zhang, Xue An, Ce Liu, Liqing Chen

PII: S2214-3173(25)00060-5
DOI: https://doi.org/10.1016/j.inpa.2025.10.001
Reference: INPA 485

To appear in: *Information Processing in Agriculture*

Received Date: 18 July 2025
Revised Date: 27 September 2025
Accepted Date: 20 October 2025

Please cite this article as: C. Yu, Y. Xue, L. Zhang, X. An, C. Liu, L. Chen, YOLO-MEST: a re-parameterized multi-scale fusion model with enhanced detection head for high-accuracy tea bud detection, *Information Processing in Agriculture* (2025), doi: https://doi.org/10.1016/j.inpa.2025.10.001

YOLO-MEST: A Re-parameterized Multi-Scale Fusion Model with Enhanced Detection Head for High-Accuracy Tea Bud Detection

Chuanyang Yu[a,b], Yi Xue[a], Liuyang Zhang[a], Xue An[a], Ce Liu[a], Liqing Chen[a,*]

[a] *College of Engineering, Anhui Agricultural University, Hefei, China*

[b] *Institute of Machinery and Electrical Engineering, Anhui Jianzhu University, Hefei, China*

* Corresponding author.

Liqing Chen: lqchen@ahau.edu.cn. The given name is Liqing, and the family name is Chen.

Abstract: *Accurate recognition of tea buds is essential for automated harvesting. However, conventional image-based methods have difficulty with complex field conditions, such as varying illumination, occlusion, and cluttered backgrounds. To overcome these challenges, we introduce YOLO-MEST, a new detection model that incorporates multi-scale feature fusion. This model introduced the RepNCSELAN4 module to enhance feature extraction capabilities and the SPPELAN module to improve feature fusion. The efficient detection head, LiteShiftHead, was added to the network output to improve the accuracy of bounding boxes and classification regression. An improved loss function dDIoU based on the difference in the width-to-height ratio of bounding boxes was designed to enhance the accuracy of bounding box localization further. To form a complete detection-to-picking pipeline, a morphological algorithm-based tea bud picking point estimation algorithm was further proposed, which effectively determined the tea bud picking points based on the detection results. We conducted performance tests on tea bud recognition based on a self-built dataset of high-quality tea. Compared to the original YOLOv8 model, the YOLO-MEST model increased mAP50 by 1.7% and mAP by 4.4%, respectively. Ablation studies confirm the contribution of each component. The proposed method significantly improves detection accuracy and supports practical intelligent tea harvesting.*

Keywords: *Computer vision, tea bud, YOLO framework, Loss function, Picking localization*

1. Introduction

Tea is regarded as one of the most important beverages in the world. As of 2020, the global tea cultivation area was approximately 5 million hectares, with China's tea cultivation area reaching approximately 3.27 million hectares, thus ranking first in the world in terms of production. Among these, high-quality teas hold the highest economic value, contributing as much as 75% of the total market value of tea[1]. Currently, tea picking is predominantly carried out through manual labour, a method that faces substantial challenges, including high labour costs and reduced productivity. Moreover, conventional mechanical harvesting often results in the damage of tea buds, thereby severely compromising tea quality and leading to substantial economic losses [2]. The adoption of intelligent harvesting machine technology is emerging as a trend, and the precise identification of tea buds is a core technology for achieving intelligent harvesting. The precise identification of tea buds has become one of the most challenging technical problems in the field of intelligent product identification. Tender buds exhibit specific characteristics that make them challenging to identify. They are typically small in size, ranging from 10 to 35 mm, and often blend in with their surroundings, such as older leaves and branches. Moreover, they can block each other, making it even more challenging to see. These factors are further complicated by the complex and variable lighting conditions found in tea gardens.

To achieve precise identification of tea buds, numerous scholars have conducted extensive research and exploration into methods for detecting tea buds. Methods can be categorized into two main frameworks: traditional image processing algorithms and deep learning-based object detection approaches. Traditional image processing algorithms primarily rely on color space analysis and morphological feature extraction for object segmentation. The typical methods include the watershed algorithm with color thresholds[3], the combination of B-G and Otsu algorithms[4], and the K-means clustering algorithm[5]. While these methods show promise in controlled environments, their dependence on manual feature engineering presents significant challenges in the complex and unstructured contexts of tea gardens. Changes in natural lighting can lead to variations in color features, which may result in misclassification. Occlusion of leaves and branches significantly affects the accuracy of morphological parameter calculations due to texture feature loss. Moreover, the tension between real-time processing needs and algorithmic complexity restricts their practical application.

Object detection methods based on deep learning have become a major research focus due to their enhanced feature extraction abilities. These methods can be classified into two main categories: (1) two-stage approaches (e.g., Faster R-CNN[6]), which first extract candidate regions and then perform detection, achieving high accuracy but requiring significant computational resources. For example, the SORC model[7] enhanced the region proposal network, achieving an mAP of 82.3% in tea bud detection. Chen et al.[8] integrated multi-task learning to achieve tea bud recognition with an mAP of 79% and to localize picking points. (2) One-stage methods (e.g., the YOLO series) directly predict target locations, resulting in faster

processing speeds. Yang et al.[9] demonstrated an enhancement to the YOLOv3 model for the purpose of tea bud detection. Li et al.[10] incorporated an attention mechanism into the YOLOv4 model to enhance its performance. Additionally, Gui et al.[11] reduced the parameter count of the YOLOv5 model through lightweight modifications.

The prevailing tea bud detection methods principally depend on visual features. However, these methods frequently overlook critical tea plant growth characteristics, including bud-leaf angle and spatial distribution. This has resulted in a high rate of false positives in scenes with high density. To address this, the present paper proposes YOLO-MEST (YOLO based on Multiple Enhanced Strategies for Tea), which is based on YOLOv8[12]. YOLO-MEST enhances the performance of tea bud picking point detection and picking point localization through the following improvements:

(i) The initial component of the proposed methodology is an improvement in the extraction of features. The replacement of the C2f module of PAFP with RepNCSPELAN4 is intended to enhance multi-scale feature fusion capabilities.

(ii) Optimized detection head: The LiteShiftHead has been developed for the purpose of enhancing the detection and selection of small objects in complex backgrounds.

(iii) The loss function has been refined. The incorporation of a shape difference penalty term into the DIoU framework is intended to enhance the accuracy of localization.

(iv) A post-processing module for harvesting point estimation is proposed, which integrates morphological processing on the detected regions to generate precise tea bud picking coordinates.

In summary, the proposed YOLO-MEST is a significantly modified version of the YOLOv8 architecture, specifically redesigned for the challenges of tea bud detection. The novelty of this work lies in the novel integration and synergistic effect of these enhancements rather than in any single component alone.

2. Materials and methods

2.1. Teas database building

This experiment used a diverse dataset collected in Huoshan County, Lu'an City, Anhui Province, China. The images were captured in early April 2024 using a Huawei Mate 60 Pro device, showcasing tea buds from premium varieties of tea, photographed from different angles and in various weather conditions. The dataset, rich in variety, comprises 4,600 high-resolution images (4096×2160 px) of four premium tea varieties (Anji White, Shu, Wuniu, and Wan Nong 95), captured under varying illumination conditions (sunny/cloudy) and angles (0 to 45-degree tilt). Fig. 1

shows an example of the collected premium tea buds. Tea buds can be classified into two categories based on their appearance: Y-type and I-type. The images were labeled using the LabelImg tool, with tea buds manually categorized as Y-type (buds and leaves unfolded) or I-type (buds and leaves not unfolded), as illustrated in the upper right corner of Fig. 1. The annotations were then saved in the YOLO format. Additionally, a subset of these images was annotated with key points—specifically, the apex of the bud and the base of the petiole—using LabelMe. This was done to evaluate the picking point estimation algorithm.

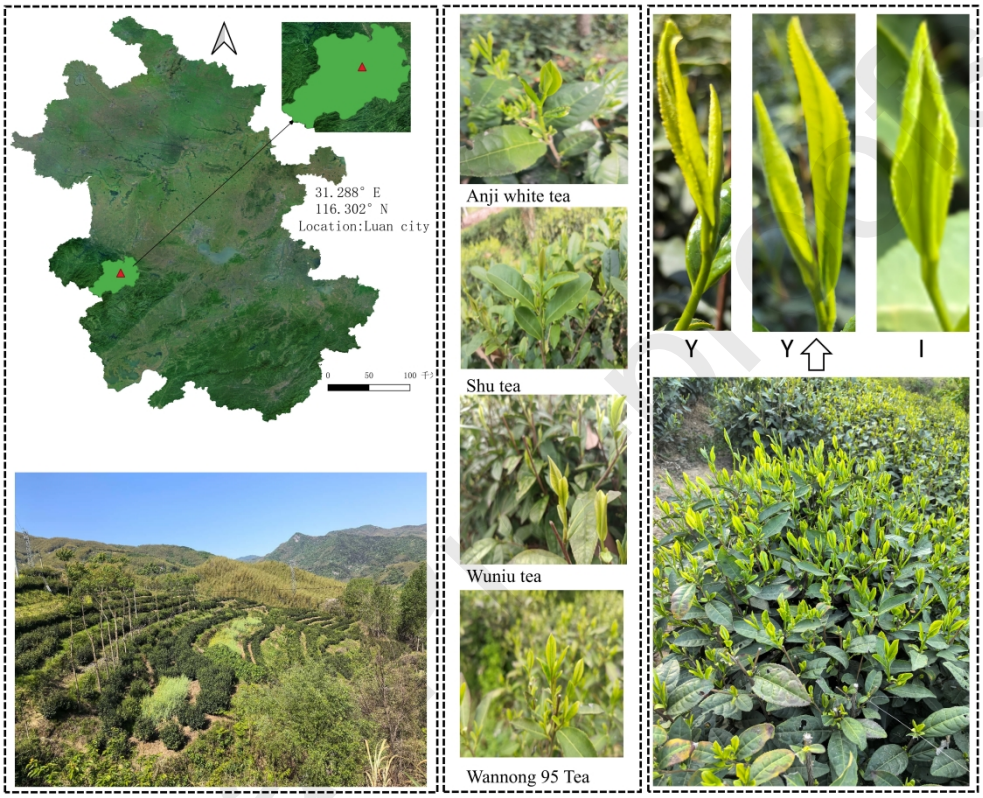


Fig. 1. Composition of the tea bud dataset.

2.2. YOLO-MEST model

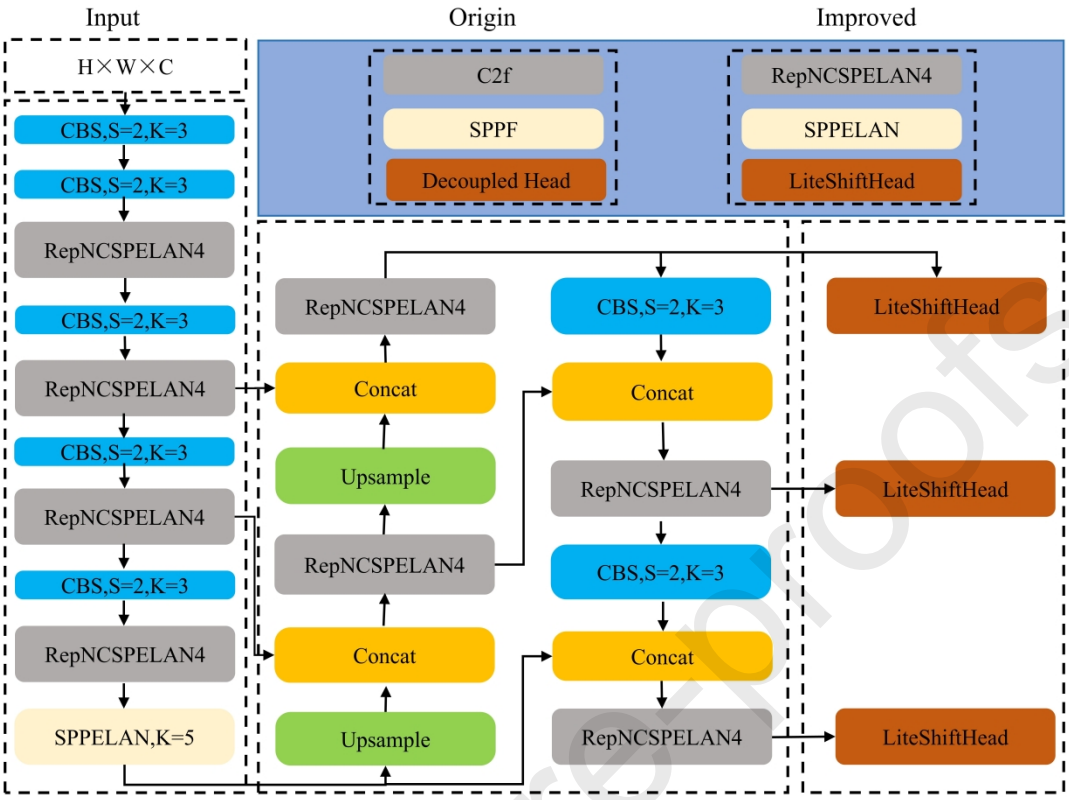


Fig. 2. Improved YOLO-MEST model.

To enhance target detection performance in complex agricultural environments, this study uses YOLOv8 as the baseline model and introduces the improved YOLO-MEST architecture. The main advancements include: (1) the reconstruction of the feature extraction network with the RepNCSELAN4 module; (2) the design of a lightweight, decoupled detection head called LiteShiftHead; and (3) the introduction of a dynamic distance intersection over union (dDIoU) loss function. As shown in Fig. 2, this architecture aims to optimize both feature representation capabilities and computational efficiency through multidimensional improvements.

2.2.1. RepNCSELAN4 module

In the detection of tea buds, the presence of old leaves and the small size of the buds can create significant background interference. Traditional convolution operations depend on fixed sampling points, which may lead to the oversight of complex spatial variations, especially when the target is intricate and affected by background noise.

The reparameterization module employs a branched structure during training and a fused single-branch structure during inference. Both structures are mathematically equivalent. The multi-branch structure aids in gradient flow and model learning, while the single-path structure is more computationally efficient. Reparameterization techniques, along with effective feature aggregation strategies, enhance the model's ability to extract features while minimizing computational overhead. As a result, using

RepNCSELAN4 allows the model to capture information more effectively, enabling it to differentiate between tea buds and old leaves.

RepNCSELAN4 features an innovative architecture, as shown in Fig. 3, combining three essential elements: (1) a reparameterized Conv-BN block that enriches training features, (2) cross-stage fusion through concatenation, and (3) a robust ELAN-based hierarchical aggregation. This advanced parameterization enables the model to learn richer feature representations while maintaining efficiency during inference.

The RepNCSP structure is similar to the C2f module, consisting of a convolutional layer and multiple RepNBottleneck modules, which are designed with a residual structure. With the YOLOv7 ELAN foundation, RepNCSELAN4 enhances its capabilities through reparameterization and cross-stage feature fusion, leading to superior multi-level feature aggregation. Consequently, the backbone network and C2f module in RepNCSELAN4 significantly improve the model's focus on tea buds.

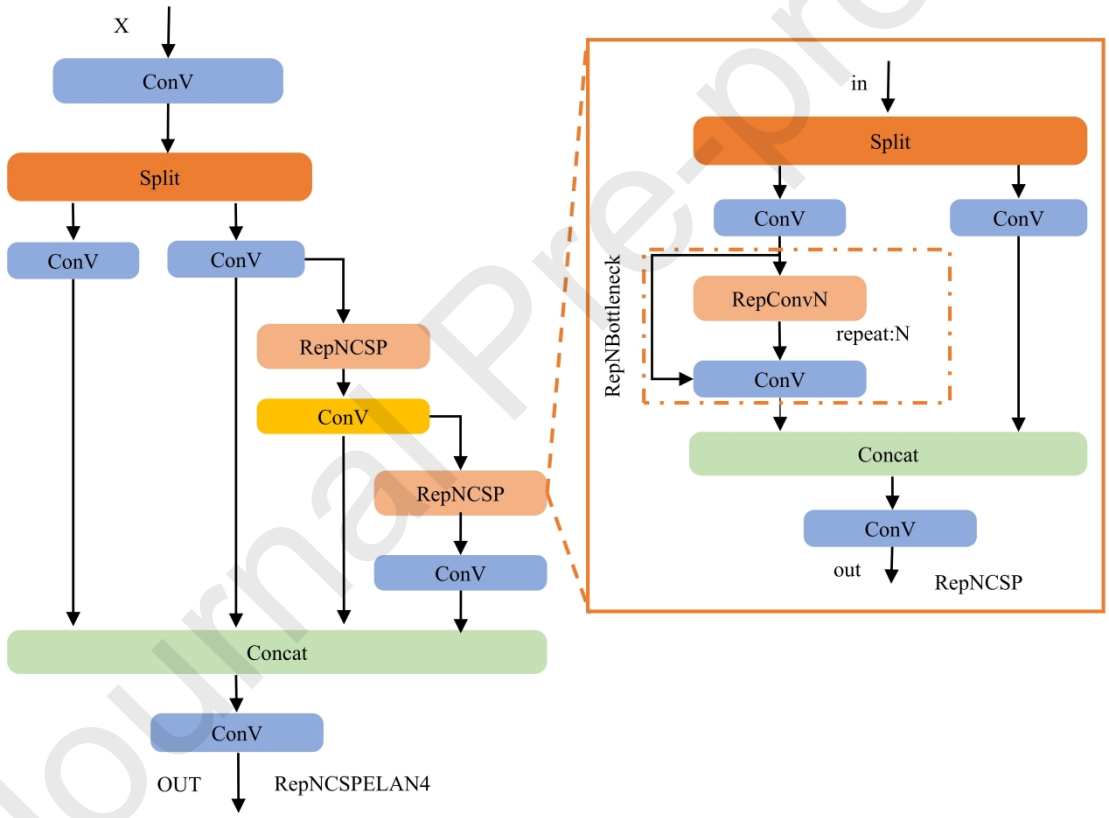


Fig. 3. RepNCSELAN4 module.

2.2.2. SPPELAN feature fusion

In the original YOLOv8 model, the backbone feature extraction uses the SPPF structure, while an enhanced approach, SPPELAN (Fig. 4), is utilized at the end. The main difference is in their processes: SPPF feeds the output of one pooling layer into the next, repeating the operation. In contrast, SPPELAN retains the results from each independent pooling branch, providing a richer multi-scale feature representation.

This method improves the model's ability to detect tender bud targets in complex scenarios.

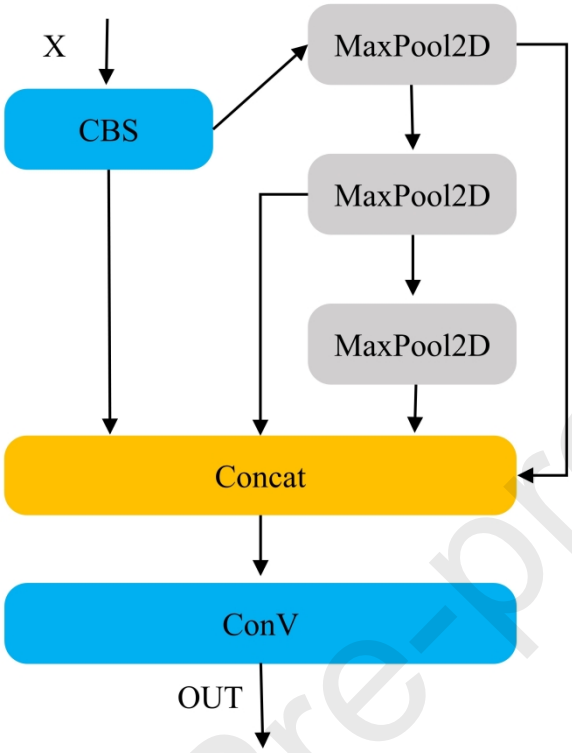


Fig. 4. SPPELAN data stream structure

2.2.3. Improved lightweight decoupling detection head

The original YoloV8 detection head, shown in Fig. 5a, combines two main parts: the REG and CLS modules. It processes features at different scales to produce detection results. The upgraded LiteShiftHead, illustrated in Fig. 5b, improves this design by adding SPConv (Spatial Channel Mixing Convolution) and DWConv (Depthwise Convolution) modules. The SPConv module uses channel segmentation and applies 3×3 group convolution along with 1×1 convolution to blend features effectively. The DWConv module uses different filters for each input channel, which helps capture spatial information better. In the REG module, there are two groups of standard convolution (Conv) and one group of SPConv to enhance regression tasks. The CLS module includes one group of Conv and one group of DWConv to improve classification. This updated design aims to boost detection performance.

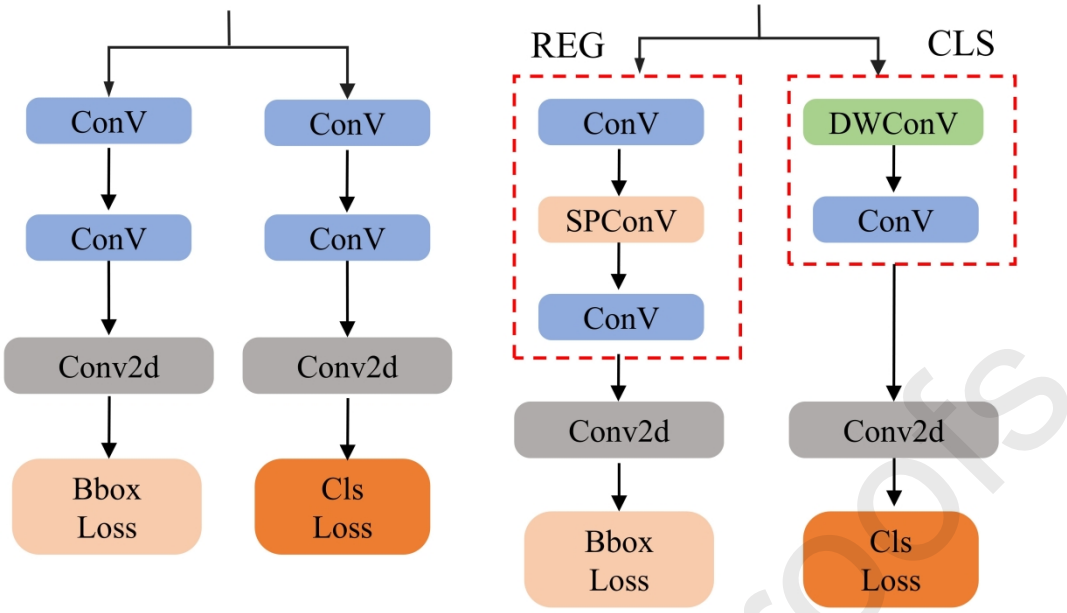


Fig. 5. LiteShiftHead structure. (a) Multi-scale features and output of detection results; (b) Improving the LiteShiftHead network.

2.2.4. dDIOU

The original DIoU[13] introduced a penalty term for the distance between center points in Intersection over Union (IoU), but it did not take into account the impact of shape differences. In this study, we enhance the DIoU by adding penalty terms for both center point distance and shape differences, employing the calculation methods outlined in Eqs (1) to (4). This improvement places greater emphasis on aligning the positions and ensuring the shape consistency of target boxes. By dynamically adjusting the penalty terms and incorporating a width-to-height ratio difference factor, we can better optimize the performance of target detection models.

$$L_{dDIoU} = 1 - IoU + \frac{\rho^2}{c^2} + \alpha ; \quad (1)$$

$$c^2 = \left(\max(w_p, w_t) + \max(h_p, h_t)\right)^2 \quad (2)$$

$$IoU = \frac{|B_p \cap B_t|}{|B_p \cup B_t|} \quad (3)$$

$$\gamma = \frac{|w_p - w_t|}{\max(w_p, w_t)} + \frac{|h_p - h_t|}{\max(h_p, h_t)} \quad (4)$$

where

(1) $\rho^2 = (x_p - x_t)^2 + (y_p - y_t)^2$ denotes the squared Euclidean distance between the centroids of predicted box B_p and ground truth box B_t, with (x_p, y_p) and (x_t, y_t) being their center coordinates.

(2) w_p, h_p and w_t, h_t are the width and height, respectively.

(3) $\alpha = \frac{\gamma}{1+\gamma}$ is a dynamic weight balancing the shape difference penalty.

(4) c is the diagonal length of the smallest enclosing rectangle covering.

The dynamic adjustment of α enables dDIoU to prioritize center alignment for compact targets (e.g., Y-type buds) while emphasizing shape consistency for elongated I-type buds, as validated in Fig. 6c.

2.3. Harvest point estimation

Building on the bounding boxes detected by the YOLO-MEST model, we propose a pipeline based on a morphological algorithm to estimate the picking points of tea buds. The process begins with foreground extraction using the GrabCut algorithm[14], as demonstrated in Fig. 6a. Given an input image I ∈ R^{m×n×3}, we first convert it to grayscale through a standard colour space transformation, denoted as I_gray. The GrabCut algorithm generates a binary mask M, where the automatic rectangular initialization region covers 95% of the image area (with a δ =5-pixel boundary margin). After five iterations of graph-cut optimization for energy function minimization, a refined binary output B ∈ {0,255}^{m×n} is generated, with foreground pixels set to 255, as shown in Fig. 6b.

Lin et al.[15] employed medial axis transform (MAT) for morphological skeletonization of the binary image. For each connected component in B, an 8-connected skeleton S ∈ {0,1}^{m×n} was computed via MAT, as shown in Fig. 6c. A convolution kernel K = [1 1 1; 1 10 1; 1 1 1] was constructed for adaptive intersection detection, where a pixel (x, y) was identified as an intersection if (S×K)(y,x) ⩾ 13 (Fig. 6d).

To obtain the final picking points, the preliminary intersections were further filtered based on two criteria.

(i) If no intersections existed, the picking point was determined by offsetting the lowest skeleton point along the negative image vertical axis by N_1 = 20 pixels.

(ii) If intersections existed, the lowest intersection point was selected as the picking point. However, if this point was located in the upper half of the skeleton, the picking point was instead set to the lowest skeleton point with an additional vertical offset of N_2 = 20 pixels (while retaining its horizontal coordinate).

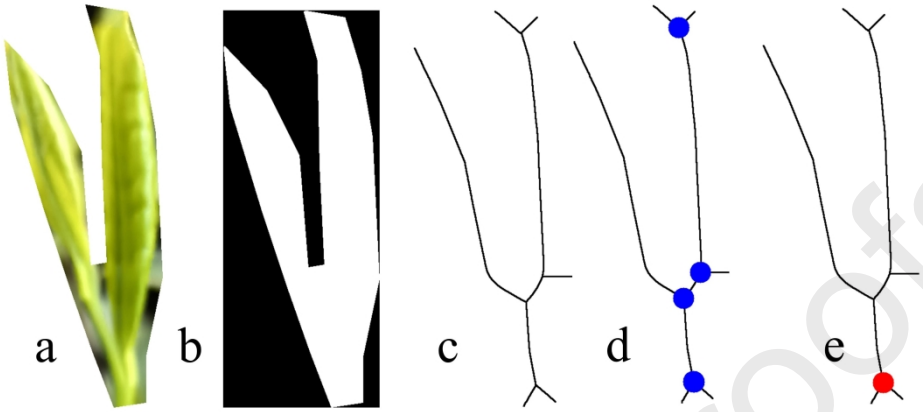


Fig. 6. Harvest point estimation process

2.4. Test procedure

2.4.1. Environment Configuration

This experiment was conducted on a Windows 10 (64-bit) system with an Intel Core i9-11900KF CPU at 3.50 GHz and an RTX 3090i GPU. We used CUDA 11.3.0 for model acceleration and CUDNN 8.2.0 to enhance GPU performance. The deep learning network was built with PyTorch 1.12.1 and trained using Python 3.9.0.

During model training, the number of epochs and batch size were set to 300 and 32, respectively, to enhance training efficiency. The initial learning rate, determined through testing, was set at 0.001. The Sto-chastic Gradient Descent (SGD) optimizer was utilized for learning rate adjustment, with the cosine annealing hyperparameter and learning rate momentum established at 0.01 and 0.947, respectively. Input images were resized to 640×640 pixels, and training lasted for 300 iterations. Hyperparameters were chosen based on standard YOLO practices and refined through a limited grid search on the validation set to optimize mAP50. The final settings were selected for their stability and superior performance.

2.4.2. Evaluation criteria

In object detection tasks, the evaluation metrics**Error! Reference source not found.** typically include accuracy, precision (P), recall (R), average precision (AP), and mean average precision (mAP). Accuracy reflects the degree of match between the detection results and the actual targets and is usually calculated using the IoU value between the predicted and ground truth bounding boxes. A higher IoU value indicates more accurate detection results. The formulas for P and R are as follows:

$$P = \frac{TP}{TP + FP} \quad (5)$$

$$Recall = \frac{TP}{TP + F} \quad (6)$$

Where *FP* stands for false positives, *FN* for false negatives, *TP* for true positives, and *TN* for true negatives.

In object detection, mAP is commonly used to evaluate the performance of detectors. It is expressed as follows:

$$AP = \int_0^1 P(R)dR \quad (7)$$

$$mAP = \frac{1}{K}\sum_{i=1}^{K} AP \quad (8)$$

Where *P* is the probability of correctly predicting tea buds and key points as positive samples. *R* is the probability of correctly identifying positive samples as such. *AP* is a combination of precision and recall, *mAP* is the average AP of different categories, and *K* is the number of categories. In this experiment, there is only one category, so $K = 1$.

3. Results and discussion

All experiments in this subsection are conducted using YOLOv8-n with an input size of 640×640. Our study utilized 4,600 annotated data samples, focusing on tea buds for keypoint annotation and frame labeling. The dataset was divided into three subsets in an 8:1:1 ratio: a training set containing 3,680 samples, a validation set with 460 samples, and a test set with 460 samples.

3.1. Ablation study on combined modules

To validate the synergy of RepNCSELAN4, LiteShiftHead, and dDIoU, we incrementally integrate each module into YOLOv8-n and measure mAP50, FLOPs, and inference speed (FPS). As shown in Table 1, the complete YOLO-MEST achieved a mAP50 of 84.9%, compared to the 83.2% baseline, with only an 8%

increase in computational overhead. This demonstrates the efficiency of our co-design strategy.

Table 1 Ablation study on module combinations.

Configuration	mAP50(%)	Params(M)	FLOPs(G)	FPS
YOLOv8-n (Baseline)	83.2	3.01	8.2	62
+ RepNCSELAN4	83.9	2.57	10.9	58
+ LiteShiftHead	83.6	3.01	8.1	60
+ dDIoU	83.9	3.01	8.2	61
Full YOLO-MEST	84.9	2.58	10.9	55

The ablation study confirms that the performance gains of YOLO-MEST result from both improvements in individual modules and their effective integration. The RepNCSELAN4 module enhances feature representation through reparameterization, boosting mAP50 by 0.7% despite reducing parameters by 33%, albeit with increased FLOPs (32.9%). LiteShiftHead achieves a 0.4% accuracy gain with negligible computational cost, leveraging shift operations for efficient spatial modeling. Furthermore, the dDIoU module contributes an additional 0.7% increase in mAP50 by dynamically optimizing bounding box regression, which aligns with the benefits observed from DIoU-Net for detecting small objects.

The full model has a mAP50 score that is 1.7% higher than YOLOv8-n (84.9% compared to 83.2%), highlighting the effectiveness of co-design. This performance surpasses that of similar lightweight detectors, such as YOLOv5-n, which achieves approximately 82.5%, while still maintaining real-time processing speed at 55 frames per second (FPS). When compared to traditional tea shoot detection methods, like support vector machines combined with histogram of oriented gradients (SVM/HOG), which yield less than 75% mAP50, YOLO-MEST excels in handling complex agricultural scenes.

3.2. Comparison of RepNCSELAN4 with other feature fusion methods

To validate the effectiveness of the proposed RepNCSELAN4 model for tea bud detection, we compared it with five representative feature fusion mechanisms: CBAM [17], Shuffle Attention (SA)[18], SimAM[19], GAM[20], and SK[21]. A self-built, high-quality tea dataset was used for testing, as shown in Table 2. YOLOv8-n was adopted as the base model, with each feature fusion module integrated into the C2f module, with outputs from layers 2 through 5.

Table 2 Comparison results of feature fusion mechanisms.

Model	mAP50 (%)	Parameters (M)	FLOPs (G)	Parameters-to-Baseline (M)
Baseline	83.2	3.01	8.2	-
CBAM	83.4	2.67	7.2	-0.34
Shuffle Attention(SA)	82.3	3.01	8.2	-0.34
SimAM	82.6	3.21	8.8	+0.20
GAM	83.8	3.45	8.5	+0.44
SK	83.6	8.56	12.6	+4.4
Our Proposed c2f-RepNCSELAN	83.9	2.58	10.9	-0.43

The RepNCSELAN4 attention mechanism presented in this paper demonstrates significant advantages in terms of model performance and parameter efficiency, as shown in Table 1. When compared to other mainstream attention mechanisms, RepNCSELAN4 reduces the number of parameters by 0.43 million while maintaining excellent detection performance and achieving a lighter network structure. Specifically, RepNCSELAN4 achieves a mean Average Precision (mAP50) of 83.9%, compared to the baseline of 83.2%, with 14.3% fewer parameters (2.58 million vs. 3.01 million) than the baseline model. This improvement is largely attributed to the C2f-RepNCSELAN module, which effectively enhances the interaction learning between channel and spatial dimensions, enabling the model to comprehensively understand feature information.

Furthermore, this improved attention mechanism allows the model to capture long-term dependencies in input sequences, which is crucial for object detection tasks in complex scenarios. The combined benefits of these features significantly enhance the model's capability for feature extraction, resulting in more accurate detections while maintaining a high inference speed. These characteristics make RepNCSELAN4 particularly well-suited for embedded devices with limited computing resources or for real-time detection applications.

3.3. Improvements to LiteShiftHead compared to other detection heads

To validate the effectiveness of the proposed improvement to LiteShiftHead for tea bud detection, several representative feature fusion mechanisms were integrated with it, including DynamicHead[Error! Reference source not found.] and LADH[Error! Reference source not found.]. Based on a self-built high-quality tea dataset, comparative tests were conducted using YOLOv8-n as the base model, and the results are shown in Table 3.

Table 3 Comparison results of detection heads

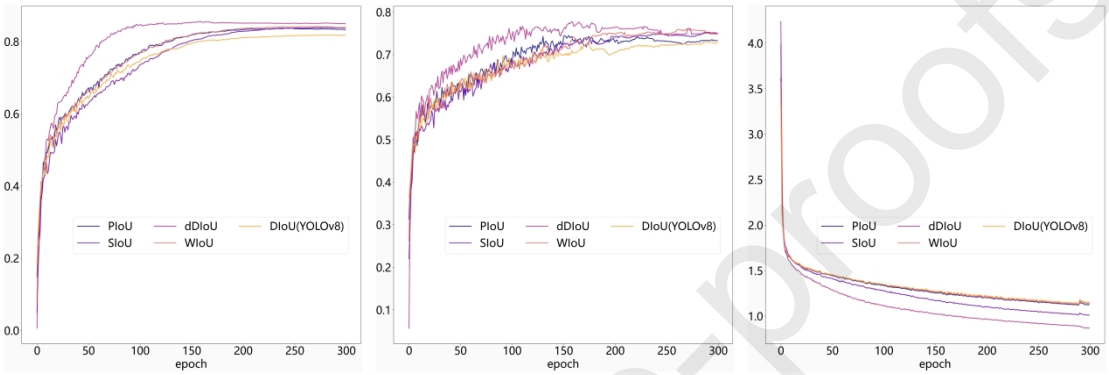
Head	mAP50 (%)	Parameters (M)	FLOPs (G)	mAP50-to-Baseline (M)
Baseline	83.2	3.01	8.2	-
DynamicHead	83.8	3.01	8.2	+0.6
LADH	83.1	3.01	8.2	-0.1
LiteShiftHead	83.9	3.01	8.1	+0.7

As shown in Table 2, the improved LiteShiftHead achieves a mAP50 score of 83.9% for the object detection model, representing a 1% improvement over the DynamicHead scheme. This enhancement is a result of the decoupling of the detection head structure. The object localization branch utilizes spatial pyramid convolutions (SPConv) to integrate contextual information from various receptive fields through parallel, multi-scale convolution kernels. This significantly improves the model's ability to detect targets of different sizes. Meanwhile, the category classification branch employs depth-separable convolution (DWConv), which combines channel decoupling with point-by-point fusion mechanisms. This approach maintains feature discrimination while reducing theoretical computational complexity. Overall, this task-specific, modular design strategy enables the detection head to effectively balance feature expression capabilities with computational efficiency.

3.4. Comparison of different loss functions

In the default configuration of the original YOLOv8 architecture, the bounding box regression task employs an intersection over union (IoU) loss function. While IoU exhibits robust geometric matching capabilities in traditional settings through center point distance penalties and aspect ratio consistency constraints, its mathematical modeling exhibits two common limitations: First, the rectangular constraint mechanism based on fixed weight coefficients makes it challenging to adapt to the

shapes of targets with irregular contours, such as ring-shaped parts and multi-branched tree branches. Second, the quadratic term of the center point distance amplifies minor localization errors (less than 5 pixels), which limits the model's localization accuracy in complex scenes. The proposed dDIoU loss function demonstrates superior performance to DIoU, WIoU, SIoU, and PIoU loss functions in feature regression tasks, specifically in terms of faster convergence speed. Fig. 7 illustrates the training curves for the recall rate and mean average precision (mAP) metrics, as well as the changes in loss values during training, for each loss function. As can be seen, the proposed dDIoU loss function is more stable and converges faster.



(a) Recall (b) mAP (c) Train/loss

Fig. 7. Function curves of recall rate and mAP as a function of training cycle under different loss functions.

We evaluated the computational efficiency of the proposed method by comparing the model parameters and computational complexity of models that incorporate DIoU, WIoU, SIoU, PIoU, and dDIoU loss functions on our self-built dataset (Table 4). Computational complexity is quantified in terms of floating-point operations per second (FLOPs).

Table 4 Loss function test results

IoU	mAP50 (%)	Parameters (M)	FLOPs (G)	mAP-to-Baseline (%)
DIoU（YOLOv8）	83.1	3.01	8.2	-
WIoU	83.3	3.01	8.2	+0.2
SIoU	83.1	3.01	8.2	0

PIoU	83.9	8.95	13.4	+0.6
dDIoU	83.9	3.01	8.2	+0.6

The introduction of the dDIoU loss function, as shown in Table 3, enables the model to achieve the best overall performance compared to other loss functions. The model records a mean Average Precision (mAP50) of 83.9%, which is 0.8%, 0.6%, and 0.8% higher than the results achieved using DIoU (YOLOv8)[24], WIoU[25], and SIoU[26], respectively. Notably, this performance is on par with the mAP50 of PIoU[27], despite the model containing 5.94 million fewer parameters. This suggests that the dDIoU loss function effectively emphasizes the distance and shape features of bounding boxes within complex foregrounds and backgrounds. It reduces the interference caused by occlusions in the background, thereby enhancing the accuracy of matching between predicted and ground-truth bounding boxes.

3.5. Feature detection inference and evaluation

To demonstrate the effectiveness of the proposed module intuitively, we utilize XGrad-CAM to visualize its performance, as shown in Fig. 8. The YOLOv8n-CBAM, YOLOv8-RepNCSELAN4, YOLOv8-dDIoU, and YOLOv8-LiteShiftHead models are improved versions of the CBAM, RepNCSELAN4, dDIoU, and LiteShiftHead models, respectively, embedded into the YOLOv8-n framework.

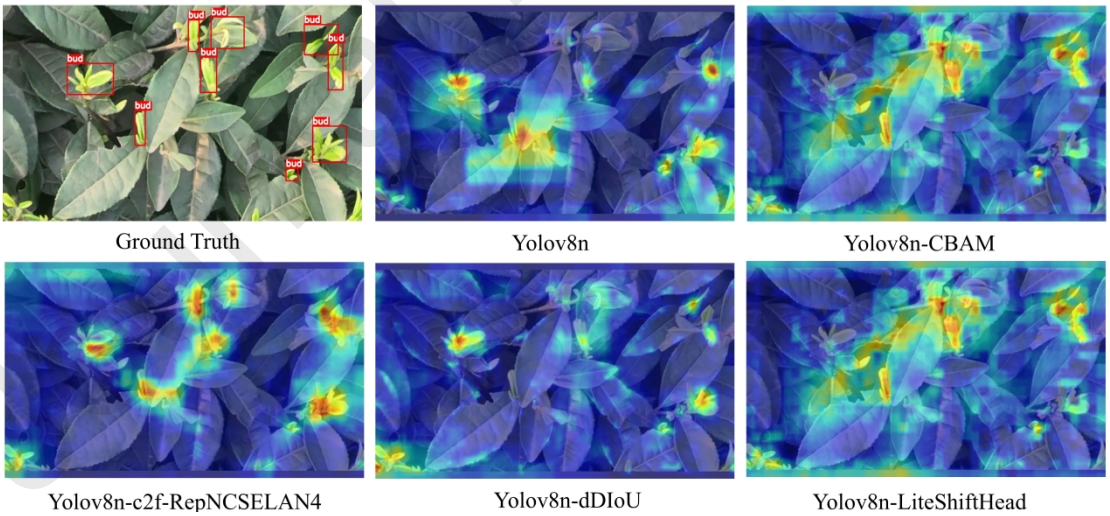


Fig. 8. Comparison of XGrad-CAM feature maps.

The visualization results show that each module performs exceptionally well in object localization, offering more continuous and comprehensive attention to object regions. This leads to enhanced feature representation capabilities. When compared to the baseline model and typical CBAM feature fusion modules, the upgraded RepNCSELAN4 module demonstrates superior detection performance. Given that tea

buds are small, dense targets, the RepNCSELAN4 module effectively fuses low-level and high-level features of tea buds, thereby improving their representation. Furthermore, the improved RepNCSELAN4 module reduces false positives and missed detections. Additionally, integrating LiteShiftHead enhances the features of foreground objects while reducing background noise, resulting in better overall detection performance.

3.6. Model comparison analysis

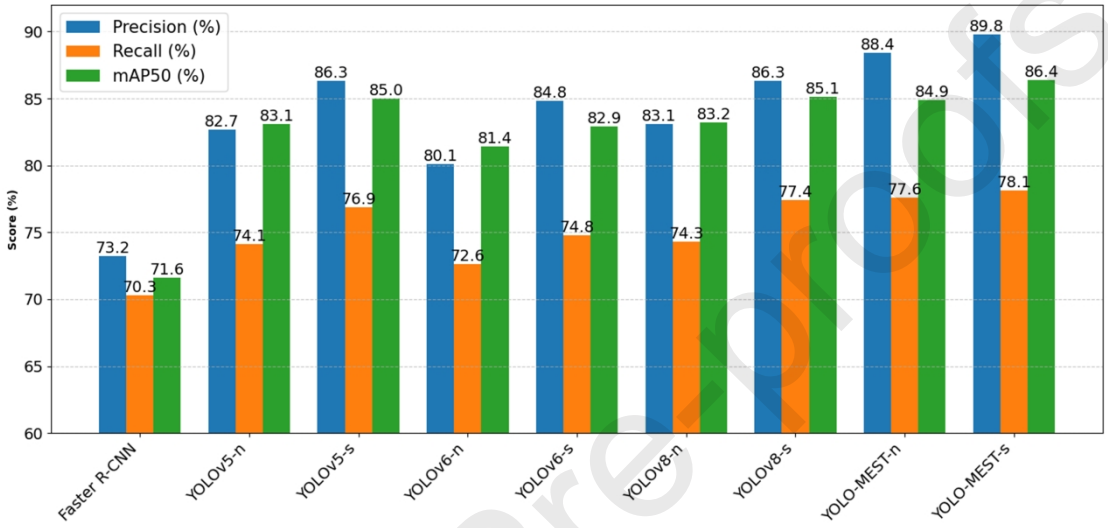


Fig. 9. Comparison of detection results for various models.

This experiment aimed to validate the technical advantages of the improved YOLO-MEST model for detecting tea buds. An automated data collection system was established, and a professional test dataset consisting of 460 high-resolution images of fresh tea leaves (4096 × 2160 pixels) was created. In a consistent deployment environment on the NVIDIA Jetson Orin NX edge computing platform, performance comparison experiments were conducted among Faster R-CNN, YOLOv5, YOLOv6, YOLOv8, and YOLO-MEST to assess their effectiveness in detecting high-quality tea buds. The results are presented in Fig. 9.

As shown in Fig. 9, YOLO-MEST-s achieved a mean Average Precision at mAP50 of 86.4% at 58 frames per second (FPS), surpassing YOLOv8-s, which recorded a mAP50 of 85.1% at 62 FPS in terms of accuracy. Additionally, the detection accuracy of YOLO-MEST-n exceeded that of YOLOv5-n, YOLOv6-n, and YOLOv8-n[28]. Our dataset indicated a mAP50 of 84.9%, which is 1.7% higher than YOLOv8-n. This data clearly demonstrates that the YOLO-MEST model significantly enhances detection accuracy, achieving state-of-the-art performance.

In contrast, the two-stage detection model, Faster R-CNN, yielded a mAP50 of only 71.6% with 41.26 million parameters, making it the least effective model in terms of overall performance and unsuitable for the tea bud detection task in this study. The YOLO series models are inherently more suited for object detection tasks due to their

innovative architectural design and optimizations. We incorporated enhancements such as RepNCSELAN4, LiteShiftHead, and dDIoU, which improve the algorithm's target detection capabilities in complex backgrounds. Overall, YOLO-MEST-s delivered the best performance across all evaluation metrics.

3.7. Adaptability test

To evaluate YOLO-MEST-n's adaptability to unknown environments and generalization capability, this study collected an additional 50 tea leaf images from complex agricultural scenes at the Teaching Demonstration Base of Anhui Agricultural University in Hefei City, Anhui Province. These scenes included factors such as lighting, background interference, leaf occlusion, individual leaf occlusion, and variations in leaf size. The images were used to compare the model's performance. Fig. 10 shows the partial detection results of YOLO-MEST-n compared with YOLOv5-n, YOLOv6-n, and YOLOv8-n.

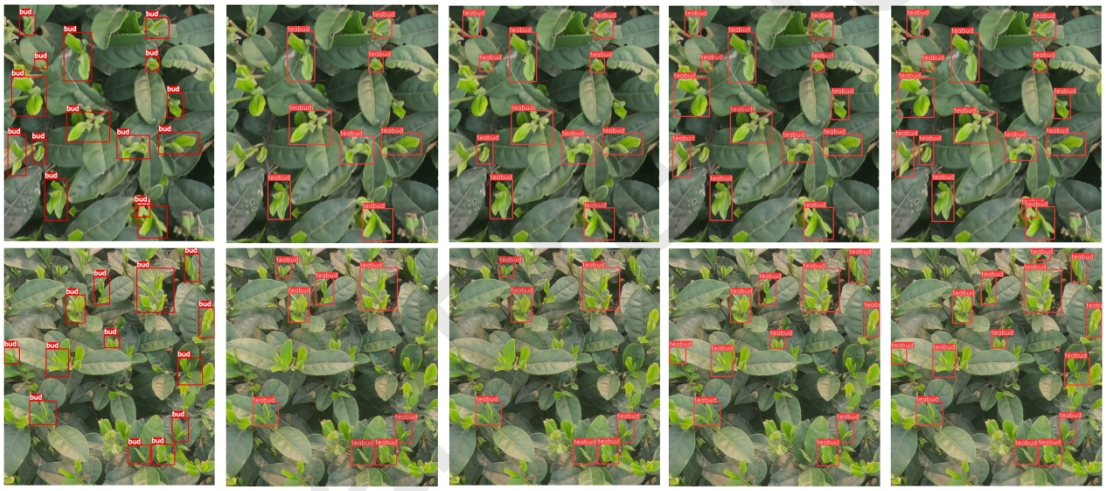


Fig.10. Visualization of detection results using different methods

As Fig. 10 shows, YOLOv5-n, YOLOv6-n, and YOLOv8-n all have false negatives. In contrast, YOLO-MEST-n provides stable detection performance regardless of background interference complexity, target size diversity, or lighting condition variability. These results demonstrate the model's flexibility and adaptability in unknown, complex agricultural scenarios, further validating its high accuracy, robustness, and excellent generalization capabilities.

3.8. Harvest point estimation

Building on the aforementioned tea bud detection algorithm (which is based on bounding box detection), this paper presents a new approach to detecting tea bud picking points. It is important to note that the keypoint annotations mentioned were used solely for the evaluation of this picking point estimation method and were not part of the training or evaluation of the core YOLO-MEST object detection model.

This method uses GrabCut segmentation, central axis skeletonization, and adaptive picking point detection technology to identify tea bud picking points. First, the YOLO-MEST algorithm detects the tea buds and generates cropped bounding boxes for the images, making the targets more focused. Then, the GrabCut algorithm [29] is used for segmentation to obtain the tea bud segmentation region effectively. Next, the binary image undergoes a central axis transformation, and the tea bud is skeletonized. This process simplifies the object into a skeleton with a width of a single pixel while preserving its topological structure. Finally, a 3×3 convolution kernel is applied to the skeleton to detect branch points. Picking points are determined based on the growth characteristics of the tea buds. Fig. 11 shows the results of the algorithm, in which the segmented tea buds exhibit Y-shaped and I-shaped structures with varying orientations. The proposed sampling point estimation algorithm effectively extracts the tea bud sampling points.

Fig.11. Harvesting point detection.

To validate the effectiveness of the proposed tea bud picking point detection algorithm, we trained and inferred the cropped tea bud dataset using the YOLOv8-pose[30] network. The dataset comprised 3,656 images, which were divided into training, validation, and test sets in an 8:1:1 ratio. Annotation was performed using the open-source software LabelMe, with four key points labeled for each instance: the apex of the one-bud-one-leaf and the base of the petiole (two points each). Examples of the annotated tea bud key points are illustrated in Figs 12a–d.

The trained YOLOv8-pose weights were applied to infer the test samples, with the results depicted in Fig. 12e-h. A noticeable discrepancy was observed between the YOLOv8-pose inference results and the ground truth annotations. Specifically, while Fig. 12e and Fig. 12f successfully detected all four key points, their positions significantly deviated from the actual values. In contrast, Figs. 12 g and 12 h exhibited missed detections of key points, along with pronounced positional inaccuracies in the detected points. There may be two reasons for this.

(1) Tea bud keypoints (e.g., leaf apex and petiole base) are inherently ambiguous due to their small size and irregular shapes. Slight inconsistencies in manual labeling (Fig. 12a-d) may confuse the model during training.

(2) This dataset (3,656 images) may not be sufficient to fully capture the variability in the appearance of tea buds under different lighting conditions, growth stages, or occlusion scenarios. Smaller datasets often result in poor generalization capabilities, especially in fine-grained tasks such as keypoint localization. However, increasing the dataset size means more work, which is not conducive to further deployment of the algorithm.

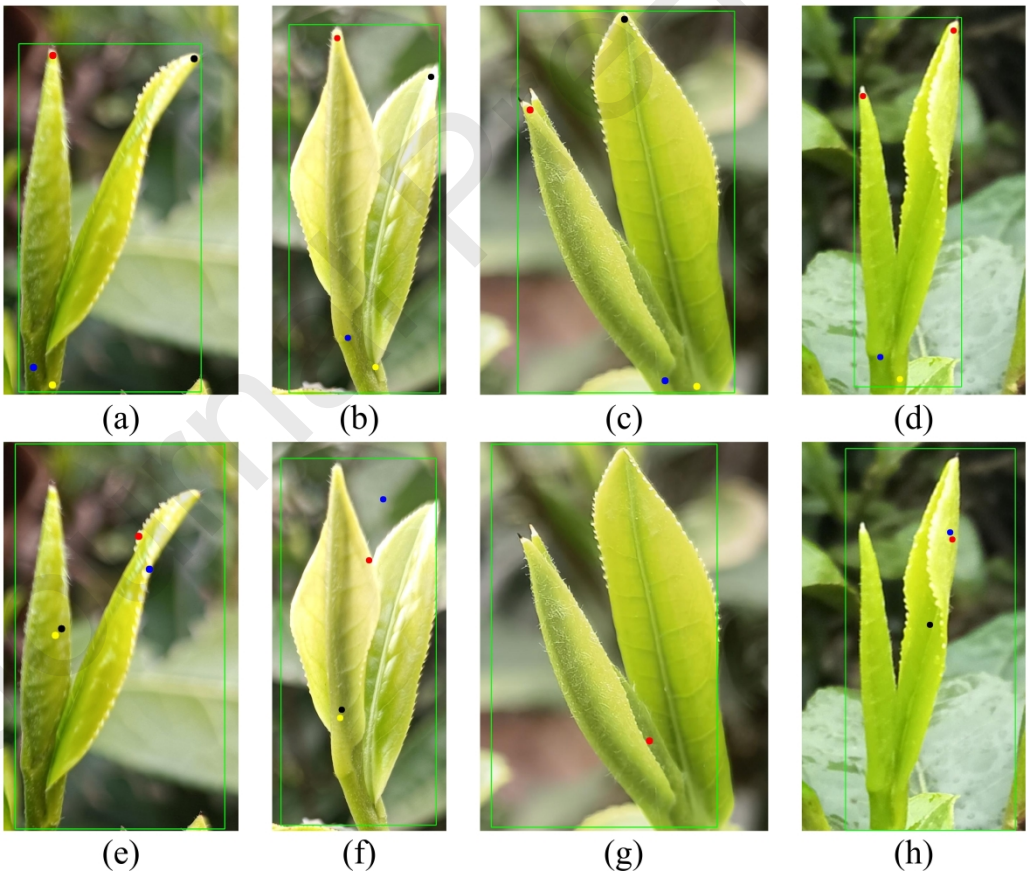


Fig. 12. Comparison between Ground truth and Yolov8-pose detection of key points (a) to (d) Ground truth (e) to (h) Yolov8-pose detection.

4. Conclusion

To accurately identify tea buds in complex agricultural environments, this study proposes a significantly modified object detection model, YOLO-MEST, which is developed based on YOLOv8n. The core of this work is a combined network architecture comprising RepNCSELAN4+LiteShiftHead and dDIOU. The YOLOv8n model was optimized for tea bud detection in complex scenarios, resulting in the YOLO-MEST object detection model. The following conclusions were drawn:

(1) Fusion of low- and high-level features improves tea bud detection accuracy. The color and texture information in low-level features enhances the network's capability for feature representation.

(2) The RepNCSELAN4+LiteShiftHead+dDIOU combination strategy outperforms other attention mechanisms in terms of detection performance. By embedding RepNCSELAN4 in the output of each layer of the C2f module and introducing the dDIOU detection box deformation penalty function at the model's output layer, the model suppresses complex background noise while strengthening important channel and spatial information, thereby improving detection accuracy.

(3) The RepNCSELAN4 replaces traditional convolutions in the C2f module, thereby expanding the convolutional receptive field and enhancing the network's ability to capture contextual information. This helps address false positives and missed detections for small tea buds.

(4) The proposed morphology-based picking point estimation algorithm serves as an effective post-processing step that operates on the detection results. This approach improves the system's applicability to tea bud picking point acquisition without requiring additional deep learning models for keypoint detection, thereby reducing dependency on large, annotated datasets and enhancing practicality for real-world deployment.

Through extensive experimentation and analysis of visualization results, we found that the proposed feature enhancement strategy effectively improves the network's ability to represent features, enabling accurate localization of tea bud targets. The experimental results show that the combined strategy model achieves an improved detection accuracy of 84.49% mAP on the self-built data set, which is 1.7% higher than that of existing detection methods. The proposed YOLO-MEST model, along with the associated picking point estimation algorithm, marks a significant advancement towards fully automated tea harvesting systems. Its balance of accuracy and speed makes it a strong candidate for integration into embedded systems on harvesting robots. This technology could help alleviate labor shortages, reduce production costs, and enhance the consistency of tea quality by enabling selective picking of high-value buds.

First, the model's computational complexity, at 10.9 GFLOPs, may still pose challenges for deployment on ultra-low-power edge devices. Future work will focus on exploring model pruning and quantization techniques to improve deployment efficiency. Second, the current model was trained and tested on a dataset from a specific region and season. Its performance under extreme weather conditions—such as heavy rain or fog—and across different tea cultivars requires further investigation. Finally, integrating this technology into real-world applications will involve challenges, including real-time processing on moving platforms and mechanical alignment, which are beyond the scope of this paper but will be the focus of our subsequent applied research.

Data availability

The data that support the findings of this study are available from the corresponding author upon reasonable request.

CRediT authorship contribution statement

Chuanyang Yu: Data curation, Formal Analysis, Methodology, Software, Writing-original draft, Conceptualization. **Yi Xue:** Data curation, Methodology, Supervision, Writing-review & editing. **Liuyang Zhang:** Validation, Data curation, Formal Analysis. **Xue An:** Supervision, Investigation, Writing-review & editing. **Ce Liu:** Validation, Data curation, Software, Writing-review & editing. **Liqing Chen:** Conceptualization, Formal Analysis, Funding acquisition, Methodology, Supervision, Project administration, Writing-review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This research was financially supported by State Key Laboratory of Tea Biology and Resource Utilization (Grant No. SKLTOF20230123; Project Holder: Ce Liu).

References

[1] Zhao X, He L, Li Y, Chen J, Wu C. Kinetostatic modeling of clamping force in a tendon-driven soft robotic gripper for tea shoot plucking. Comput Electron Agric 2025;236:110441. https://doi.org/10.1016/j.compag.2025.110441.

[2] Lin G, Xiong J, Zhao R, Li X, Hu H, Zhu L, Zhang R. Efficient detection and picking sequence planning of tea buds in a high-density canopy. Comput Electron Agric 2023;213:108213. https://doi.org/10.1016/j.compag.2023.108213.

[3] Xie D, Chen L, Liu L, Chen L, Wang H. Actuators and sensors for applic ation in agricultural robots: A review. Machines 2022;10(10):913. https://do i.org/10.3390/machines10100913.

[4] 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. https://doi.org/10.1016/j.compag.2021.106108.

[5] Zhang L, Zhang H, Chen Y, Dai S, Li X, Kenji I, Li M. Real-time monitoring of optimum timing for harvesting fresh tea leaves based on machine vision. Int J Agric Biol Eng 2019;12(1):6-9. https://doi.org/10.25165/j.ijabe.20191201.3418.

[6] Ren S, He K, Girshick R, Sun J. Faster r-cnn: Towards real-time object detection with region proposal networks. IEEE Trans Pattern Anal Mach Intell 2016;39(6):1137-1149. https://doi.org/10.1109/TPAMI.2016.2577031.

[7] 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. Journal of King Saud University-Computer and Information Sciences 2024;36(10):102262. https://doi.org/10.1016/j.jksuci.2024.102262.

[8] Chen J, Chen Y, Jin X, Che J, Gao F, Li N. Research on a parallel robot for tea flushes plucking. In: 2015 International Conference on Education, Management, Information and Medicine. 2015, p. 22-26. https://doi.org/10.2 991/emim-15.2015.5.

[9] Chen C, Lu J, Zhou M, 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. https://doi.org/10.1016/j.compag.2022.107116.

[10] Li J, Li J, Zhao X, Su X, Wu W. Lightweight detection networks for tea bud on complex agricultural environment via improved YOLO v4. Comput Electron Agric 2023;211:107955. https://doi.org/10.1016/j.compag.2023.107955.

[11] Gui Z, Chen J, Li Y, Chen Z, Wu C, Dong C. A lightweight tea bud detection model based on Yolov5. Comput Electron Agric 2023;205:107636. https://doi.org/10.1016/j.compag.2023.107636.

[12] Varghese R, Sambath M. Yolov8: A novel object detection algorithm with enhanced performance and robustness. In: 2024 International Conference on Advances in Data Engineering and Intelligent Computing Systems (ADICS). IEEE; 2024, p. 1-6. https://doi.org/10.1109/ADICS58448.2024.10533619.

[13] Wu Y, Zhu Y, Guo J, Yu Y. YOLOv5 detection algorithm of steel Defects based on introducing light convolution network and DIOU function. In: 2023 IEEE 12th Data Driven Control and Learning Systems Conference (DDCLS), Xiangtan, China; 2023, p. 118-122. doi:10.1109/DDCLS58216.2023.10165997.

[14] Boykov Y, Jolly M P. Interactive organ segmentation using graph cuts. In: International conference on medical image computing and computer-assisted intervention. Berlin, Heidelberg: Springer Berlin Heidelberg; 2000, p. 276-286. https://doi.org/10.1007/978-3-540-40899-4_28.

[15] Lin C, Liu L, Li C, Kobbelt L, Wang B, Xin S, Wang W. Seg-mat: 3d shape segmentation using medial axis transform. IEEE Trans Visual Comput. Graphics 2020;28(6):2430-2444. https://doi.org/10.1109/TVCG.2020.3032566.

[16] Li S, Meng J, Lv D, Chen X, Liu J. Defect Detection Method Based on Improved YOLOv5s Model. In: 2023 2nd International Conference on 3D Immersion, Interaction and Multi-sensory Experiences (ICDIIME). IEEE; 2023, p. 331-336. https://doi.org/10.1109/ICDIIME59043.2023.00070.

[17] Qiao S, Chen L C, Yuille A. DetectoRS: Detecting objects with recursive feature pyramid and switchable atrous convolution. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021, p. 10213-10224. https://doi.org/10.1109/CVPR46437.2021.01008.

[18] Norkobil Saydirasulovich S, Abdusalomov A, Jamil M K, Nasimov R, Kozhamzharova D, Cho Y I. A YOLOv6-based improved fire detection approach for smart city environments. Sensors 2023;23(6):3161. https://doi.org/10.3390/s23063161.

[19] Yang D, Huang Z, Zheng C, Chen H, Jiang X. Detecting tea shoots usin g improved YOLOv8n. Transactions of the Chinese Society of Agricultural Engineering 2024;40(12):165-173. https://doi.org/10.11975/j.issn.1002-6819.202401155.

[20] Wang Y, Xiao M, Wang S, Jiang Q, Wang X, Zhang Y. Detection of famous tea buds based on improved YOLOv7 network. Agriculture 2023;13(6):1190. https://doi.org/10.3390/agriculture13061190.

[21] Zhang Q L, Yang Y B. Sa-net: Shuffle attention for deep convolutional neural networks. In: ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE; 2021, p. 2235-2239. https://doi.org/10.1109/ICASSP39728.2021.9414568.

[22] Dai X, Chen Y, Xiao B, Chen D, Liu M, Yuan L, Zhang L. Dynamic head: unifying object detection heads with attentions. In: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA; 2021, p. 7369-7378, doi: 10.1109/CVPR46437.2021.00729.

[23] Guo Y, Shen Q. Zhang S, Zhang C, Wang X. An airborne target recognition model based on SPD, PConv and LADH detection heads. In: International

Conference on Autonomous Unmanned Systems. Singapore, Springer Nature Singapore; 2023, p. 325-337. https://doi.org/10.1007/978-981-97-1087-4_31.

[24] Zheng Z, Wang P, Liu W, Li J, Ye R, Ren D. Distance-IoU loss: Faster and better learning for bounding box regression. In: Proceedings of the A AAI conference on artificial intelligence. 2020;34(07):12993-13000. https://doi.org/10.1609/aaai.v34i07.6999.

[25] Dai X, Chen Y, Xiao B, Chen D, Liu M, Yuan L Zhang L. Dynamic head: Unifying object detection heads with attentions. In: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2021, p. 7373-7382. https://doi.org/10.48550/arXiv.2106.08322.

[26] Zhang J, Chen Z, Yan G, Wang Y, Hu B. Faster and lightweight: an improved YOLOv5 object detector for remote sensing images. Remote Sens 2023;15(20):4974. https://doi.org/10.3390/rs15204974.

[27] Liu C, Wang K, Li Q, Zhao F, Zhao K, Ma H. Powerful-IoU: More strai ghtforward and faster bounding box regression loss with a nonmonotonic f ocusing mechanism. Neural Networks 2024;170:276-284. https://doi.org/10.1 016/j.neunet.2023.11.041.

[28] Xiao S, Zhao Q, Chen Y, Li T. A dual-backbone lightweight detection and depth position picking system for multiple occlusions Camellia oleifera fruit. Comput Electron Agric 2025;233:110157. https://doi.org/10.1016/j.compag.2025.110157.

[29] Li Y, Zhang J, Gao P, Jiang L, Chen M. Grab cut image segmentation b ased on image region. In 2018 IEEE 3rd international conference on imag e, vision and computing (ICIVC). IEEE; 2018, p. 311-315. https://doi.org/1 0.1109/ICIVC.2018.8492818.

[30] Dong C, Tang Y, Zhang L. HDA-pose: a real-time 2D human pose estimation method based on modified YOLOv8. Signal, Image and Video Processing. 2024;18(8):5823-5839. https://doi.org/10.1007/s11760-024-03274-2.

■ **Ethics Statement**

☑ Not applicable: This manuscript does not include human or animal research.

☐ If this manuscript involves research on animals or humans, it is imperative to disclose all approval details.

If Yes, please provide your text here:

[31]

■ Authors' Biographies

Chuanyang Yu (chuanyang@ahjzu.edu.cn) is currently a Ph.D. candidate at the College of Engineering, Anhui Agricultural University. He received his B.Eng. (2011) and M.Eng. (2014) degrees in Vehicle Engineering from Anhui Agricultural University. His research focuses on intelligent agricultural machinery and smart farming equipment, particularly exploring the applications of vehicle engineering in modern agriculture.

Yi Xue (xueyi123@stu.ahau.edu.cn) is currently pursuing his Ph.D. degree at the College of Engineering, Anhui Agricultural University. He obtained his Master's degree in Agricultural Engineering from Anhui Agricultural University in 2025 and his Bachelor's degree in Mechanical Engineering from the Anhui Agricultural University College of Economics and Technology in 2020. His research primarily focuses on visual image processing and its applications in agricultural engineering.

Liuyang Zhang (liuyangzhang@stu.ahau.edu.cn) is currently a Ph.D. candidate at the College of Engineering, Anhui Agricultural University. He received his Bachelor's degree in Vehicle Engineering from Anhui Agricultural University in 2023. His research focuses on intelligent agricultural machinery, with a particular interest in the development and optimization of smart farming equipment.

Xue An (anxue@ahau.edu.cn) is an Associate Professor and master's supervisor at the Department of Agricultural Machinery, School of Engineering, Anhui Agricultural University. She received her bachelor's degree in Agricultural Mechanization and Automation (2015) and her master's degree in Agricultural Engineering (2018) from Henan Agricultural University. She obtained her Ph.D. in Agricultural Mechanization Engineering from the College of Mechanical and Electronic Engineering, Northwest A&F University in 2023. From March 2022 to February 2023, she was a CSC-funded joint Ph.D. researcher at the Department of Precision Horticulture, Leibniz Institute for Agricultural Engineering and Bioeconomy (ATB), Germany. From March to September 2023, she continued her research as an assistant researcher in the Department of Mechatronics in Agriculture at ATB. Dr. An's research interests include intelligent agricultural machinery, equipment for protected horticulture, and smart agricultural systems.

Ce Liu (liuce@ahau.edu.cn) is a Lecturer at the School of Engineering, Anhui Agricultural University. He received his bachelor's degree in Vehicle Engineering from Anhui Agricultural University in 2014, and his master's and doctoral degrees in Mechanical Manufacturing and Automation (2017) and Mechanical Design and Theory (2021), respectively, from Hefei University of Technology. He holds a Ph.D. in Engineering and is a council member of the Anhui Society for Vibration Engineering.

His research focuses on intelligent agricultural harvesting machinery, structural vibration, and noise control. He has published more than ten SCI/EI-indexed papers as the first or corresponding author and has applied for ten invention patents and one software copyright.

Liqing Chen (lqchen@ahau.edu.cn). He received the B.E. degree, the M.S. and Ph.D. degrees in automotive engineering from Hefei University of Technology, China, in 2000, 2005, and 2015, respectively. He is a Professor and doctoral supervisor at the School of Engineering, Anhui Agricultural University. He currently serves as the Director of the Anhui Provincial Engineering Technology Center for Intelligent Agricultural Machinery. He is also a member of the Youth Talent Support Program of the Ministry of Agriculture and Rural Affairs of China, a post scientist in the National Wheat Industry Technology System, and a recipient of

742 the State Council Special Government Allowance. In 2024, he was listed among the world's
743 top 2% of scientists.

744

745

746 [32]