

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>

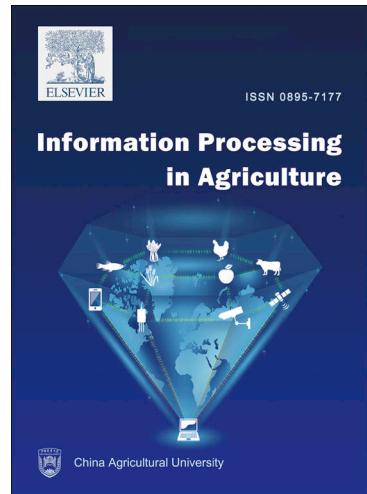
Reference: INPA 485

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>

This is a PDF file of an article that has undergone enhancements after acceptance, such as the addition of a cover page and metadata, and formatting for readability, but it is not yet the definitive version of record. This version will undergo additional copyediting, typesetting and review before it is published in its final form, but we are providing this version to give early visibility of the article. Please note that, during the production process, errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

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,*}

^aCollege 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.

8

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.

28

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

32 **1. Introduction**

33 Tea is regarded as one of the most important beverages in the world. As of 2020,
 34 the global tea cultivation area was approximately 5 million hectares, with China's tea
 35 cultivation area reaching approximately 3.27 million hectares, thus ranking first in the
 36 world in terms of production. Among these, high-quality teas hold the highest
 37 economic value, contributing as much as 75% of the total market value of tea^[1].
 38 Currently, tea picking is predominantly carried out through manual labour, a method
 39 that faces substantial challenges, including high labour costs and reduced productivity.
 40 Moreover, conventional mechanical harvesting often results in the damage of tea buds,
 41 thereby severely compromising tea quality and leading to substantial economic losses
 42 [2]. The adoption of intelligent harvesting machine technology is emerging as a trend,
 43 and the precise identification of tea buds is a core technology for achieving intelligent
 44 harvesting. The precise identification of tea buds has become one of the most
 45 challenging technical problems in the field of intelligent product identification.
 46 Tender buds exhibit specific characteristics that make them challenging to identify.
 47 They are typically small in size, ranging from 10 to 35 mm, and often blend in with
 48 their surroundings, such as older leaves and branches. Moreover, they can block each
 49 other, making it even more challenging to see. These factors are further complicated
 50 by the complex and variable lighting conditions found in tea gardens.

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

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

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

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

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

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

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

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

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

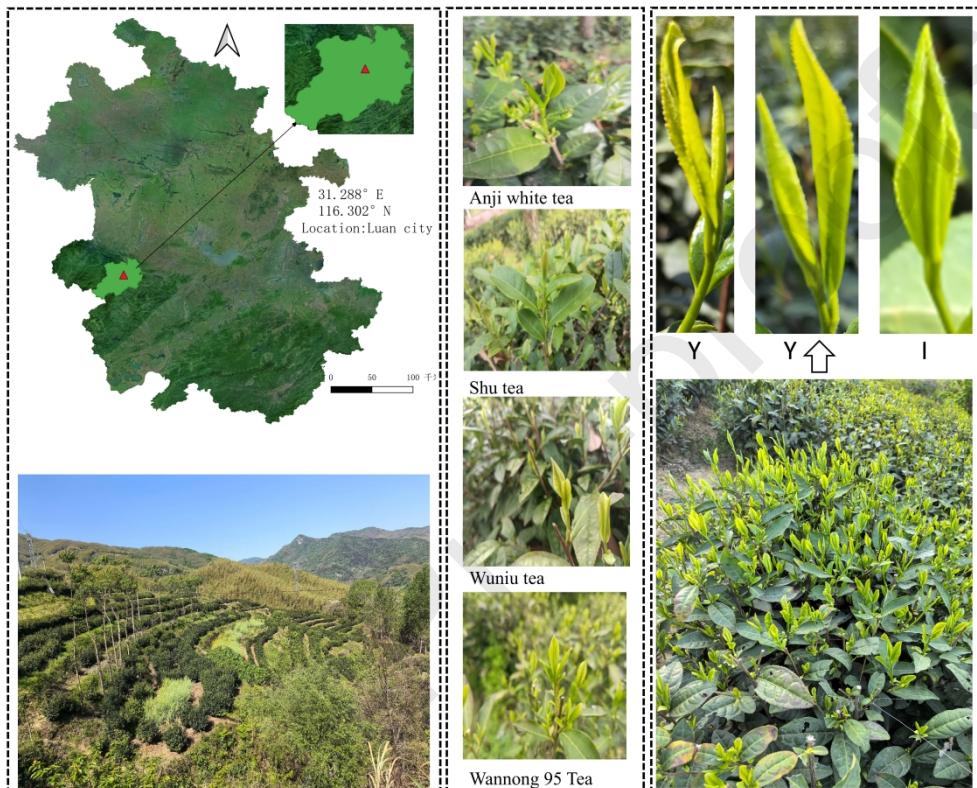
102

103 2. Materials and methods

104 2.1. Teas database building

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

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

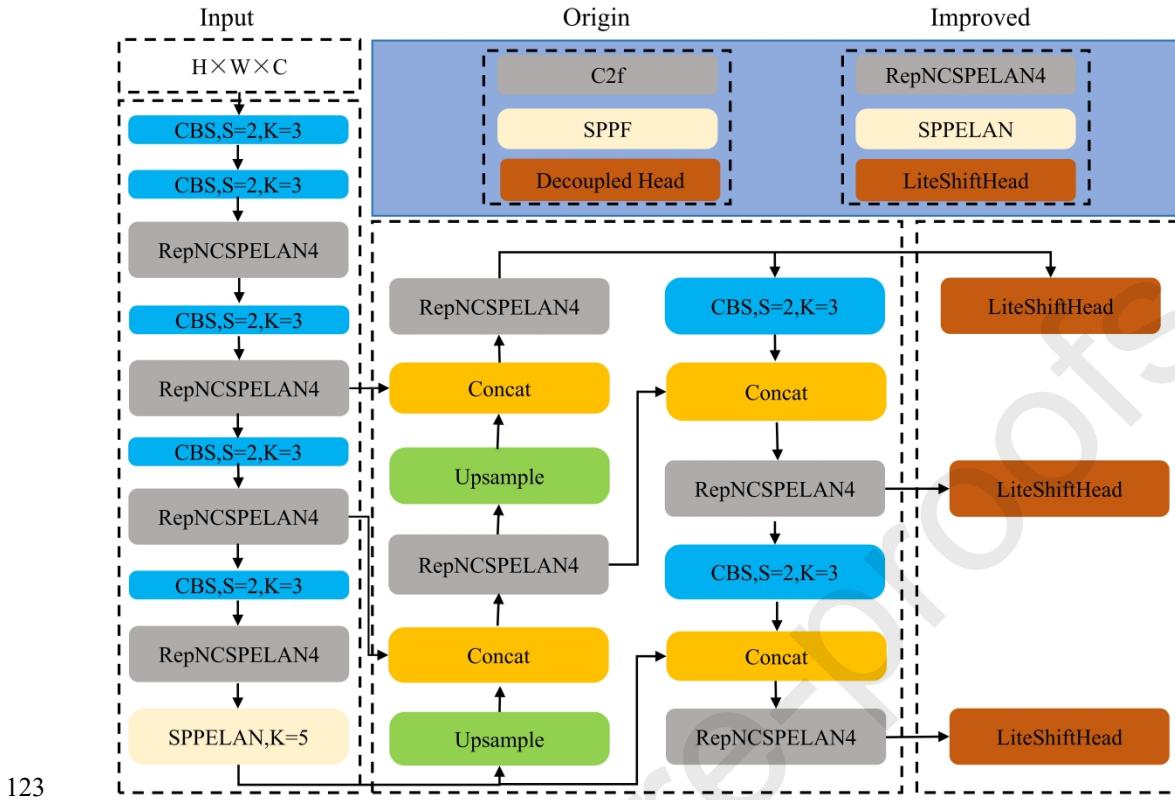


120

121

Fig. 1. Composition of the tea bud dataset.

122 2.2. YOLO-MEST model

124 **Fig. 2.** Improved YOLO-MEST model.

125 To enhance target detection performance in complex agricultural environments,
 126 this study uses YOLOv8 as the baseline model and introduces the improved
 127 YOLO-MEST architecture. The main advancements include: (1) the reconstruction of
 128 the feature extraction network with the RepNCSELAN4 module; (2) the design of a
 129 lightweight, decoupled detection head called LiteShiftHead; and (3) the introduction
 130 of a dynamic distance intersection over union (dDIoU) loss function. As shown in Fig.
 131 2, this architecture aims to optimize both feature representation capabilities and
 132 computational efficiency through multidimensional improvements.

133 **2.2.1. RepNCSELAN4 module**

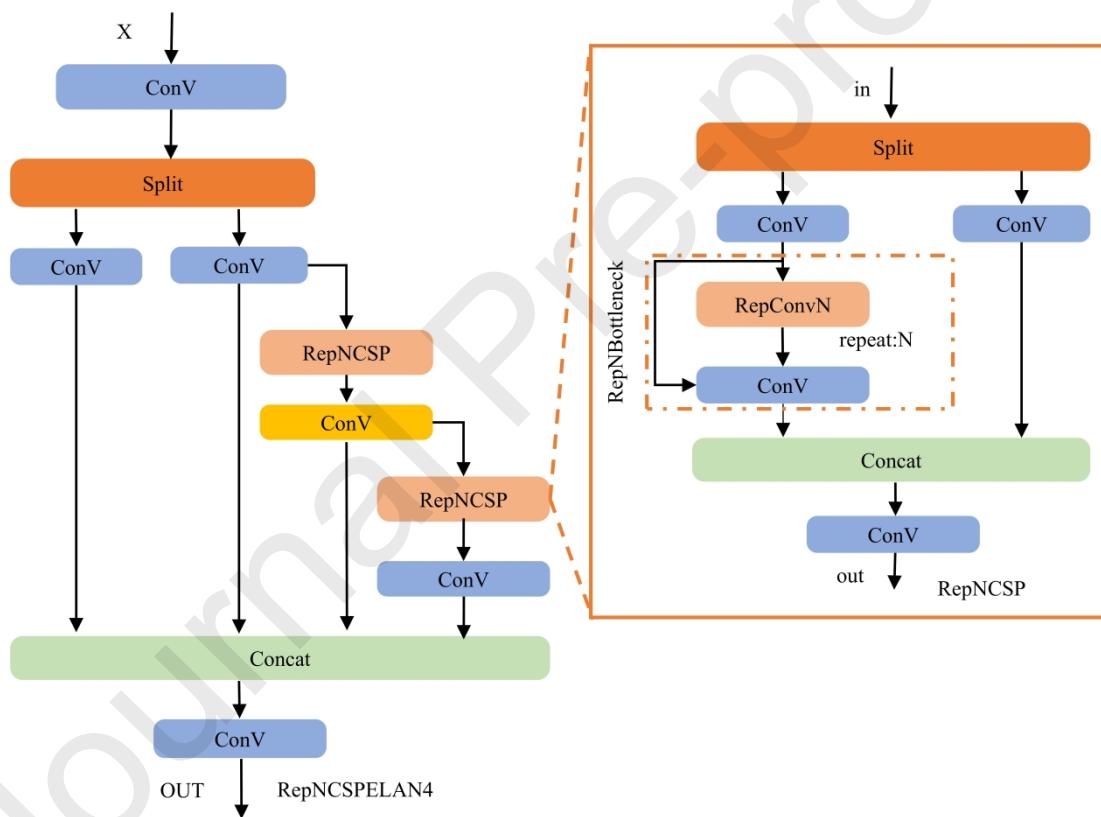
134 In the detection of tea buds, the presence of old leaves and the small size of the
 135 buds can create significant background interference. Traditional convolution
 136 operations depend on fixed sampling points, which may lead to the oversight of
 137 complex spatial variations, especially when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

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

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

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

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

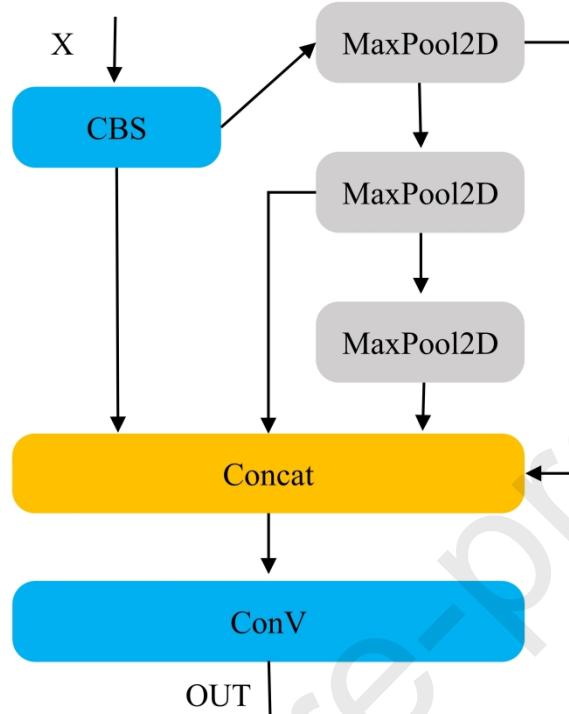


160 **Fig. 3.** RepNCSELAN4 module.

161 **2.2.2. SPPELAN feature fusion**

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

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

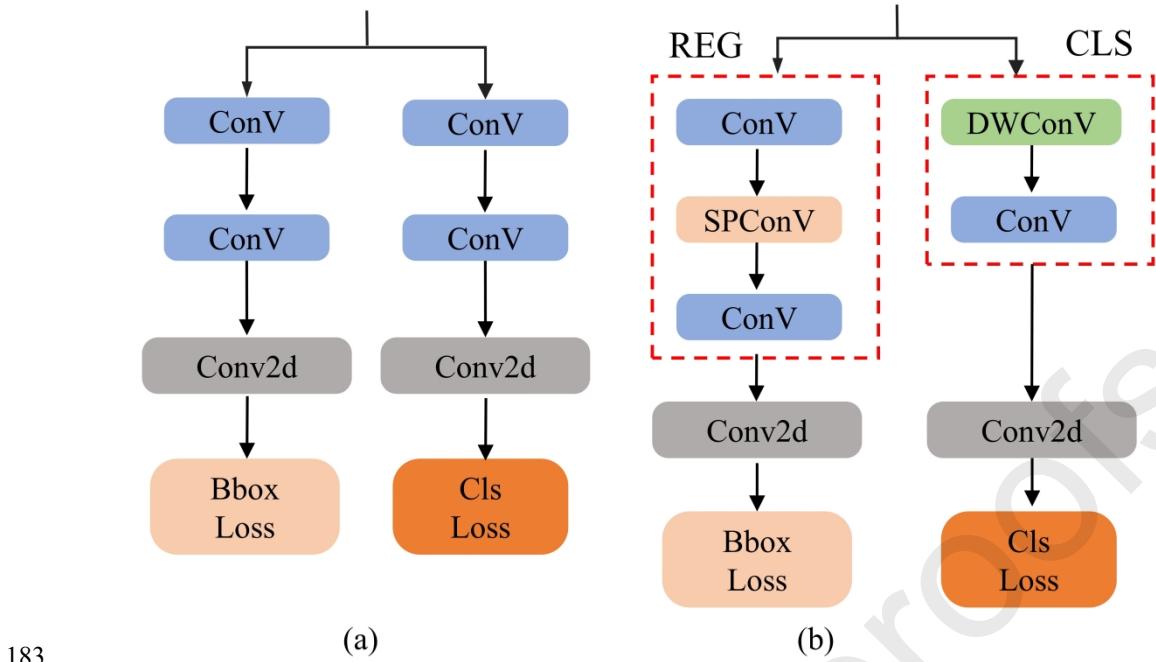


169

170 **Fig. 4.** SPPELAN data stream structure

171 *2.2.3. Improved lightweight decoupling detection head*

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



183

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

186 **2.2.4. *d*DIOU**

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

$$L_{dD Io U} = 1 - Io U + \frac{\rho^2}{c^2} + \alpha ; (1)$$

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

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

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

195 where

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

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

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

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

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

205 2.3. Harvest point estimation

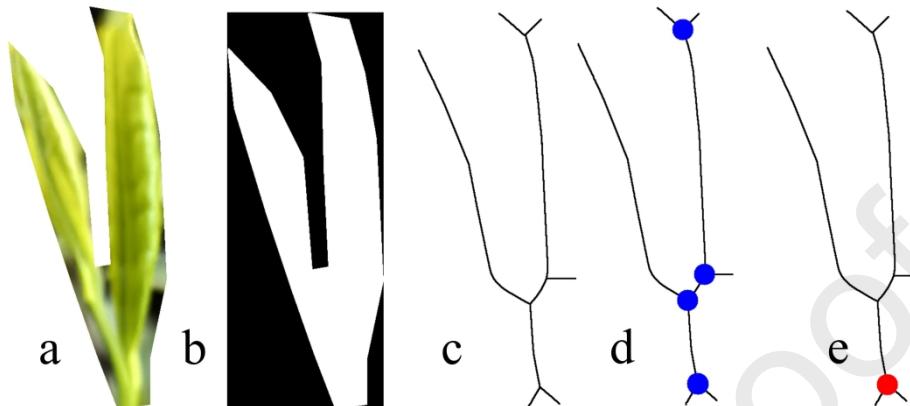
206 Building on the bounding boxes detected by the YOLO-MEST model, we
207 propose a pipeline based on a morphological algorithm to estimate the picking points
208 of tea buds. The process begins with foreground extraction using the GrabCut
209 algorithm^[14], as demonstrated in Fig. 6a. Given an input image $I \in \mathbb{R}^{m \times n \times 3}$,
210 we first convert it to grayscale through a standard colour space transformation,
211 denoted as I_{gray} . The GrabCut algorithm generates a binary mask M , where the
212 automatic rectangular initialization region covers 95% of the image area (with a
213 $\delta = 5$ -pixel boundary margin). After five iterations of graph-cut optimization for energy
214 function minimization, a refined binary output $B \in \{0, 255\}^{m \times n}$ is generated,
215 with foreground pixels set to 255, as shown in Fig. 6b.

216 Lin et al.^[15] employed medial axis transform (MAT) for morphological
217 skeletonization of the binary image. For each connected component in B , an
218 8-connected skeleton $S \in \{0, 1\}^{m \times n}$ was computed via MAT, as shown in Fig. 6c.
219 A convolution kernel $K = [1 \ 1 \ 1; \ 1 \ 10 \ 1; \ 1 \ 1 \ 1]$ was constructed for adaptive
220 intersection detection, where a pixel (x, y) was identified as an intersection if
221 $(S \times K)(y, x) \geq 13$ (Fig. 6d).

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

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

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



230

231 **Fig. 6.** Harvest point estimation process

232 *2.4. Test procedure*

233 *2.4.1. Environment Configuration*

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

238 During model training, the number of epochs and batch size were set to 300 and
 239 32, respectively, to enhance training efficiency. The initial learning rate, determined
 240 through testing, was set at 0.001. The Sto-chastic Gradient Descent (SGD) optimizer
 241 was utilized for learning rate adjustment, with the cosine annealing hyperparameter
 242 and learning rate momentum established at 0.01 and 0.947, respectively. Input images
 243 were resized to 640×640 pixels, and training lasted for 300 iterations.
 244 Hyperparameters were chosen based on standard YOLO practices and refined through
 245 a limited grid search on the validation set to optimize mAP50. The final settings were
 246 selected for their stability and superior performance.

247 *2.4.2. Evaluation criteria*

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

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

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

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

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

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

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

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

263 3. Results and discussion

264 All experiments in this subsection are conducted using YOLOv8-n with an input
 265 size of 640×640 . Our study utilized 4,600 annotated data samples, focusing on tea
 266 buds for keypoint annotation and frame labeling. The dataset was divided into three
 267 subsets in an 8:1:1 ratio: a training set containing 3,680 samples, a validation set with
 268 460 samples, and a test set with 460 samples.

269 3.1. Ablation study on combined modules

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

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

276 **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

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

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

294 *3.2. Comparison of RepNCSELAN4 with other feature fusion methods*

295 To validate the effectiveness of the proposed RepNCSELAN4 model for tea bud
 296 detection, we compared it with five representative feature fusion mechanisms: CBAM
 297 [17], Shuffle Attention (SA)[18], SimAM[19], GAM[20], and SK[21]. A self-built,
 298 high-quality tea dataset was used for testing, as shown in [Table 2](#). YOLOv8-n was
 299 adopted as the base model, with each feature fusion module integrated into the C2f
 300 module, with outputs from layers 2 through 5.

301

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

302 The RepNCSELAN4 attention mechanism presented in this paper demonstrates
 303 significant advantages in terms of model performance and parameter efficiency, as
 304 shown in [Table 1](#). When compared to other mainstream attention mechanisms,
 305 RepNCSELAN4 reduces the number of parameters by 0.43 million while maintaining
 306 excellent detection performance and achieving a lighter network structure.
 307 Specifically, RepNCSELAN4 achieves a mean Average Precision (mAP50) of 83.9%,
 308 compared to the baseline of 83.2%, with 14.3% fewer parameters (2.58 million vs.
 309 3.01 million) than the baseline model. This improvement is largely attributed to the
 310 C2f-RepNCSELAN module, which effectively enhances the interaction learning
 311 between channel and spatial dimensions, enabling the model to comprehensively
 312 understand feature information.

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

320 *3.3. Improvements to LiteShiftHead compared to other detection heads*

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

326 **Table 3** Comparison results of detection heads

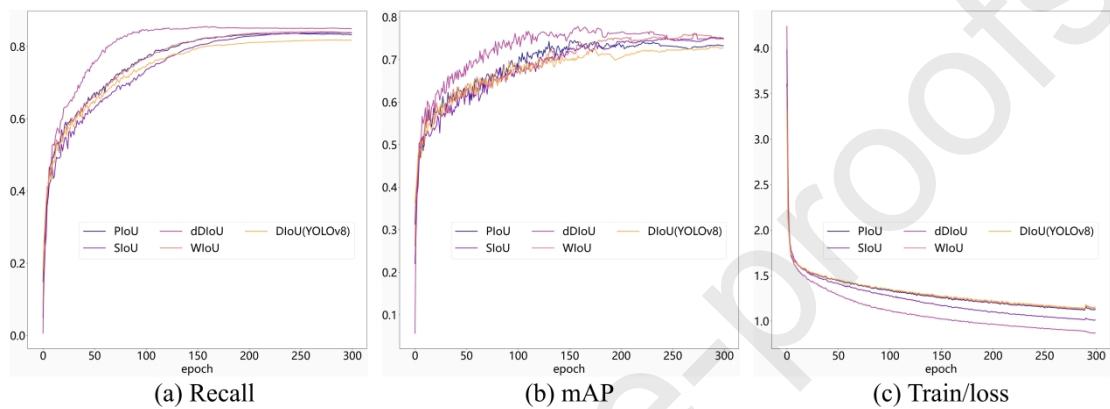
Head	mAP50 (%)	Parameters (M)	FLOPs (G)	mAP50-to-Baseline (M)
Baseline	83.2	3.01	8.2	-
DynamicHead	83.8	3.01	8.2	+0.6
LADH	83.1	3.01	8.2	-0.1
LiteShiftHead	83.9	3.01	8.1	+0.7

327 As shown in [Table 2](#), the improved LiteShiftHead achieves a mAP50 score of
 328 83.9% for the object detection model, representing a 1% improvement over the
 329 DynamicHead scheme. This enhancement is a result of the decoupling of the detection
 330 head structure. The object localization branch utilizes spatial pyramid convolutions
 331 (SPConv) to integrate contextual information from various receptive fields through
 332 parallel, multi-scale convolution kernels. This significantly improves the model's
 333 ability to detect targets of different sizes. Meanwhile, the category classification
 334 branch employs depth-separable convolution (DWConv), which combines channel
 335 decoupling with point-by-point fusion mechanisms. This approach maintains feature
 336 discrimination while reducing theoretical computational complexity. Overall, this
 337 task-specific, modular design strategy enables the detection head to effectively
 338 balance feature expression capabilities with computational efficiency.

339 *3.4. Comparison of different loss functions*

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

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



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

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

363 **Table 4** Loss function test results

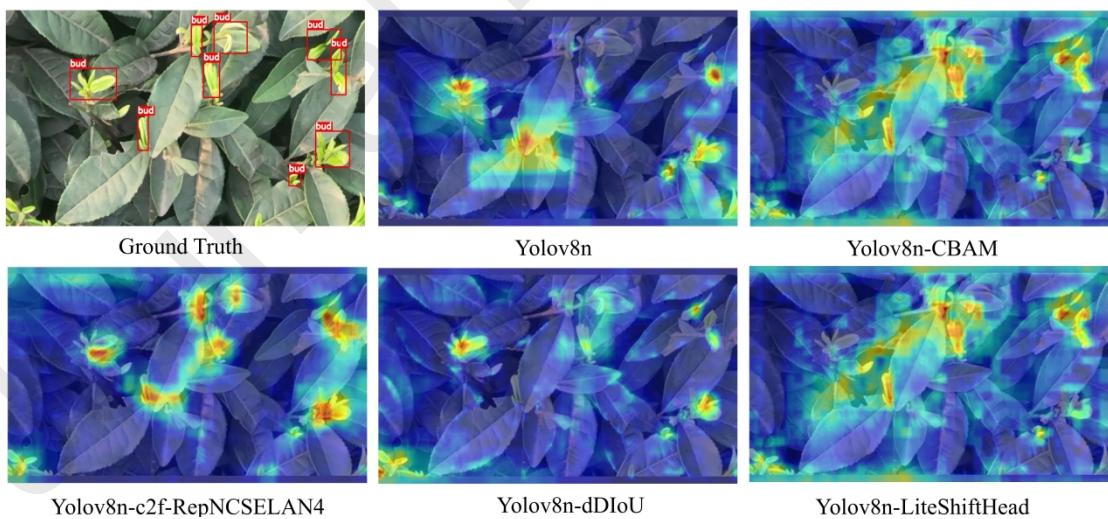
IoU	mAP50 (%)	Parameters (M)	FLOPs (G)	mAP-to-Baseline (%)
DIoU (YOLOv8)	83.1	3.01	8.2	-
WIoU	83.3	3.01	8.2	+0.2
SIoU	83.1	3.01	8.2	0

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

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

374 *3.5. Feature detection inference and evaluation*

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

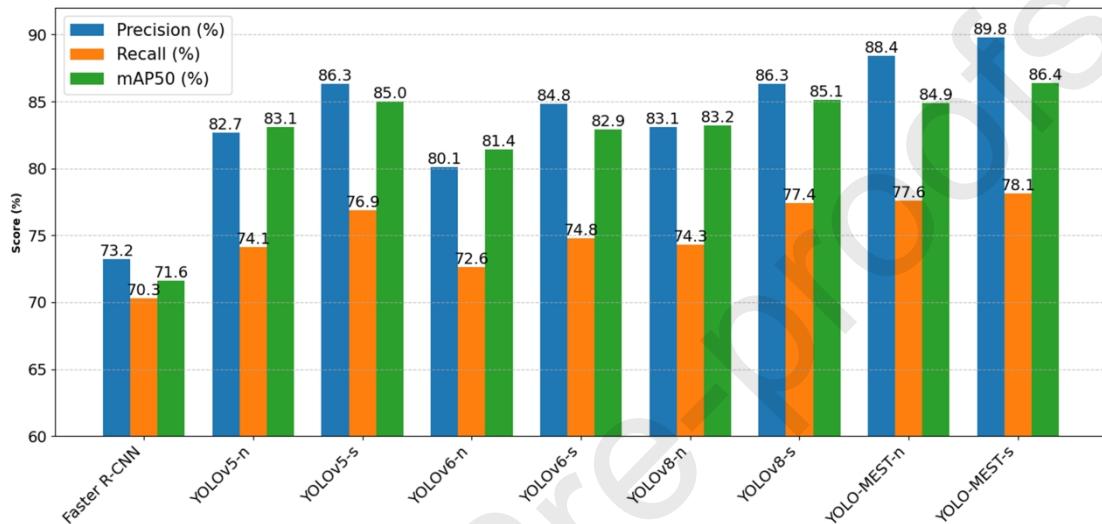


381 **Fig. 8.** Comparison of XGrad-CAM feature maps.

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

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

393 *3.6. Model comparison analysis*



394

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

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

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

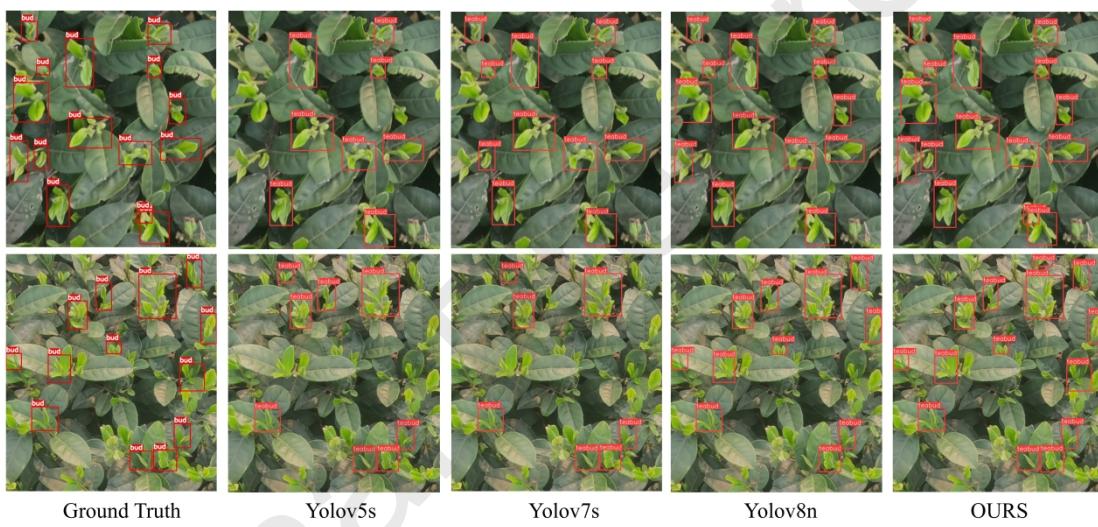
411

412 In contrast, the two-stage detection model, Faster R-CNN, yielded a mAP50 of only
 413 71.6% with 41.26 million parameters, making it the least effective model in terms of
 414 overall performance and unsuitable for the tea bud detection task in this study. The
 415 YOLO series models are inherently more suited for object detection tasks due to their

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

420 *3.7. Adaptability test*

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



429
 430 **Fig.10.** Visualization of detection results using different methods

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

437 *3.8. Harvest point estimation*

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

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



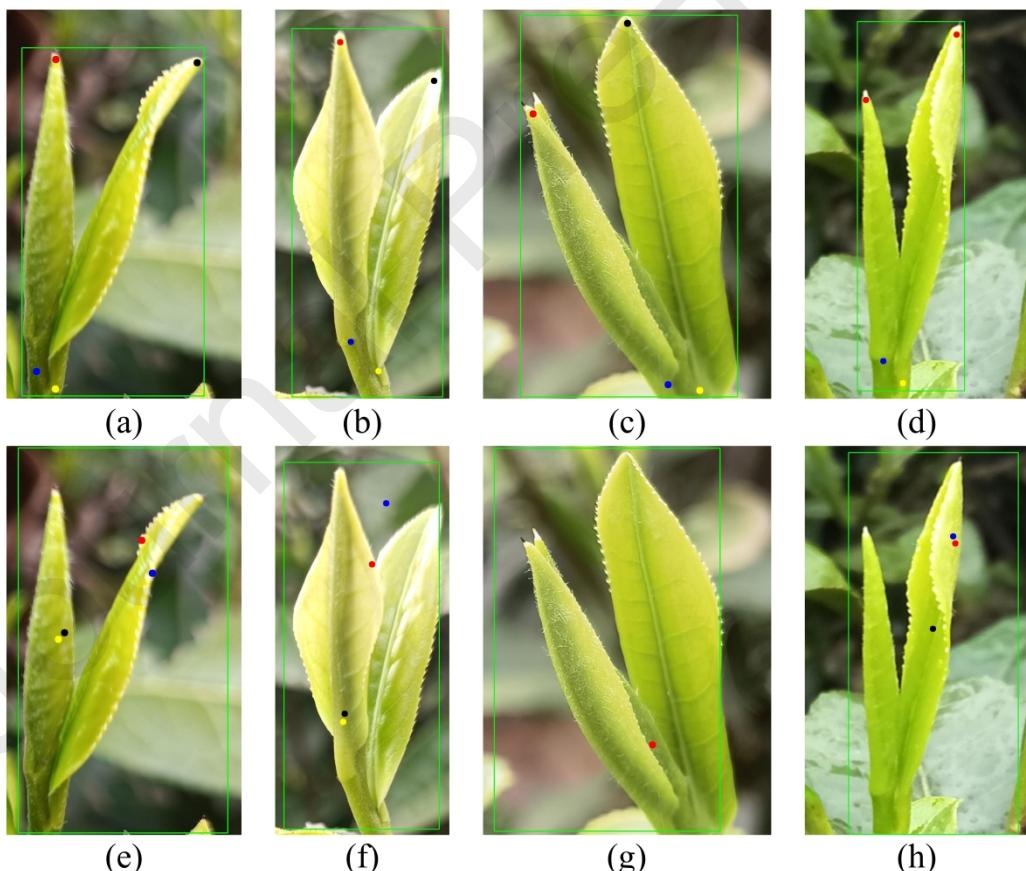
457 **Fig.11.** Harvesting point detection.

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

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

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

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



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

484 **4. Conclusion**

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

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

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

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

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

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

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

532 **Data availability**

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

535 **CRediT authorship contribution statement**

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

543 **Declaration of competing interest**

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

547 **Acknowledgments**

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

550 **References**

- 551 [1] Zhao X, He L, Li Y, Chen J, Wu C. Kinetostatic modeling of clamping force in a
 552 tendon-driven soft robotic gripper for tea shoot plucking. Comput Electron Agric
 553 2025;236:110441. <https://doi.org/10.1016/j.compag.2025.110441>.
- 554 [2] Lin G, Xiong J, Zhao R, Li X, Hu H, Zhu L, Zhang R. Efficient detection and
 555 picking sequence planning of tea buds in a high-density canopy. Comput
 556 Electron Agric 2023;213:108213. <https://doi.org/10.1016/j.compag.2023.108213>.

- 557 [3] Xie D, Chen L, Liu L, Chen L, Wang H. Actuators and sensors for applic
558 ation in agricultural robots: A review. *Machines* 2022;10(10):913. <https://doi.org/10.3390/machines10100913>.
- 560 [4] Zhang L, Zou L, Wu C, Jia J, Chen J. Method of famous tea sprout identification
561 and segmentation based on improved watershed algorithm. *Comput Electron
562 Agric* 2021;184:106108. <https://doi.org/10.1016/j.compag.2021.106108>.
- 563 [5] Zhang L, Zhang H, Chen Y, Dai S, Li X, Kenji I, Li M. Real-time monitoring of
564 optimum timing for harvesting fresh tea leaves based on machine vision. *Int J
565 Agric Biol Eng* 2019;12(1):6-9. <https://doi.org/10.25165/j.ijabe.20191201.3418>.
- 566 [6] Ren S, He K, Girshick R, Sun J. Faster r-cnn: Towards real-time object detection
567 with region proposal networks. *IEEE Trans Pattern Anal Mach Intell*
568 2016;39(6):1137-1149. <https://doi.org/10.1109/TPAMI.2016.2577031>.
- 569 [7] Pan Z, Gu J, Wang W, Fang X, Xia Z, Wang Q, Wang M. Picking point
570 identification and localization method based on swin-transformer for
571 high-quality tea. *Journal of King Saud University-Computer and Information
572 Sciences* 2024;36(10):102262. <https://doi.org/10.1016/j.jksuci.2024.102262>.
- 573 [8] Chen J, Chen Y, Jin X, Che J, Gao F, Li N. Research on a parallel robot
574 for tea flushes plucking. In: 2015 International Conference on Education,
575 Management, Information and Medicine. 2015, p. 22-26. <https://doi.org/10.2991/emim-15.2015.5>.
- 577 [9] Chen C, Lu J, Zhou M, Yi J, Liao M, Gao Z. A YOLOv3-based computer vision
578 system for identification of tea buds and the picking point. *Comput Electron
579 Agric* 2022;198:107116. <https://doi.org/10.1016/j.compag.2022.107116>.
- 580 [10] Li J, Li J, Zhao X, Su X, Wu W. Lightweight detection networks for tea bud on
581 complex agricultural environment via improved YOLO v4. *Comput Electron
582 Agric* 2023;211:107955. <https://doi.org/10.1016/j.compag.2023.107955>.
- 583 [11] Gui Z, Chen J, Li Y, Chen Z, Wu C, Dong C. A lightweight tea bud detection
584 model based on Yolov5. *Comput Electron Agric* 2023;205:107636.
585 <https://doi.org/10.1016/j.compag.2023.107636>.
- 586 [12] Varghese R, Sambath M. Yolov8: A novel object detection algorithm with
587 enhanced performance and robustness. In: 2024 International Conference on
588 Advances in Data Engineering and Intelligent Computing Systems (ADICS).
589 IEEE; 2024, p. 1-6. <https://doi.org/10.1109/ADICS58448.2024.10533619>.
- 590 [13] Wu Y, Zhu Y, Guo J, Yu Y. YOLOv5 detection algorithm of steel Defects based
591 on introducing light convolution network and DIOU function. In: 2023 IEEE
592 12th Data Driven Control and Learning Systems Conference (DDCLS), Xiangtan,
593 China; 2023, p. 118-122. doi:10.1109/DDCLS58216.2023.10165997.

- 594 [14] Boykov Y, Jolly M P. Interactive organ segmentation using graph cuts. In:
 595 International conference on medical image computing and computer-assisted
 596 intervention. Berlin, Heidelberg: Springer Berlin Heidelberg; 2000, p. 276-286.
 597 https://doi.org/10.1007/978-3-540-40899-4_28.
- 598 [15] Lin C, Liu L, Li C, Kobbelt L, Wang B, Xin S, Wang W. Seg-mat: 3d shape
 599 segmentation using medial axis transform. IEEE Trans Visual Comput. Graphics
 600 2020;28(6):2430-2444. <https://doi.org/10.1109/TVCG.2020.3032566>.
- 601 [16] Li S, Meng J, Lv D, Chen X, Liu J. Defect Detection Method Based on Improved
 602 YOLOv5s Model. In: 2023 2nd International Conference on 3D Immersion,
 603 Interaction and Multi-sensory Experiences (ICDIIME). IEEE; 2023, p. 331-336.
 604 <https://doi.org/10.1109/ICDIIME59043.2023.00070>.
- 605 [17] Qiao S, Chen L C, Yuille A. DetectoRS: Detecting objects with recursive feature
 606 pyramid and switchable atrous convolution. In: Proceedings of the IEEE/CVF
 607 conference on computer vision and pattern recognition. 2021, p. 10213-10224.
 608 <https://doi.org/10.1109/CVPR46437.2021.01008>.
- 609 [18] Norkobil Saydirasulovich S, Abdusalomov A, Jamil M K, Nasimov R,
 610 Kozhamzharova D, Cho Y I. A YOLOv6-based improved fire detection
 611 approach for smart city environments. Sensors 2023;23(6):3161.
 612 <https://doi.org/10.3390/s23063161>.
- 613 [19] Yang D, Huang Z, Zheng C, Chen H, Jiang X. Detecting tea shoots usin
 614 g improved YOLOv8n. Transactions of the Chinese Society of Agricultural
 615 Engineering 2024;40(12):165-173. [https://doi.org/10.11975/j.issn.1002-6819.2
 02401155](https://doi.org/10.11975/j.issn.1002-6819.2

 616 02401155).
- 617 [20] Wang Y, Xiao M, Wang S, Jiang Q, Wang X, Zhang Y. Detection of famous tea
 618 buds based on improved YOLOv7 network. Agriculture 2023;13(6):1190.
 619 <https://doi.org/10.3390/agriculture13061190>.
- 620 [21] Zhang Q L, Yang Y B. Sa-net: Shuffle attention for deep convolutional neural
 621 networks. In: ICASSP 2021-2021 IEEE International Conference on Acoustics,
 622 Speech and Signal Processing (ICASSP). IEEE; 2021, p. 2235-2239.
 623 <https://doi.org/10.1109/ICASSP39728.2021.9414568>.
- 624 [22] Dai X, Chen Y, Xiao B, Chen D, Liu M, Yuan L, Zhang L. Dynamic head:
 625 unifying object detection heads with attentions. In: 2021 IEEE/CVF Conference
 626 on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA;
 627 2021, p. 7369-7378, doi: 10.1109/CVPR46437.2021.00729.
- 628 [23] Guo Y, Shen Q, Zhang S, Zhang C, Wang X. An airborne target recognition
 629 model based on SPD, PConv and LADH detection heads. In: International

- 630 Conference on Autonomous Unmanned Systems. Singapore, Springer Nature
 631 Singapore; 2023, p. 325-337. https://doi.org/10.1007/978-981-97-1087-4_31.
- 632 [24] Zheng Z, Wang P, Liu W, Li J, Ye R, Ren D. Distance-IoU loss: Faster
 633 and better learning for bounding box regression. In: Proceedings of the A
 634 AAI conference on artificial intelligence. 2020;34(07):12993-13000. <https://doi.org/10.1609/aaai.v34i07.6999>.
- 636 [25] Dai X, Chen Y, Xiao B, Chen D, Liu M, Yuan L Zhang L. Dynamic head:
 637 Unifying object detection heads with attentions. In: 2021 IEEE/CVF Conference
 638 on Computer Vision and Pattern Recognition (CVPR). 2021, p. 7373-7382.
 639 <https://doi.org/10.48550/arXiv.2106.08322>.
- 640 [26] Zhang J, Chen Z, Yan G, Wang Y, Hu B. Faster and lightweight: an improved
 641 YOLOv5 object detector for remote sensing images. Remote Sens
 642 2023;15(20):4974. <https://doi.org/10.3390/rs15204974>.
- 643 [27] Liu C, Wang K, Li Q, Zhao F, Zhao K, Ma H. Powerful-IoU: More strai
 644 ghtforward and faster bounding box regression loss with a nonmonotonic f
 645 ocusing mechanism. Neural Networks 2024;170:276-284. <https://doi.org/10.1016/j.neunet.2023.11.041>.
- 647 [28] Xiao S, Zhao Q, Chen Y, Li T. A dual-backbone lightweight detection and depth
 648 position picking system for multiple occlusions Camellia oleifera fruit. Comput
 649 Electron Agric 2025;233:110157. <https://doi.org/10.1016/j.compag.2025.110157>.
- 650 [29] Li Y, Zhang J, Gao P, Jiang L, Chen M. Grab cut image segmentation b
 651 ased on image region. In 2018 IEEE 3rd international conference on imag
 652 e, vision and computing (ICIVC). IEEE; 2018, p. 311-315. <https://doi.org/10.1109/ICIVC.2018.8492818>.
- 654 [30] Dong C, Tang Y, Zhang L. HDA-pose: a real-time 2D human pose estimation
 655 method based on modified YOLOv8. Signal, Image and Video Processing.
 656 2024;18(8):5823-5839. <https://doi.org/10.1007/s11760-024-03274-2>.

657 ■ **Ethics Statement**

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

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

661

662

663 If Yes, please provide your text here:

664 [31]

665 ■ Authors' Biographies

666

667



668 **Chuanyang Yu** (chuanyang@ahjzu.edu.cn) is currently a Ph.D.
669 candidate at the College of Engineering, Anhui Agricultural University.
670 He received his B.Eng. (2011) and M.Eng. (2014) degrees in Vehicle
671 Engineering from Anhui Agricultural University. His research focuses on intelligent
672 agricultural machinery and smart farming equipment, particularly exploring the applications
673 of vehicle engineering in modern agriculture.
674

675

676

677



678 **Yi Xue** (xueyi123@stu.ahau.edu.cn) is currently pursuing his
679 degree at the College of Engineering, Anhui Agricultural University.
680 He obtained his Master's degree in Agricultural Engineering from Anhui
681 Agricultural University in 2025 and his Bachelor's degree in Mechanical Engineering from
682 the Anhui Agricultural University College of Economics and Technology in 2020. His
683 research primarily focuses on visual image processing and its applications in agricultural
684 engineering.
685

686

687



688 **Liuyang Zhang** (liuyangzhang@stu.ahau.edu.cn) is currently a
689 candidate at the College of Engineering, Anhui Agricultural University.
690 He received his Bachelor's degree in Vehicle Engineering from Anhui
691 Agricultural University in 2023. His research focuses on intelligent
692 agricultural machinery, with a particular interest in the development and optimization of
693 smart farming equipment.
694

695

696

697



702 **Xue An** (anxue@ahau.edu.cn) is an Associate Professor and master's supervisor at the
 703 Department of Agricultural Machinery, School of Engineering, Anhui Agricultural University.
 704 She received her bachelor's degree in Agricultural Mechanization and Automation (2015) and
 705 her master's degree in Agricultural Engineering (2018) from Henan Agricultural University.
 706 She obtained her Ph.D. in Agricultural Mechanization Engineering from the College of
 707 Mechanical and Electronic Engineering, Northwest A&F University in 2023. From March
 708 2022 to February 2023, she was a CSC-funded joint Ph.D. researcher at the Department of
 709 Precision Horticulture, Leibniz Institute for Agricultural Engineering and Bioeconomy (ATB),
 710 Germany. From March to September 2023, she continued her research as an assistant
 711 researcher in the Department of Mechatronics in Agriculture at ATB. Dr. An's research
 712 interests include intelligent agricultural machinery, equipment for protected horticulture, and
 713 smart agricultural systems.

714

715

716

717



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

726

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

730

731

732 **Liqing Chen** (lqchen@ahau.edu.cn). He received the B.E. degree,
 733 the M.S. and Ph.D. degrees in automotive engineering from Hefei
 734 University of Technology, China, in 2000, 2005, and 2015, respectively.
 735 He is a Professor and doctoral supervisor at the School of Engineering,
 736 Anhui Agricultural University. He currently serves as the Director of the Anhui Provincial
 737 Engineering Technology Center for Intelligent Agricultural Machinery. He is also a member
 738 of the Youth Talent Support Program of the Ministry of Agriculture and Rural Affairs of
 739 China, a post scientist in the National Wheat Industry Technology System, and a recipient of
 740
 741



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]