ISSN 1330-3651 (Print), ISSN 1848-6339 (Online)
Received: 11 July 2025; Accepted: 30 August 2025

https://doi.org/10.17559/TV-20250711002819
Original scientific paper

Enhanced YOLO Architecture with Attention Mechanism for Accurate Tobacco Plant Counting from UAV Images

Chuanzhi MA, Yuehan LI, Yilong PENG, Ran WANG, Jiani LIU, Shaofan TANG, Fu WANG, Jianhua LI*

Abstract: (1) Background: This study investigates the construction and optimization of the You Only Look Once (YOLO) deep learning model for high-precision identification of suitable tobacco leaves. (2) Methods: Using tobacco fields in Xiaoxin Street, Niulanjiang Town, Songming County, Kunming as the study area, a total of 1200 UAV images collected during the planting, growth, and harvesting stages were employed as the training dataset to train object detection models such as YOLO v3. After 200 training iterations, the recognition performance of each model was compared and analyzed. (3) Results: YOLO v5 and YOLO v7 were selected as baseline models, and a channel attention mechanism was integrated to develop the improved YOLO v5-EN model. Ablation experiments were conducted by incorporating the attention module, dynamic rectified linear unit (DReLu) activation function, and a feature refinement module. YOLO v7 en was designed as a backbone network, and metrics such as precision, recall, and accuracy were comprehensively evaluated to assess the performance of both the baseline and improved models in identifying the number of tobacco plants. Compared to the baseline, the improved YOLO v5 model demonstrated a 0.36% increase in precision and a 1.55% increase in recall, achieving an overall recognition accuracy of 91.41%. The improved YOLO v7 model achieved a precision of 99.16% and a mean average precision (map) of 95.86%. These results indicate that the enhanced YOLO v5 model with channel attention effectively addresses the issues of missed and false detections in tobacco plant recognition. Furthermore, the improved YOLO v7 model, integrated with collaborative optimization strategies and an enhanced backbone, significantly improves the performance and efficiency of the detection model, particularly in terms of accuracy and processing speed for complex visual tasks. (4) Conclusions: The improved YOLO models significantly enhance the accuracy of tobacco plant count recognition and offer a practical solution for efficient tobacco plant statistics, serving as a reference for intelligent agriculture.

Keywords: deep learning; number of trees; tobacco leaf; YOLO

1 INTRODUCTION

Tobacco is an important cash crop. Yunnan Province ranks among the top in China in both tobacco yield and total planting area, making a substantial contribution to national economic development [1]. The tobacco industry in China urgently requires the adoption of advanced technologies and scientific management approaches to enhance the technological sophistication and intelligent level of tobacco cultivation, while continually increasing demands for accuracy and efficiency in information extraction technologies [2, 3]. Leaf count serves as a crucial indicator for evaluating tobacco plant growth. During the sowing stage, it is used to assess emergence rate and seeding quality; during the growth stage, it helps evaluate plant damage; and during the harvesting stage, it supports biomass analysis and yield estimation [4]. The accuracy of leaf count directly influences the allocation of agricultural resources and crop yield outcomes. Therefore, precise monitoring of leaf number is of great significance for field management, productivity enhancement, and income improvement. The conventional method of tobacco plant counting primarily relies on manual observation, which is time-consuming, labor-consuming, and inefficient. Its accuracy is constrained by personnel experience and counting precision. Thus, there is an urgent need for a convenient and efficient method for extracting tobacco plant counts to address these limitations.

With the rapid advancement of UAV remote sensing technology, UAVs equipped with high-resolution cameras have emerged as a novel approach for data acquisition, offering an efficient and accurate method for extracting tobacco plant counts and significantly enhancing the efficiency and spatial coverage of data collection [5]. The integration of UAVs and deep learning models enables precise and efficient crop monitoring, pest and disease detection, and yield prediction, thereby improving the accuracy and operational efficiency of agricultural production [6]. This approach also enhances the real-time responsiveness of crop management and holds substantial potential in agricultural big data applications and intelligent management systems, facilitating the transition from traditional to modern agriculture [7]. YOLO (You Only Look Once), a real-time object detection framework based on deep learning, has been extensively adopted across various domains due to its speed and accuracy. By utilizing imagery captured by UAVs equipped with high-definition camera in combination with the YOLO model, automated identification and counting of tobacco plants can be achieved, improving the efficiency and accuracy of population estimation while eliminating subjective errors associated with manual counting.

In recent years, numerous researchers have focused on modifying the YOLO series algorithms to enhance target detection performance, resulting in notable improvements in recognition accuracy and operational efficiency. Cong et al. employed the YOLO v5s convolutional neural network to detect Camellia oleifera fruits under natural conditions, demonstrating high accuracy and robustness in handling dense fruit clusters, occlusion, low lighting, and image blur [8]. Bai et al. introduced detailed adaptations of the YOLO model tailored to diverse application scenarios and requirements, indicating the model can more flexibly and effectively handle various complex and changeable visual recognition tasks [9]. Wang et al. conducted a series of experiments on cherry datasets using the YOLOV5 model, which substantially improved recognition accuracy in cherry classification tasks [10]. Han et al. developed a YOLOv5s detection model through transfer learning and applied a channel pruning algorithm for simplification, resulting in enhanced detection speed and accuracy for apple fruits [11]. Wang et al. proposed a method utilizing YOLO and deep learning techniques to count, train, and test UAV images of greenhouse-grown red tomatoes, green tomatoes, and tomato flowers, achieving significantly higher recognition accuracy while reducing manual labor [12]. Yunus et al. designed an end-to-end recognition approach for UAV images of navel orange trees using the YOLOV4 deep learning algorithm, specifically addressing

small target detection in complex natural environments [13]. Loddo et al. applied deep learning techniques to extract recognition features from plant images and used a linear support vector machine for classification, with deep neural networks exhibiting markedly superior performance compared to manually extracted features [14].

Although the YOLO model has demonstrated strong performance in plant recognition, it continues to face several technical challenges. Owing to the large diversity of plant species and significant morphological variations, a single YOLO model may be insufficient to accurately identify all categories. Therefore, optimizing the YOLO-based methodology is essential for improving the accuracy of tobacco plant recognition. Additionally, the complexity of backgrounds in plant images presents further obstacles to model precision. Enhancing the training dataset through the incorporation of images with diverse and dynamic backgrounds can significantly improve the robustness of the model. This study aims to address existing limitations in tobacco plant counting, including constraints related to operational range and time, high costs of acquisition equipment, and lack of timeliness. By using tobacco images collected by UAV, extracting orthophoto imagery, and refining the detection model based on the YOLO algorithm, this research enables accurate estimation of the number of tobacco plants in the field. The proposed approach provides substantial theoretical and practical value for the precise identification of crops and the efficient management of agricultural production.

2 MATERIALS AND DATA SOURCES

2.1 Study Area

The study area is situated in Xiaoxin street, Niulanjiang Town, Songming County, Kunming City, Yunnan Province, within the geographical location is 102°29'to 102°58'e and 24°59'to 25°21'n. It lies on the eastern margin of the Yunnan-Guizhou Plateau, with most of the terrain ranging in elevation from 1500 m to 2100 m. The region exhibits distinct characteristics of a plateau climate and is classified as a subtropical plateau monsoon climate. The annual average temperature ranges from 15° C to 17° C, with mild winters and cool summers. Annual precipitation is approximately 1000-1200 mm, primarily concentrated between May and October, creating favorable conditions for the cultivation of tobacco and other crops.

2.2 Data Acquisition and Preprocessing

The primary device used for remote sensing image acquisition was the Dajiang spirit 4 multi spectral UAV. The total weight of the fuselage is 1487 g, with a maximum battery capacity of 5870 mAh, allowing a maximum flight duration of 30 min per battery. The horizontal flight speed reaches 50 km/h in positioning mode and 58 km/h in attitude mode. The operational temperature range is 0° to 40°. The camera features a field of view of 62.7°, a focal length of 5.74 mm, and a maximum image resolution of 1600 × 1300 (4:3.25). The gimbal offers a controllable pitch range from –90° to +30°. The lens filters are centered at 450 nm ± 16 nm (blue), 560 nm ± 16 nm (green), 650 nm ± 16 nm (red), 736 nm ± 16 nm (red edge), and 840 nm ± 16 nm (near-infrared). Data were collected on four occasions between May and July 2023.

Based on the conditions of the experimental area, four ground control points were established for each data acquisition session to assist in aerial navigation. RTK-based real-time positioning tools were employed to ensure high spatial accuracy and data reliability of the UAV imagery. Flight paths were programmed using Dajiang software, and images were captured at three altitudes - 10 m, 15 m, and 25 m - resulting in a total of 7054 images. The ground sampling distance was 3 cm, referenced to the World Geodetic System 1984 (WGS-84). Pix4d Mapper was utilized for parameter configuration and subsequent processing to generate high-resolution digital orthophoto maps (DOM) [15] and digital surface models (DSM) [16].

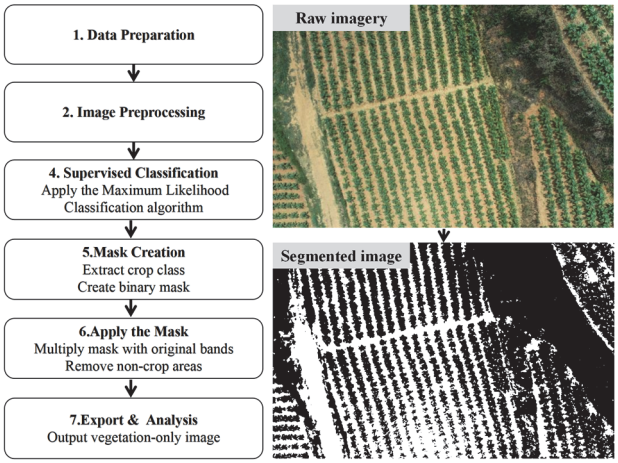


Figure 1 Supervised classification-based soil background elimination

To calibrate the characteristics of the UAV image dataset accurately, tailored optimization strategies were applied based on stage-specific features. As presented in Tab. 1, during the tobacco planting stage, each plant typically exhibits 3-6 leaves with small individual leaf areas, resulting in low coverage.

Table 1 Problems and solutions for different growth cycles of tobacco

Growth period	Problems in growth period	Optimization measures
Planting period	Tobacco seedlings are small in size, light in leaf color, and the color contrast with soil is not obvious	Using contour feature fusion
Growth period	Feature recognition of growth freshman inconsistent impact	Using data enhancement to increase canopy feature learning
Picking period	Uneven leaf overlapping is serious at flowering stage	Integrate edge feature module

This leads to high color contrast with the surrounding soil, facilitating clear visualization of leaf contours. The overall plant morphology and edge outlines serve as key indicators for identifying the seedling stage. During the leaf growth phase, plant overlap increases and vegetation coverage becomes denser. The tops of the leaves often form distinctive trumpet-shaped structures that are seldom obscured by adjacent foliage, making them critical features for detection in this period. In the harvesting stage, the canopy reaches full development with increased leaf stacking and maximum plant height. Tobacco flowers generally emerge at the apex of the plants, and as maturity progresses, some leaves gradually turn yellow. These

phenotypic traits provide reliable cues for identifying plants in the mature stage.

In deep learning, increasing the volume of training data has been demonstrated as an effective approach to mitigate overfitting [17]. Data augmentation not only significantly improves algorithm accuracy but also enhances model robustness [18]. In this study, tobacco images from various growth stages were annotated to create an image dataset, which was then segmented into 640 × 640 patches. Over 7000 images were collected, and the input dataset was prepared with labeled samples. Due to the low saliency of certain UAV-captured features during some growth stages, which may lead to reduced accuracy, a total of 3000 segmented images were selected as the initial dataset. Among these, 1000 images from three distinct stages were used as the validation set, while 2000 images were used for training. Additionally, geometric and pixel-level transformations [19] were applied to enhance dataset quality and improve model performance and adaptability.

2.3 Basic Model and Improvement

2.3.1 Experimental Environment

The experimental platform is configured with Windows10 (Professional Edition) as the operating system, an Intel Core i9-13900 CPU operating at 5.60 GHz, 64 GB of RAM, an NVIDIA GeForce RTX 3090 GPU, and 24 GB of video memory. The development environment uses Python 3.9, and the deep learning framework is PyTorch GPU version 2.0.1. The experimental settings are as follows: the input image size is 640 × 640, the batch size is 16, the initial learning rate is 0.01, the weight decay is 0.0005, the optimizer is Adam, and the number of training epochs is 200.

2.3.2 YOLO Basic Model

In deep neural network, object detection is a critical task that involves locating the positions of objects within images or videos. The YOLO network is a representative object detection algorithm capable of detecting all objects in an image simultaneously through a single forward pass, significantly enhancing detection efficiency and speed. The input data is processed by a backbone network composed of convolutional modules, residual connections, and multi-scale pooling, followed by a neck structure that integrates features through a fusion mechanism. Finally, the output is generated via anchor boxes, a decoupled head, and a loss calculation detection head. The overall network architecture is illustrated in Fig. 1.

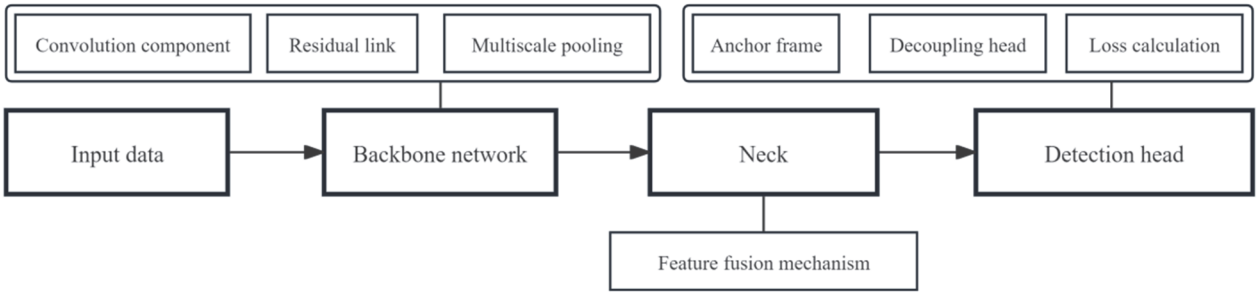


Figure 2 YOLO network structure

This study compares and analyzes widely used models including YOLO v3, YOLO v5, and YOLO v7 in terms of backbone network architecture, convolutional modules, and residual connections. YOLO v3 employs a 53-layer deep convolutional backbone with unidirectional feature fusion and a feature pyramid network. Its detection head utilizes non-maximum suppression with manually predefined anchor boxes. While its strengths lie in structural simplicity and broad compatibility, making it suitable for basic object detection tasks in simple environments, it suffers from high computational cost, limited performance in detecting small objects, and suboptimal real-time capability. YOLO v5 incorporates bidirectional integration via cross-stage partial networks, spatial pyramid pooling, and a path aggregation network. It simplifies the activation function, supports adaptive anchor box generation, and introduces a partially decoupled detection head. These enhancements improve inference speed and multi-scale feature utilization, enabling lightweight deployment without sacrificing accuracy. YOLO v5 is well-suited for real-time detection, shows improved performance on small to medium targets, and allows flexible deployment and parameter tuning. YOLO v7 is constructed using an efficient layer aggregation network, spatial pyramid pooling with cross-phase partial connections, and a fully decoupled detection head. It integrates distance-based Intersection over Union for non-maximum suppression, optimized transmission allocation, and dynamic loss functions. Feature reuse, bounding box regression, and sample assignment strategies are further reinforced. YOLOv7 excels in complex scenarios and is particularly effective for high-precision tasks, though it demands higher computational resources due to its model complexity.

In the baseline model, the augmented dataset is randomly partitioned, with 90% allocated for training, 10% reserved for testing, and 10% of the training portion further separated as the validation set. To reduce training time from random initialization, this study initializes model parameters using pretrained weights obtained from superior performance on a relevant dataset, thereby accelerating convergence and enhancing generalization. During training, the loss value is recorded after each epoch to facilitate monitoring and analysis. Stochastic gradient descent is employed as the optimization algorithm, with weight decay incorporated to mitigate overfitting. Additionally, a cosine annealing strategy is applied to progressively decrease the learning rate, enabling dynamic parameter adjustment throughout the training process. The final performance evaluation is conducted using the model configuration that yields the best results.

2.3.3 YOLO Model Improvement

Compared with the single shot multibox detector (SSD) and fast region-based convolutional neural network (fast r-cnn), which are widely used in agricultural target detection, this study integrates a channel attention mechanism [20, 21]. By applying global average pooling to the input feature map channels and replacing the fully connected layer with a 1 × 1 convolution layer, the traditional YOLOv5 architecture is modified to enhance inter-channel interaction and information flow. This modification enhances the capability of the model to extract the texture and structural features of tobacco heads, while suppressing interference from irrelevant background information, thereby improving both the precision and accuracy of tobacco leaf recognition.

In order to improve the performance of YOLO v5 in tobacco small target detection task, this paper introduces a lightweight channel attention mechanism Efficient Channel Attention Module (ECAM) in the backbone network, which establishes the dependency between channels through global average pooling and one-dimensional convolution, and can improve the feature expression ability without introducing additional parameters. This module is inserted into the Cross Phase Partial Structure Module (CPPSM) in the middle of the backbone network to enhance the characteristic response ability between channels, so as to obtain the improved YOLO v5-EN model. The introduction of the model significantly enhances the ability of the network to extract small target features, and effectively improves the detection accuracy while maintaining the lightweight of the model.

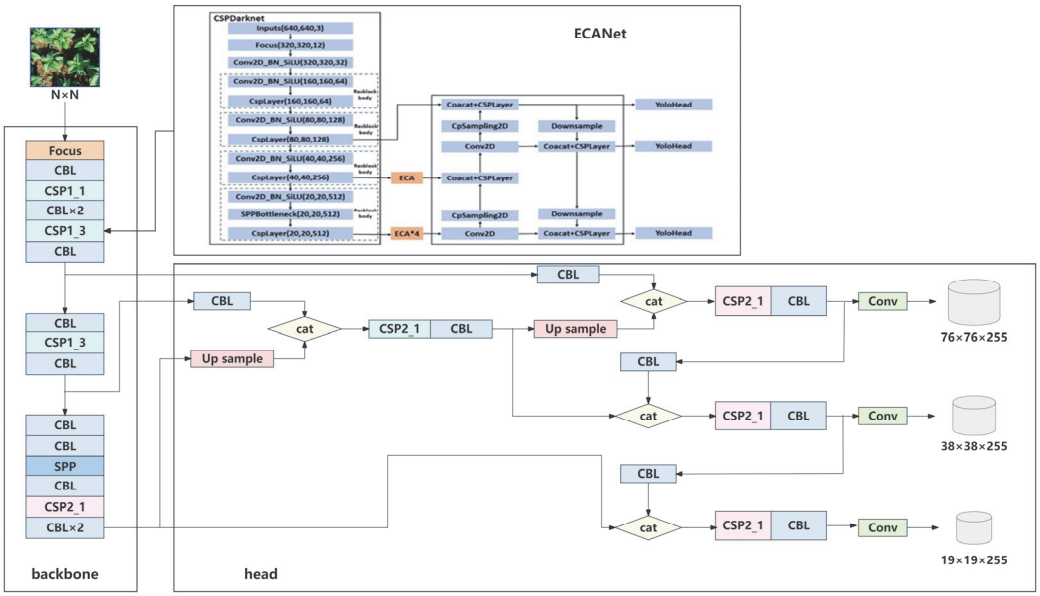


Figure 3 YOLO v5-EN network structure

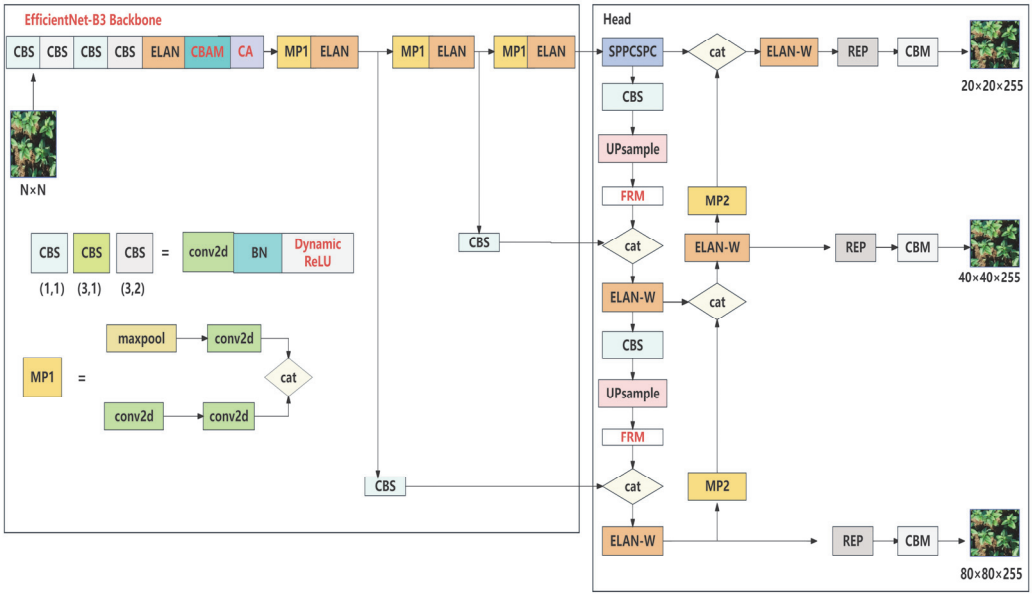


Figure 4 YOLO v7-EN network structure

In field environments, noise caused by weeds and surrounding leaves, together with the small size of tobacco targets, often hinders accurate detection. These factors substantially increase the difficulty of recognition based on neural networks and may result in missed detections or misclassification of non-tobacco targets. In this study,

EfficientNet is employed as a replacement for the backbone network in the traditional YOLOv7 model [22]. In addition, the convolutional block attention module is introduced, the dynamic ReLU activation function is embedded within convolutional components, and both a feature refinement module and a feature fusion mechanism are incorporated [23-25]. These improvements collectively lead to a significant enhancement in the overall recognition performance of the modified YOLO v7 model.

For the YOLOv7 model, this paper optimizes it from multiple structural levels to improve its robustness and accuracy in small target detection scenarios. Firstly, after the convolution module, the Convolution Block Attention Module (CBAM) and coordinate attention mechanism are introduced to enhance the fusion ability of spatial and channel features. Secondly, the original Silu activation function is replaced by dynamic relu with learnable parameters, so that the network has stronger adaptive expression ability. Thirdly, in the process of multi-scale feature fusion, the Feature Refinement Module (FRM) is added to further strengthen the expression ability of deep features through channel purification and spatial reweighting double branch design. Finally, the backbone network is replaced by a more efficient convolutional neural network (efficient net) from the traditional cross stage partially structured dark network (csparknet), which reduces the size of the parameters and improves the accuracy, so as to obtain the improved YOLOv7-EN model. The multi module joint experiment results show that the improved structure has good complementarity, and the overall performance of YOLOv7 in the target detection task is improved.

2.4 Evaluating Indicator

In the model experiments conducted in this study, precision, recall, and accuracy were selected as evaluation metrics. Accuracy reflects the degree of alignment between the detected targets and the actual targets. Precision represents the proportion of correctly identified targets among all detected targets in an image recognition task, serving as a measure of the ability of the algorithm to distinguish true positives from false positives or missed detections. This metric plays a critical role in the assessment of target detection algorithm performance. Recall is defined as the ratio of correctly detected true targets to the total number of actual targets, indicating the effectiveness of the algorithm in minimizing missed detections and enhancing comprehensive target identification. In target detection tasks, the mean average precision (mAP) is commonly used as an overall indicator of detection accuracy. For single-category detection tasks such as tobacco, mAP provides a direct and intuitive reflection of model performance on the specific category, offering a convenient and effective benchmark for evaluation. These metrics collectively provide a comprehensive assessment of model performance in target detection tasks and form a solid basis for performance evaluation.

The formula of Accuracy rate (A), Precision Rate (P) and Recall rate (R) is as follows:

$$A = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

$$P = \frac{TP}{TP + FP} \tag{2}$$

$$R = \frac{TP}{TP + FN} \tag{3}$$

In Eqs. (1), (2), and (3), TP represents the number of tobacco leaves accurately identified by the model, TN indicates the number of irrelevant targets correctly classified, FP refers to the number of unrelated objects incorrectly classified as tobacco leaves by the model or instances where the same tobacco leaf is detected multiple times, FN represents the number of tobacco leaves missed by the model, including those not detected within the same prediction frame.

3 RESULTS

3.1 Tobacco Number Extraction Based on Basic Model

In this study, 1200 images were employed as the training set, and the model was trained for 200 iterations. As the number of iterations increased, the accuracy loss on the training set gradually decreased. By the 100th iteration, the performance evaluation metrics of the YOLO model had stabilized, and no significant changes were observed thereafter, leading to the termination of training. At this stage, the model had reached a satisfactory training state and fulfilled the predefined training objectives.

Table 2 Comparison of model performance across different object detection algorithms for tobacco plant counting

	Model				
	SSD	Faster R-CNN	YOLO v3	YOLO v5	YOLO v7
Actual quantity	1350	1350	1350	1350	1350
Identification quantity	1212	1196	1216	1234	1265
Number of missed inspections	138	154	134	116	85
Number of false inspections	37	39	30	25	18
Correctly identify quantity	1175	1157	1186	1209	1247
Accuracy rate / %	96.95%	96.74%	97.53%	97.97%	98.58%
Precision Rate / %	87.04%	85.70%	87.85%	89.56%	92.37%

As shown in Fig. 2, the results indicate that during the training phase, the boundary box loss (box loss) steadily decreased from 0.03 to 0.01, the target detection loss (obj loss) converged from 0.04 to 0.00, and the classification loss (CLS loss) was optimized from 0.08 to 0.02. Concurrently, the detection precision and recall increased from 0.4 and 0.2 to 0.8, demonstrating an effective balance between positioning accuracy and target recognition ability in the model. In the validation phase, the losses for box, obj, and cls were stabilized at approximately 0.02, 0.01, and 0.03, respectively, with mAP@0.5 reaching 0.5, and a more stringent mAP@0.5: 0.95 at 0.3. These results confirm the model reliability in conventional detection

scenarios, although there is still room for improvement in fine-tuning the bounding box positioning.

In complex field environments, the recognition of tobacco leaves by the model is interfered by various factors, such as the background, which may lead to a decline in detection performance and increase the difficulty and complexity of tobacco leaf recognition. When validating the model detection results, it is necessary to manually check the tobacco leaf recognition and counting outcomes for each image, counting the instances of missed and false detections. In this study, manual verification is employed to ensure the accuracy of the model test results, identify issues related to missed and false detections, and guarantee the model reliability and effectiveness in practical applications. By comparing and analyzing the detection results of different models on the same field tobacco images, as shown in Tab. 2, significant differences in performance are observed among the models in terms of tobacco recognition.

As shown in Fig. 3, the recognition performance of different models varies across tobacco plants in different growth stages. Instances of missed detections and false merging of multiple plants into a single one are observed. The recognition performance of SSD, Faster R-CNN, and YOLO v3 is suboptimal, whereas YOLO v5 and YOLO v7 demonstrate better recognition accuracy, though a few cases of missed and mistaken detections still occur.

3.2 Extraction of Tobacco Number Based on Improved Model

3.2.1 YOLO v5 Improvements

In this study, an enhanced YOLO v5-EN model integrated with a channel attention mechanism was introduced. The performance of the improved model was evaluated using 30 test images collected from the field. The experimental results are shown in Tab. 3. A set of 30 representative test images was selected to ensure coverage of different growth stages and scene complexities while maintaining manual verification feasibility.

Fig. 4 illustrates the counting performance of tobacco leaf recognition. The comparison between the original and the improved model was conducted using the same image, with red boxes indicating successful recognition. The original model exhibited significant omissions in identifying certain plants. In contrast, the improved model successfully recognized the majority of plants, although a small number of missed detections still occurred.

Table 3 Performance comparison of YOLO v5 before and after improvement

Model	Actual quantity	Detected Count	Number of missed inspections	Number of false inspections	Precision rate / %	Recall rate / %	Accuracy rate / %
YOLO v5	1350	1234	116	25	97.97	91.41	89.56
YOLOv5-EN	1350	1255	95	21	98.33	92.96	91.41

As shown in Tab. 3, the improved model demonstrates a 0.36% increase in accuracy, a 1.55% increase in recall, and an overall recognition accuracy improvement of 1.85%, reaching 91.41%. This results in a substantial enhancement in the efficiency of tobacco recognition and counting. These findings indicate that the model with a channel attention mechanism significantly improves performance in addressing the issues of missed and false detections in tobacco leaf recognition.

3.2.2 YOLO v7 Improvements

In this study, the standard Convolutional Block Attention Mechanism (CBAM), Coordinated Attention Mechanism (CAM), Dynamic Rectified Linear Unit (DReLU) activation function, and Feature Refinement (FR) module were employed to enhance the original YOLO v7 en model and validate its effectiveness. Ablation experiments were conducted under identical configurations, using the same tobacco UAV images, to compare detection performance. Based on the YOLO v7 model, the improvements involving various modules and the replacement of the backbone network are introduced sequentially.

The results of the ablation experiments are presented in Tab. 4, where the performance of the basic YOLO v7 model and its modified variants on the target detection task are evaluated. The YOLO v7 model achieves an accuracy of 98.58% and the mAP of 93.45%. The processing time on the Graphics Processing Unit (GPU) is 142 MS, while the Central Processing Unit (CPU) processing time is 320 Ms. To enhance the model performance, the CBAM module was introduced, resulting in a slight increase in accuracy to 98.61% (an increase of 0.03 percentage points) and a corresponding rise in mAP to 93.75% (an increase of 0.3 percentage points). The GPU processing time increased slightly to 145 ms, while the CPU processing time significantly decreased to 294 ms, indicating that the computational efficiency was optimized alongside an improved attention mechanism. With the inclusion of the CAM module, accuracy slightly decreased to 98.52% (a reduction of 0.06%), while mAP showed a minor increase to 93.46% (an increase of 0.01%).

Table 4 Ablation study results of improved YOLO v7 with attention and refinement modules

Model	CBAM	CAM	DReLU	FR	Precision rate / %	Recall rate / %	mAP / %	GPU / ms	CPU / ms
YOLO v7	×	×	×	×	98.58	93.70	93.45	142	320
	√	×	×	×	98.61	93.93	93.75	145	294
	×	√	×	×	98.52	93.67	93.46	136	334
	×	×	√	×	98.53	93.76	93.67	143	386
	×	×	×	√	98.46	93.78	93.54	139	303
	√	×	√	√	98.82	94.51	94.10	128	258
YOLO v7-EN	×	×	×	×	98.84	96.07	95.06	151	312
	√	×	×	×	98.87	96.09	95.16	139	286
	×	√	×	×	98.85	96.08	95.08	147	294
	×	×	√	×	98.84	96.14	95.12	130	297
	×	×	×	√	98.88	96.16	95.15	145	301
	√	×	√	√	99.16	96.45	95.86	121	241

This suggests that while channel attention can enhance the model overall recognition capabilities, it may come at the cost of some accuracy. The GPU processing time reduced to 136 ms, while the CPU time increased to 334 ms. The introduction of the DReLU activation function led to a slight decrease in accuracy to 98.53% (a 0.05% reduction), but mAP increased to 93.67% (a 0.22%

improvement). This result reflects that DReLU increases computational complexity but also enhances the model ability for nonlinear processing. The addition of the FR module resulted in a decrease in accuracy to 98.46% (a 0.12% reduction), while mAP increased to 93.54% (a 0.09% increase). The GPU and CPU processing times were slightly reduced to 139 ms and 303 ms, respectively. When CBAM, DReLU, and FR were combined, accuracy significantly increased to 98.82%, and mAP rose to 94.10%, with a substantial improvement in image processing speed. The GPU and CPU processing times were reduced to 128 ms and 258 ms, respectively. These results demonstrate that the performance and efficiency of the target detection model can be significantly improved through a variety of collaborative optimization measures, particularly in terms of accuracy and processing speed for complex visual tasks.

As shown in Fig. 5, the ablation experiment of the YOLO v7-en model, which replaced the backbone network, significantly enhanced the edge recognition of tobacco leaf numbers, improving both accuracy and precision. As depicted in Fig. 6, the YOLO v7-en model achieved an accuracy of 99.16% and the mAP of 95.86% in identifying tobacco fields during both the planting and growth stages. Additionally, the processing time of the improved model on both GPU and CPU was notably reduced. The results demonstrate that, with the comprehensive integration of model variants such as CBAM, DReLU, and FR, and the enhanced backbone network, the performance of the improved model substantially outperforms the baseline model.

Table 5 Performance comparison of YOLO v7 before and after improvement

Model	Actual quantity	Identification quantity	Number of missed inspections	Number of false inspections	Precision rate / %	Recall rate / %	Accuracy rate / %
YOLO v7	1350	1265	85	18	98.58	93.70	92.37
YOLO v7-EN	1350	1302	48	11	99.16	96.45	95.63

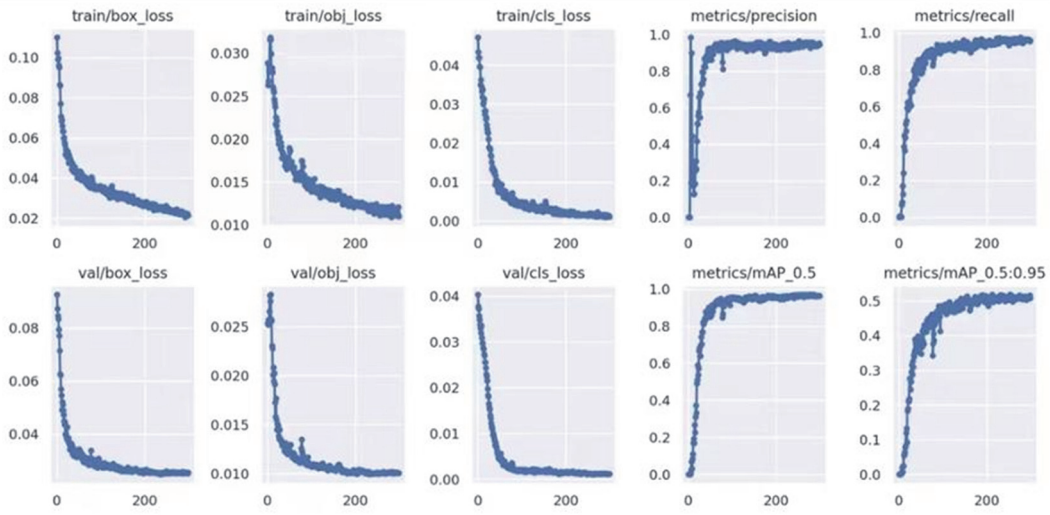


Figure 5 Effectiveness of detection of each network model

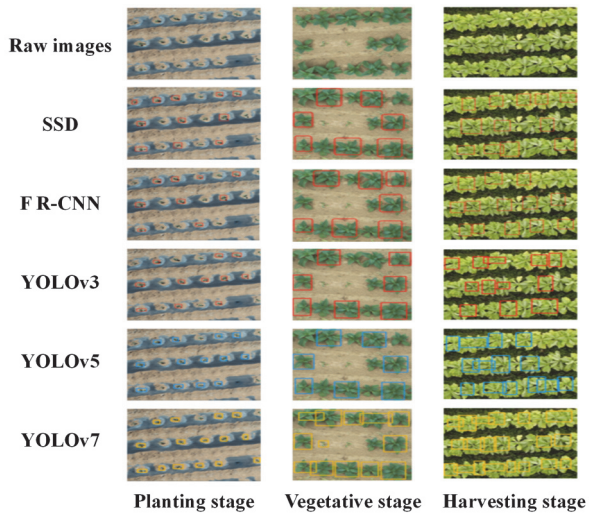
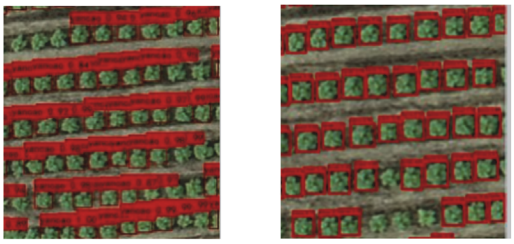
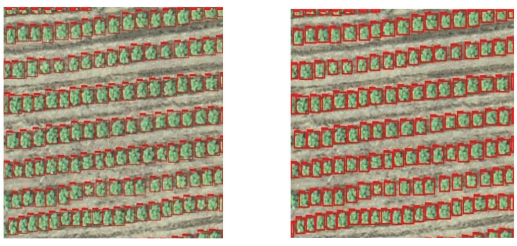


Figure 6 Detection results of tobacco plants using different base models (SSD, Faster R-CNN, YOLO v3, YOLO v5, and YOLO v7). Red bounding boxes indicate successfully detected plants. The figure highlights variations in model performance, where YOLO v5 and YOLO v7 show superior recognition with fewer missed or merged detections compared to SSD and Faster R-CNN



Recognition based on YOLO v5 model Recognition based on YOLO v5-EN model

Figure 7 Comparison of the effect of identification and counting before and after YOLO v5 model improvement



Recognition based on YOLO v7 model Recognition based on YOLO v7-EN model

Figure 8 Comparison of the effect of identification and counting before and after YOLO v7 model improvement

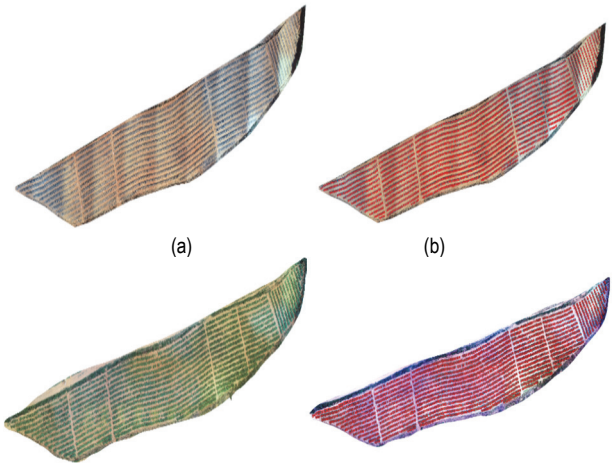


Figure 9 Comparison of the extraction effect of tobacco plants at different time periods. (a) Mosaic image during planting period; (b) Effect drawing of extraction during planting period; (c) Mosaic image of growth period; (d) Effect drawing of extraction in growth period

4 DISCUSSION

In this study, deep learning methods were applied to identify and count tobacco plants at different growth stages. It was observed that the recognition performance of the improved model on the tobacco identification dataset was enhanced. However, the dataset used in this study contains only one variety of tobacco. In contrast, different tobacco varieties, even within the same growth stage in different regions, possess unique characteristics, which may lead to deviations in accurately extracting the number of tobacco varieties.

The dataset has a relatively fixed planting density, and the row spacing of tobacco plants was manually predetermined prior to planting. As a result, it is difficult to assess the model performance when simulating tobacco planting density in different fields during the validation stage. Additionally, the study area is characterized by a limited presence of weeds and other crops near the tobacco plants, and the planting environment is minimally disturbed during the planting and growth phases. If other external factors are present, the robustness and accuracy of the tobacco recognition model may differ from actual recognition and detection scenarios.

The data augmentation approach in this study has some limitations. By introducing various rotation angles or using different branching subnet works to train the tobacco plant features from different directions, additional data augmentation techniques could be incorporated. Furthermore, the original loss function could be improved by replacing it with a unique rotation loss function, thereby enabling recognition of tobacco plants from multiple orientations.

This study focuses on only one tobacco field in the selected region. If the recognition model, designed based on the characteristics of tobacco growth in this specific field topography, is applied to other regions, its performance may be affected. Additionally, the improved model was designed specifically for tobacco; thus, if applied to other crops, further research may be required. Expanding the testing range to include the identification of other crops from UAV images will necessitate additional steps to eliminate interference factors and enhance the model accuracy and general applicability.

5 CONCLUSION

This study focused on tobacco plants at various growth stages, utilizing deep learning methods for common object recognition to compare and analyze tobacco plant counting methods. An improved YOLO model was proposed. The results demonstrated a significant improvement in both accuracy and precision of the modified YOLO model, alongside faster training speed and a considerable reduction in training time. These improvements also contributed to a reduction in labor costs for farmers, offering valuable insights for the development of smart agriculture. Future studies may explore adapting the proposed model to other crop types with similar phenotypic structures.

Acknowledgements

This research was funded by the Yunnan Provincial Agricultural Joint Special Fund - General Project, grant number 202401BD070001-068.

6 REFERENCE

[1] Shao, J., Zhang, Q., & Wang, J. (2025). Mapping and modelling impacts of tobacco farming on local higher plant diversity: A case study in Yunnan Province, China. *Geography and Sustainability*, *6*(1), 100212. https://doi.org/10.1016/j.geosus.2024.06.009

[2] Qiu, Z., Sattayakorn, N., & Pansuwan, C. (2023). The significance and impact of digital transformation on tobacco supply chain procurement in China: An empirical study. *Proceedings of the 8th International Conference on Information Systems Engineering (ICISE 2023)*, 170-176. https://doi.org/10.1145/3641032.3641041

[3] He, L., Liao, K., Li, Y., Li, B., Zhang, J., Wang, Y., Lu, L., Jian, S., Qin, R., & Fu, X. (2024). Extraction of tobacco planting information based on UAV high-resolution remote sensing images. *Remote Sensing*, *16*(2), 359. https://doi.org/10.3390/rs16020359

[4] Dobrescu, A., Giuffrida, M. V., & Tsaftaris, S. A. (2017). Leveraging multiple datasets for deep leaf counting. *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2072-2079. https://doi.org/10.1109/ICCVW.2017.243

[5] Valente, J., Sari, B., Kooistra, L., Kramer, H., & Mücher, S. (2020). Automated crop plant counting from very high-resolution aerial imagery. *Precision Agriculture*, *21*(6), 1366-1384. https://doi.org/10.1007/s11119-020-09725-3

[6] Zhu, H., Lin, C., Liu, G., Wang, D., Qiu, S., Li, A., Xu, J.-L., & Hu, Y. (2024). Intelligent agriculture: Deep learning in UAV-based remote sensing imagery for crop diseases and pests detection. *Frontiers in Plant Science*, *15*, 1435016. https://doi.org/10.3389/fpls.2024.1435016

[7] Agrawal, J. & Arafat, M. Y. (2024). Transforming farming: A review of AI-powered UAV technologies in precision agriculture. *Drones*, *8*(11), 664. https://doi.org/10.3390/drones8110664

[8] Cong, X., Li, S., Chen, F., Liu, C., & Meng, Y. (2023). A review of YOLO object detection algorithms based on deep learning. *Frontiers in Computational Intelligence Systems*, *4*(2), 17-20. https://doi.org/10.54097/fcis.v4i2.9730

[9] Bai, R., Shen, F., Wang, M., Lu, J., & Zhang, Z. (2023). Improving detection capabilities of YOLOv8-n for small objects in remote sensing imagery: Towards better precision with simplified model complexity. *Remote Sensing*, *15*(21), 5184. https://doi.org/10.3390/rs15215184

[10] Wang, X., Huang, Y., Wei, S., Xu, W., Zhao, X., Ma, J., & Chen, X. (2025). ELD-YOLO: A lightweight framework for detecting occluded mandarin fruits in plant research. *Plants*, *14*(11), 1729. https://doi.org/10.3390/plants14111729
[11] Han, W., Jiang, F., & Zhu, Z. (2022). Detection of cherry quality using YOLOv5 model based on flood filling algorithm. *Foods*, *11*(8), 1127. https://doi.org/10.3390/foods11081127
[12] Wang, D. D. & He, D. J. (2021). Channel pruned YOLO v5s-based deep learning approach for rapid and accurate apple fruitlet detection before fruit thinning. *Biosystems Engineering*, 210, 271-281. https://doi.org/10.1016/j.biosystemseng.2021.08.015
[13] Yunus, E. G. I., Hajyzadeh, M., & Eyceyurt, E. (2022). Drone-computer communication based tomato generative organ counting model using YOLO v5 and Deep-Sort. *Agriculture*, *12*(9), 1290. https://doi.org/10.3390/agriculture12091290
[14] Loddo, A. & Di Ruberto, C. (2021). On the efficacy of handcrafted and deep features for seed image classification. *Journal of Imaging*, *7*(9), 171. https://doi.org/10.3390/jimaging7090171
[15] Zhu, H., Liu, Q., Qi, Y., Hou, X., Jiang, F., & Zhang, S. (2018). Plant identification based on very deep convolutional neural networks. *Multimedia Tools and Applications*, 77, 29779-29797. https://doi.org/10.1007/s11042-017-5578-9
[16] Xu, L., Wang, Y., Shi, X., Tang, Z., Chen, X., Wang, Y., Zhang, Z., Hu, Y., Peng, H., Bi, L., Ning, Y., Li, Z., Hu, Y., & Zhang, Y. (2023). Real-time and accurate detection of citrus in complex scenes based on HPL-YOLOv4. *Computers and Electronics in Agriculture*, *205*, 107590. https://doi.org/10.1016/j.compag.2022.107590
[17] Hernández-García, A. & König, P. (2018). Data augmentation instead of explicit regularization. arXiv preprint arXiv:1806.03852.
[18] Gai, R., Chen, N., & Yuan, H. (2023). A detection algorithm for cherry fruits based on the improved YOLO-v4 model. *Neural Computing and Applications*, *35*(19), 13895-13906. https://doi.org/10.1007/s00521-021-06029-z
[19] Chen, C., Zheng, Z., Xu, T., Guo, S., Fu, S., Yu, W., & Liu, Y. (2023). YOLO-based UAV technology: A review of the research and its applications. *Drones*, *7*(3), 190. https://doi.org/10.3390/drones7030190
[20] Betti, A. & Tucci, M. (2023). YOLO-s: A lightweight and accurate YOLO-like network for small target detection in aerial imagery. *Sensors*, *23*(4), 1865. https://doi.org/10.3390/s23041865
[21] Li, J., Liu, X., Shen, S., Xu, T., Feng, J., & Yan, S. (2018). Scale-aware fast R-CNN for pedestrian detection. *IEEE Transactions on Multimedia*, *20*(4), 985-996.
[22] Wang, P., Li, S., Wang, W., Zhang, Y., Liu, Q., Chen, H., & Xu, Y. (2024). Metal defect detection models fused EfficientNet and involution. *Journal of Sensors*, *2024*(1), 6074853. https://doi.org/10.1155/2024/6074853
[23] Kang, Z., Liao, Y., Du, S., Zhang, H., Li, J., Wang, X., & Liu, Z. (2024). SE-CBAM-YOLOv7: An improved lightweight attention mechanism-based YOLOv7 for real-time detection of small aircraft targets in microsatellite remote sensing imaging. *Aerospace*, *11*(8), 605. https://doi.org/10.3390/aerospace11080605
[24] Chen, Y., Dai, X., Liu, M., Chen, D., Luo, Z., & Yuan, L. (2020). Dynamic ReLU. *Proceedings of the European Conference on Computer Vision (ECCV 2020)*, 351-367. https://doi.org/10.1007/978-3-030-58539-6_21
[25] Li, K., Wang, Y., & Hu, Z. (2023). Improved YOLOv7 for small object detection algorithm based on attention and dynamic convolution. *Applied Sciences*, *13*(16), 9316. https://doi.org/10.3390/app13169316

Contact information:

Chuanzhi MA
College of Agronomy and biotechnology,
Yunnan Agricultural University,
Kunming 650201, China

Yuehan LI
College of Water Conservancy,
Yunnan Agricultural University,
Kunming 650201, China

Yilong PENG
Faculty of Science,
St Lucia Campus the University of Queensland,
Brisbane QLD 4072, Australia

Ran WANG
College of Water Conservancy,
Yunnan Agricultural University,
Kunming 650201, China

Jiani LIU
College of Water Conservancy,
Yunnan Agricultural University,
Kunming 650201, China

Shaofan TANG
College of Water Conservancy,
Yunnan Agricultural University,
Kunming 650201, China

Fu WANG
College of Water Conservancy,
Yunnan Agricultural University,
Kunming 650201, China

Jianhua LI
(Corresponding author)
1) College of Water Conservancy,
Yunnan Agricultural University,
Kunming 650201, China
2) Luliang Mountain Basin Land Use Field Scientific Observation Station of Yunnan Province, Luliang 655600, Yunnan, China
3) Yunnan Mountain Basin Field Science Observation and Research Station, Ministry of Natural Resources of China,
Kunming 650000, China
E-mail: wenniforever@126.com