Mathematical Modeling and Algorithm Application
ISSN: 3006-0842 | Vol. 6, No. 2, 2025

Research on an Improved YOLO Model-Based Algorithm for UAV Poppy Recognition

Hu Li, Yushuang Feng*, Jiajun Su and Bo Yuan

School of Mechanical Engineering, Tianjin University of Technology and Education, Tianjin 300222, China
* **Corresponding author**: Yushuang Feng

Abstract: To address the challenges faced in drone-based poppy detection tasks, for instance small size, confusing backgrounds, and high recognition accuracy requirements, this paper come up with a poppy small target recognition method based on enhanced YOLOv9. Initially, the CAM unit was added to the backbone network, which enhances the ability of model to express fine-grained features of poppies by integrating Channel Attention and spatial Mixing mechanisms. Secondly, a content-guided attention model is incorporated into the neck of the feature pyramid, leveraging deep semantic information to adaptively weight features of different scales. This enhancement leads to an improvement in multifaceted feature integration and a substantial enhancement in the responsiveness of the detection system to diminutive objects. Finally, the issue of existing object detection loss functions not being sufficiently sensitive to bounding box regression is addressed by the introduction of the Inner IoU loss function. Finally, the issue of existing object detection loss functions not being sufficiently sensitive to bounding box regression is addressed by the creation of the Inner IoU loss as a solution. The latter focuses more on maximizing the overlap between the actual box and the predicted box, which can further improve the accuracy of position regression. Experiments conducted on a dataset of aerial photographs of poppies demonstrate that the designed algorithm enhances mAP by 13.5% and detection accuracy by 16.8% in comparison with the original YOLOv9 model. While ensuring real-time processing, it delivers higher recognition accuracy and stronger robustness, providing an effective solution for rapid and accurate poppy identification in drug control operations. Ensuring real-time processing, it delivers higher recognition accuracy and stronger robustness, offering an adequate means by which to swiftly and accurately identify poppies in drug control operations.

Keywords: Poppy detection; YOLOv9; Loss function; Small target detection; Recognition accuracy

1. Introduction

Poppies serve as a critical source of narcotics such as opium and heroin, prompting governments worldwide to impose strict legal controls on their cultivation, processing, and distribution. The consumption of illegal drugs has been demonstrated to have a detrimental effect on both personal health and social development. Consequently, the effective identification of illicit poppy cultivation represents a pivotal measure in the fight against such drug-related criminal activities. To effectively counter illegal cultivation, law enforcement agencies urgently require efficient and accurate poppy identification technologies. Traditional methods—such as manual patrols and ground-based surveillance—suffer from high time costs, limited coverage, and significant manpower demands, rendering them inadequate for rapid large-scale and multi-location detection. Consequently, computer vision and UAV aerial photography-based poppy recognition algorithms have emerged as a pivotal research direction.

Currently, UAV-based vegetation recognition systems predominantly rely on traditional image processing and machine learning algorithms. While these methods demonstrate certain effectiveness in simple environments, their accuracy and efficiency often fall short of practical requirements when confronted with complex terrains, vegetation coverage, and varying illumination conditions. The primary challenges in UAV-based poppy detection include: (1) Low recognition accuracy due to the small and scattered nature of poppy targets; (2) Difficulty in distinguishing poppies from vegetables and ornamental flowers cultivated in proximity; (3) Variable environmental conditions—such as illumination intensity, weather, and seasonal changes—that substantially increase image recognition complexity. Thus, achieving precise and efficient identification of poppy plants remains a significant research challenge.

Many scholars have applied object detection to the identification of UAV aerial images. Scholars such as Xiangyu Liu[1] proposed an recognition method based on the SSD model, which uses satellite remote sensing images to identify the locations of poppy fields with a detection accuracy of 95%. However, this approach primarily focuses on large-scale poppy field identification and does not address the issue of feature extraction for small targets. Scholars such as Kody R. Rominger[2] improved the YOLOv5 model to identify and count individual dwarf bear poppy plants at an altitude of 40 meters, achieving an average precision of 0.83. However, the experiment was conducted in open fields for poppy identification, and its capability for local feature extraction in occluded scenarios requires further enhancement. Jun Zhou et al.[3] employed the MobileNetv2 backbone network and introduced the Generalized Intersection over Union (GIoU) loss function, which increased the average precision by 1.62%. Nevertheless, when detecting small targets, this method exhibits errors between predicted results and ground truth labels, leading to information loss. Matija Radovic et al.[4] proposed a CNN method suitable for object detection and classification in aerial images, achieving a detection accuracy of 97.8%. However, the detection speed is relatively slow, significantly lower than that of the YOLO algorithm. Wang Qian et al.[5] proposed a two-stage object detection method that reduced the number of falsely detected images by 73.88%. Despite

this improvement, the detection speed remains slow and fails to meet real-time detection requirements. Zhang Z et al.[6] introduced a poppy detection network named YOLOHLA, which uses two branches to extract features at different scales. The method achieved a mean average precision (mAP) of 88.2%. However, it is prone to false detections in densely vegetated areas. Chunshan Wang et al.[7] utilized ResNeXt group convolution to reduce the number of model parameters and added an ASPP module before the small-scale detection box, resulting in a 0.44% improvement in mAP. However, the method performs poorly in scenarios with dense and occluded targets, leading to detection errors.

The proposed model is an enhancement to the existing small target recognition model based on YOLOv9, specifically designed for the task of poppy identification in drone aerial photography images, with the objective of enhancing the accuracy of drone recognition of targets in aerial photography images while reducing false positive and false negative rates. The innovation of the research lies in:

1) The model network has been augmented with CAM, a technique that integrates global and local features, with the objective of enhancing the model's noise reduction performance and feature extraction capabilities for aerial images.

2) The inner-IoU loss function is to be designed in order to improve the model's detection accuracy and convergence speed, reduce the error between the predicted results and the true labels, and improve detection performance.

3) The incorporation of a content-guided attention mechanism within the model's neck facilitates the construction of a spatial weight distribution map, thereby guiding the model to prioritize the extraction of useful information from the image. This enhancement in the model's capacity to recognize small-sized targets is particularly effective in addressing issues such as target occlusion and high density.

2. Algorithmic model

2.1. YOLOv9 Model

The YOLOv9 model, introduced in 2024 by the original YOLOv7 research team, represents the latest advancement in the You Only Look Once (YOLO) series of object detection frameworks. YOLOv9 conducts an in-depth investigation into information bottlenecks and reversible functions, introducing the concept of Programmable Gradient Information (PGI)[8] and devising a novel Generalized Efficient Layer Aggregation Network (GELAN) architecture

Programmable Gradient Information (PGI) enables deep networks to adapt to the diverse variations required for multi-target detection by providing complete input information for computing objective functions. This ensures reliable gradients for updating network weights. PGI comprises three key components: (1) A main branch dedicated to the inference process; (2) This auxiliary reversible branch has been specifically designed to address the challenges posed by the increasing depth of neural networks; (3) Multi-level auxiliary information[9] that effectively mitigates error accumulation in deep supervision schemes—particularly critical for networks with multiple pathways and lightweight models.

The Generalized Efficient Layer Aggregation Network (GELAN)[10] is a synthesis of the merits of the Cross Stage Partial Network (CSPNet) and the Efficient Layer Aggregation Network (ELAN), designing a flexible as well as efficient framework for the integration of features. This novel framework reduces parameter volume, enhances feature reuse, and improves the model's feature representation capacity. Compared to alternative approaches, GELAN achieves superior parameter utilization using only conventional convolution operators while extending ELAN's functionality—originally limited to stacked convolution layers—to a generalized architecture compatible with any computational blocks. The YOLOv9 network architecture is illustrated in Figure 1 below.

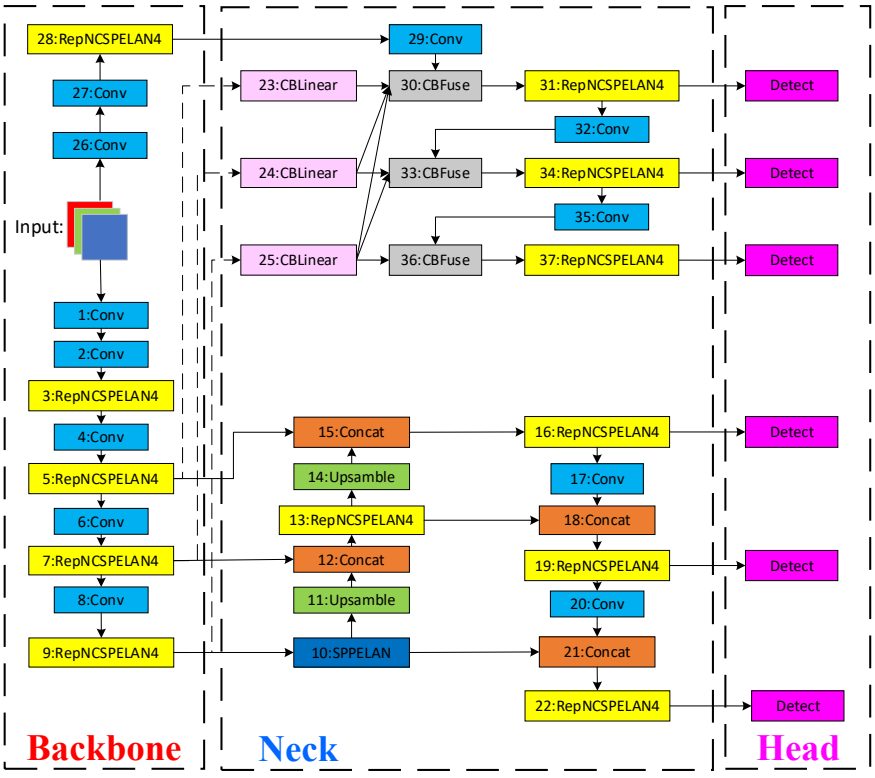


Figure 1. YOLOv9 Network architecture

2.2. Enhanced YOLOv9 model

The UAV poppy dataset employed in this study contains poppy samples of varying scales against complex backgrounds. The original YOLOv9 model demonstrates suboptimal recognition accuracy for small poppy targets in aerial imagery, exhibiting frequent false positives and missed detections. The original YOLOv9 model has a low accuracy rate for identifying small poppy targets in aerial images, with false positives and false negatives occurring. In response to the aforementioned challenges, the paper implements architectural enhancements to the YOLOv9 algorithm, proposing the HCU-YOLOv9 network model.

Firstly, this paper proposes an optimization of the backbone structure of YOLOv9 by integrating convolution and attention mechanisms. The approach involves the integration of a convolutional attention mixing (CAM) unit in sequence with the characteristic acquisition module RepNCSPELAN4, with the objective of increasing the model's capacity to extract both global and local features.

Secondly, a Content-guided Attention (CGA) model is introduced. This model dynamically draws a spatial importance map (SIM) through the utilization of adaptive learning feature tensors. The CGA assigns a designated SIM to individual channels, thereby enabling the focus to be directed towards specific areas within the channel.

The system utilizes multi-dimensional representations of input features as guidance signals, employing a parametric mapping function to independently generate spatial weight distribution maps for each feature channel. This approach amplifies semantically rich information within feature maps, significantly enhancing the discriminative power of feature representations.

Regarding the loss function, this paper innovatively replaces Complete Intersection over Union Loss (CIoU) with Inner-IoU. This approach prioritizes weight allocation across prediction boxes of varying quality, dynamically shifting the model's focus from low-quality predictions to high-quality ones to enhance detection performance. The architecture of MCI-YOLOv9 is illustrated in Figure 2:

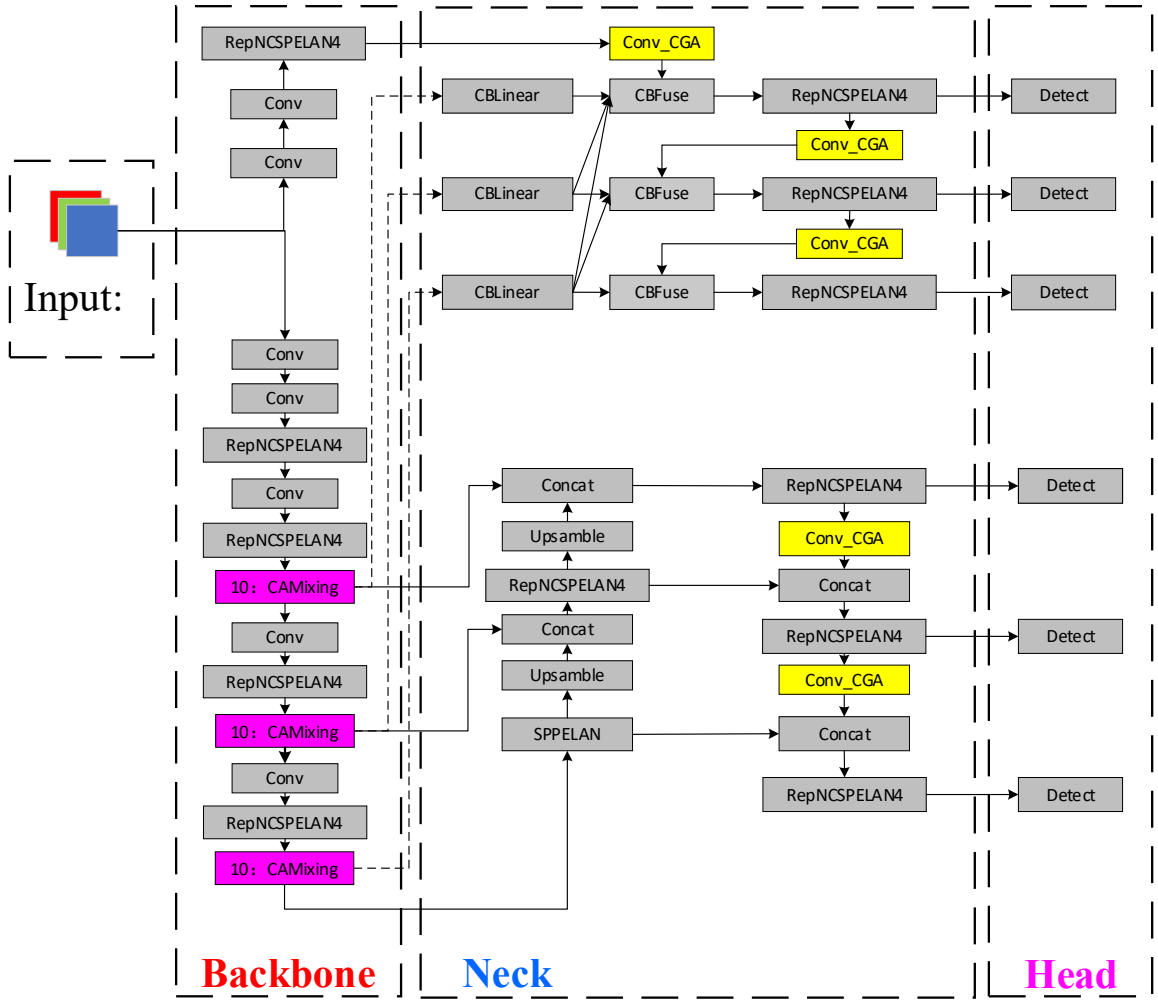


Figure 2 MCI-YOLOv9 Network

2.2.1. CAM module

To address the low recognition accuracy of small targets in aerial imagery with the YOLOv9 model, this paper introduces the Convolutional Attention Mixing (CAM) module[11]. The CAM module comprises two parts: Convolution-attention Fusion Module (CAFM) and Multi-scale Feed-forward Network (MSFN). The CAFM has both regional and global branches. The regional branch employs 1×1 convolution operations integrated with a channel shuffle mechanism to perform feature transformation. Specifically, channel shuffling first divides the feature tensor of the input image into multiple regions going in the direction of the channels. It then applies convolution to each region to facilitate cross-channel information exchange. The system concatenates the

output feature tensors of each part in the channel direction after completing the region segmentation process, forming a new feature representation. To further enhance feature extraction capability, we introduce 3×3×3 three-dimensional convolutional kernels for spatial feature learning. The calculation process for this region branch might be mathematically expressed as follows:

$$F_{conv} = W_{3\times3\times3}(CS(W_{1\times1}(Y))) \tag{1}$$

F_{conv} indicates the outcome of the regional branch., $W_{1\times1}$ represents convolution, $W_{3\times3\times3}$ denotes the weights of a 3×3×3 convolutional kernel, CS is short for channel switching operation, and Y indicates the input information. In the context of the branch as a whole, 1×1 convolutions and 3×3 depthwise convolutions first generate Query (Q), Key (K), and Value (V) tensors, yielding three tensors of dimension $\hat{H}\times\hat{W}\times\hat{C}$. Subsequently, Q convert into $\hat{Q}\in\mathbb{R}^{\hat{H}\hat{W}\times\hat{C}}$, K transform into $\hat{K}\in\mathbb{R}^{\hat{C}\times\hat{H}\hat{W}}$ [12], followed by standard attention computation $A\in\mathbb{R}^{\hat{C}\times\hat{C}}$ through $\hat{Q}$ and $\hat{K}$ interaction. The output F_{att} of the global branch is formally expressed as:

$$F_{att} = W_{1\times1}Attention(\hat{Q},\hat{K},\hat{V}) + Y \tag{2}$$

The computational formulation of the CAFM module is mathematically expressed as:

$$F_{out} = F_{att} + F_{conv} \tag{3}$$

Traditional fully connected feedforward networks (FFNs) comprise two dense layers that aggregate single-scale information. Nevertheless, the content included in single-scale information aggregation is incomplete. To promote nonlinear transformation of information, a new multi-scale network for performing forward feedback has been proposed (MSFN). Following CAM structuring, the CAFM results are entered into the MSFN, where they are processed through two parallel paths. A gating mechanism integrates these pathways to strengthen nonlinear transformations. The lower path employs depthwise convolution for local feature extraction. The upper layer uses two 3×3 dilated convolutions to obtain multi-scale feature information.

The MSFN operation is formally expressed as:

$$Gating(X) = \phi(W_{3\times3\times3}W_{1\times1}(X)) \odot (W^2_{3\times3}W_{1\times1}(X) + W^3_{3\times3}W_{1\times1}(X)) \tag{4}$$

$$X_{out} = W_{1\times1}Gating(X) \tag{5}$$

$\odot$ denotes element-wise multiplication, ϕ represents a nonlinear activation function , $W^2_{3\times3}$ indicates a 3×3 empty convolution, with the 2 in the upper right corner being the dilation rate, $W^3_{3\times3}$ denotes a 3×3 Atrous Convolution and expansion rate is three.

Figure 3 illustrates the network architecture of the CAM module.

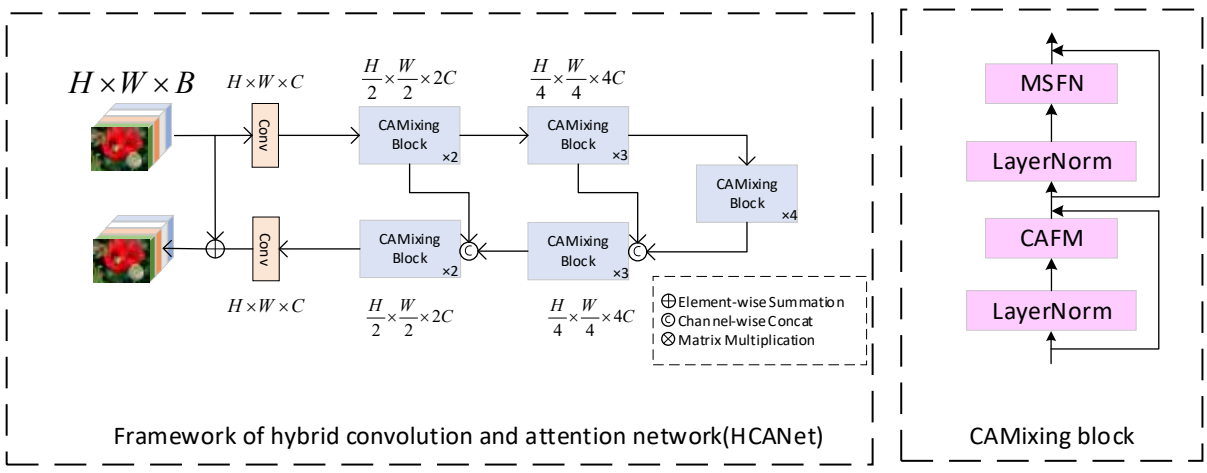


Figure 3 CAM Module Architecture Diagram

2.2.2. Content-Guided Attention (CGA) model

In order to address feature-level distribution imbalances in attention modules, this paper introduces CGA, which acquires channel-specific Spatial Importance Maps (SIM) for each input property data channel in a coarse-to-fine way. Simultaneously, it comprehensively combines channel and spatial attention weights[13] to ensure the reliability of information exchange.

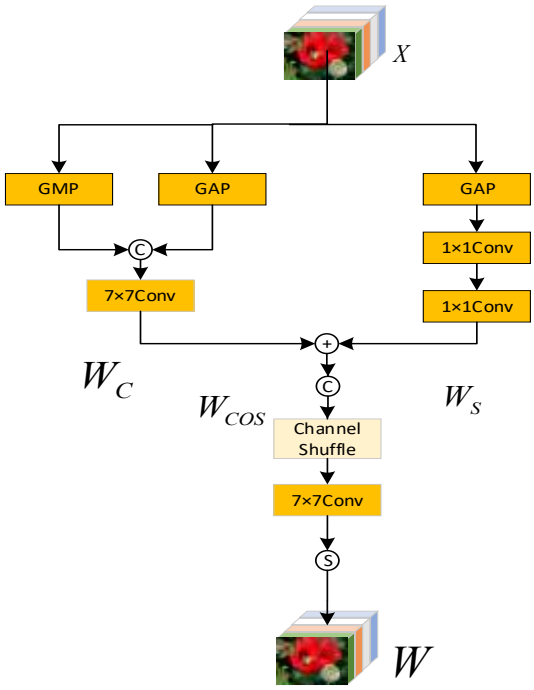


Figure 4 CGA Module Structure

The components of CGA are shown in Figure 4, $X \in \mathbb{R}^{C\times H\times W}$ represents the input feature, The aim of CGA is to produce channel-specific SIM. ($W \in \mathbb{R}^{C\times H\times W}$) It has the same dimension as X. The calculation formula of W_C, W_S in the figure:

$$W_C = Conv_{1\times1}(\max(0, Conv(X^C_{GAP}))) \tag{6}$$

$$W_S = Conv_{7\times7}([X^s_{GAP}, X^s_{GMP}]) \tag{7}$$

In the formula, $\max(0,c)$ represents the ReLU activity function, $Conv_{1\times1}(\bullet)$ denotes a convolution operation[14], with a convolution core size of 1×1., X^c_{GAP}、X^s_{GAP}、X^s_{GMP} denote features obtained by performing global average pooling[15] in the spatial direction, global average pooling along the channel direction, and global max pooling operations, respectively. The features W_C and W_S are then fused to generate a coarse Spatial Importance Map (SIM), formally expressed as:

$$W_{cou} = W_C + W_S \tag{8}$$

Leveraging the content of input features as guidance, the final channel-specific Spatial Importance Map (SIM)[16] is formed by alternately interleaving the channels of x and y through channel shuffle operations, formally expressed as:

$$W = \sigma(GConv_{7\times7}(CS([X, W_{coa}]))) \tag{9}$$

In the formula, σ denotes the sigmoid function, CS is short for channel switching operation, $GConv_{7\times7}(\bullet)$ denotes a group convolutional layer with a kernel size of 7×7. The CGA mechanism draws a different SIM for different channels, directing the model to focus its attention on key areas in each channel. This emphasizes more discriminative information within features, thereby improving the discriminative power of model.

2.2.3. Improvement of the loss function

The loss function of YOLOv9 is CIOU Loss[17], which considers the relationship between the center point of the bounding box and its length as well as width. The functions of IoU together with loss L_{CIoU} are shown in formulas (10) and (11) respectively.

$$IoU = \frac{|B \cap B^{gt}|}{|B \cup B^{gt}|} \tag{10}$$

$$L_{CIoU} = 1 - IoU + \frac{\rho^2(b, b^{gt})}{C^2} + \alpha v \tag{11}$$

In the formula, B and B^{gt} refer to two different bounding boxes: one is predicted and the other is actual; b and b^{gt} denote the center locations of the bounding boxes, one is predicted, and the other is actual; ρ indicate the distance between the different core points; α represent the weighting coefficient; v denotes how consistent the width-to-height ratio is between two different bounding boxes.

Notices (12) and (13) are the definitions of α and v[18]:

$$\alpha = \frac{v}{(1 - IoU) + v} \tag{12}$$

$$v = \frac{4}{\pi^2}(\arctan\frac{w^{gt}}{h^{gt}} - \arctan\frac{w}{h})^2 \tag{13}$$

However, due to significant size variations and uneven distribution of aerial poppies, the CIoU loss function exhibits weak generalization capability, adversely affecting poppy recognition accuracy. To address this limitation, Inner-IoU designed a scaling factor to adjust the dimensions of auxiliary bounding boxes, which can improve the model's generalization ability. For high-IoU samples, the absolute gradient value of IoU for smaller-scale auxiliary boxes exceeds that of the actual box. Conversely, for low-IoU samples, larger-scale auxiliary boxes yield greater absolute IoU gradients than the actual box. The findings of the analysis indicate that the utilization of smaller auxiliary containers for the calculation of IoU loss can accelerate the regression process for specific samples, which generally exhibit elevated intersection-over-union values. This enables the algorithm to converge more quickly. Using bigger auxiliary boxes when calculating samples with low IoU can speed up convergence. The formulation is expressed as:

$$b^{gt}_t = x^{gt}_c - \frac{w^{gt} * k_r}{2}, \quad b^{gt}_r = x^{gt}_c + \frac{w^{gt} * k_r}{2} \tag{14}$$

$$b^{gt}_t = y^{gt}_c - \frac{h^{gt} * k_r}{2}, \quad b^{gt}_t = y^{gt}_c + \frac{h^{gt} * k_r}{2} \tag{15}$$

$$b_l = x_c - \frac{w * k_r}{2}, \quad b_l = x_c + \frac{w * k_r}{2} \tag{16}$$

$$b_t = y_c - \frac{h * k_r}{2}, \quad b_t = y_c + \frac{h * k_r}{2} \tag{17}$$

$$inter = (min(b_r^{gt}, b_r)) - max(b_l^{gt}, b_l) * (min(b_b^{gt}, b_b) - max(b_t^{gt}, b_t)) \tag{18}$$

$$union = (w^{gt} * h^{gt}) * (k_r)^2 + (w * h) * (k_r)^2 - inter \tag{19}$$

$$IoU^{inner} = \frac{inter}{union} \tag{20}$$

$$L_{Inner-IoU} = 1 - IoU^{inner} \tag{21}$$

b_t^{gt} and b means the target box and the anchor box respectively; (x_c^{gt}, y_c^{gt}) denotes the X- and Y-coordinates of the core location inside the target box; (x_c, y_c) indicates the position of the anchor box core.; w^{gt} stand that is equivalent in width to that of the target box and h^{gt} represents its height. w denotes the width of the anchor frame and h denotes its height; k_r is the corresponding scaling factor, and its value range is usually $[0.5,1.5]$. The schematic diagram is shown below.

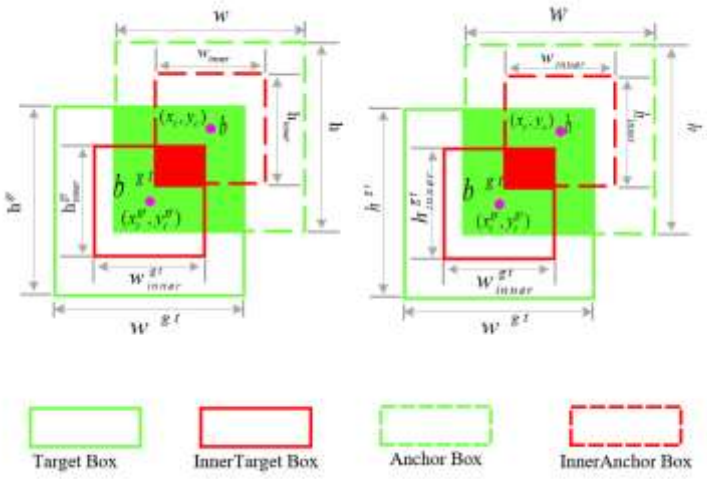


Figure 5 Inner-IoU diagram

3. Experiments and Analysis

3.1. Experimental Datasets

Cultivating poppies is strictly prohibited in China, and unauthorized poppy cultivation constitutes illegal activity. To evade supervision, growers typically employ highly concealed methods such as scattered planting, single-plant cultivation, intercropping with vegetables, and potted plant camouflage. These covert practices significantly increase identification difficulties, posing substantial challenges to anti-drug efforts. From March to July annually—the peak period for counter-narcotics field surveys—law enforcement agencies deploy drones for poppy monitoring operations. Compared to satellite remote sensing, drone-based surveillance offers greater operational flexibility and lower costs, significantly enhancing the efficiency of anti-drug inspections.

Individual poppy plants exhibit a diameter of approximately 30 cm. At low altitudes (relative height<50 meters), poppy textures are clearly observable. Between 50-100 meters, individual poppies and smaller features remain distinguishable. However, at altitudes >100 meters, distinctive features become severely blurred, leaving only color characteristics and observer experience as identification cues. During actual image capture, scenarios include both images containing multiple poppy plants and frames with isolated poppies positioned in corners. To enhance model training efficacy, images were segmented and cropped based on poppy locations, ensuring each processed image contained at least one poppy plant. Non-poppy images were filtered out, and all images were resized to a standardized 640×640 resolution. A total of 1,289 photos were obtained: 334 photos were taken at 30 meters, 432 at 60 meters, and 413 at 150 meters.

To further expand the dataset scale, this study employs randomized data augmentation techniques. First, perform horizontal mirror flipping and vertical flipping on the images; second, adjust the color, brightness, and contrast of the pictures; finally, use the mosaic method to zoom the images. The employment of these methodologies serves to improve the model's capacity for feature extraction and its generalization abilities, consequently enhancing the accuracy of detection. The final augmented dataset was partitioned using stratified random sampling: 80% for training and 20% for validation. Representative aerial poppy images are illustrated in Figure 6.



Figure 6.1 Poppy (growing period)



Figure 6.2 Poppy flower (flowering period)



Figure 6.3 Poppy (fruiting period)

Figure 6 Partial dataset of poppy plant images

3.2. Experimental environment and metrics

The present study evaluates the efficacy of the algorithm by employing accuracy and inference speed as the primary performance indicators. The value of mean average precision (mAP) is indicative of the detection accuracy, whilst mAP reflects the mAP value when the confidence threshold is set to 0.5[19]. The mAP value is obtained by adding together the average accuracy rates of all categories and then dividing this sum by the number of categories. For each class, the AP corresponds to the area under the Precision-Recall (P-R) curve, derived from the class-specific precision and recall values. The computational methods are formalized in Equations (10), (11), and (12) respectively.

$$Precision = \frac{TP}{TP + FP} \tag{22}$$

$$Recall = \frac{TP}{TP + FN} \tag{23}$$

$$mAP = \frac{\sum AP}{N(Class)} \tag{24}$$

When a sample tests positive, the model also predicts a positive result. This is known as a true positive (TP). False positive (FP) occurs when a sample tests positive but the prediction of model is incorrect. Conversely, when a negative sample is detected and the model correctly predicts a negative result, this is known as a true negative (TN). The model forecasts incorrectly when testing negative samples, resulting in a false negative (FN).

Metrics for evaluating model complexity and detection speed include Parameters (total trainable weights), GFLOPs (operations involving floating point numbers), and Frames Per Second (FPS). These parameters can effectively reflect the performance of the algorithm. The more complex the model architecture, the higher its parameter count and GFLOPs, resulting in significantly reduced FPS. Consequently, under identical hardware conditions, both training and inference speeds decrease. The FPS is a measurement of the number of images processed in a second—higher FPS values indicate faster detection speeds[20]. Three primary factors affect FPS: Image preprocessing time, Model inference time, Non-Maximum Suppression (NMS) processing time[21], with model inference time being the most significant contributor.

3.3. Experimental Environment

The experiments reported in this paper were conducted using a system equipped with a central processing unit (CPU) of 14 vCPU Intel(R) Xeon(R) Gold 6330 CPU @ 2.00GHz, and running on the 64-bit Ubuntu 20.04 operating system. The graphics processing unit (GPU) used for accelerating training was a GeForce RTX 3090 with 24GB of video memory, and CUDA version 11.3. The deep learning framework used was PyTorch version 1.10.0. The training cycle comprises 250 epochs, and each batch contains 16 images.

3.4. Ablation Study on Model Performance

The objective of this paper is to verify the efficacy of the MCI-YOLOv9 model in detecting small targets in aerial imagery through the implementation of ablation studies on enhanced modules. The experimental procedure is as follows: First, sequentially introduce the CAM and Conv_CGA modules based on the original algorithm to enhance feature fusion capability; Ultimately, the loss function was refined to a more appropriate Inner-IoU; Performance comparisons between this improved model and the baseline network are conducted using mAP@0.5, precision, parameter count, and FPS. The symbol '√' in the table means that the module has been added. The enhancement outcomes are exhibited in Table 1.

Table 1 Ablation Experiment

Network Model	CAM	Conv_CGA	Inner-IoU	mAP@0.5(%)	precision	Parameters（×106）	FPSf(/s)
a				69.7	0.727	60.49	80.82
b	√			72.3	0.814	54.77	66.10
c	√	√		75.6	0.831	56.71	74.23
e	√	√	√	83.2	0.895	56.75	81.04

Experimental data based on the YOLOv9 baseline model: Incorporating the CAM module into the backbone network elevates mAP@0.5 to 72.3%, demonstrating enhanced global and local feature extraction capabilities. Concurrently, Model b achieves a 9.4% reduction in parameter count, showcasing CAM's efficacy in model compression through channel shuffling. After adding the Content-Guided Attention (CGA) mechanism, model c was formed. Compared to model b, the mAP@0.5 shows an increase of 4.56%, the recognition precision increased by 2.03%, a marginal rise of 3.54% was

observed in the number of parameters, and the FPS increased by 12.29%. After improving the loss function, model e was generated. For model e, the mAP@0.5 shows an increase of 10.05%, the accuracy increased by 7.7%, the number of parameters remained unchanged, the FPS improved by 8.4%, and the optimized loss function significantly enhanced the convergence speed of the model. In summary, the effectiveness and superiority of this algorithm in identifying poppies using drone aerial photography has been proven.

3.5. Algorithm Comparison Experiment

The following study aims to provide further validation of the enhanced algorithm model's superior performance; comparative experiments were conducted between our proposed MCI-YOLOv9 network and various contemporary YOLO versions on this dataset. The primary algorithms compared include YOLOv5, YOLOv7, YOLOv8, YOLOv9, YOLOv10, YOLOv11, and YOLOv12, with the results presented in Table 2.

Table 2 Comparison Experiment

Network Model	P	R	mAP@0.5(%)	FPS(f/s)
YOLOv5	0.660	0.566	73.63	81.26
YOLOv7	0.710	0.574	74.50	75.43
YOLOv8[22]	0.727	0.574	68.88	135.89
YOLOv9	0.736	0.597	69.72	80.82
YOLOv10	0.685	0.611	65.62	303.03
YOLOv11	0.7482	0.5833	70.90	312.50
YOLOv12	0.8295	0.6678	79.98	243.90
MCI-YOLOv9	0.8955	0.6908	83.22	81.04

On the UAV aerial poppy dataset, the proposed MCI-YOLOv9 model achieves 83.22% recognition accuracy—significantly outperforming comparative models. It surpasses the closest competitor YOLOv12 by 3.24% and exceeds the baseline YOLOv9 by 13.5%, demonstrating superior target localization and classification capabilities. Particularly under the stringent IoU=0.5 evaluation criterion, MCI-YOLOv9 exhibits enhanced detection performance for small and occluded targets.

While its inference speed (FPS) remains moderate, it fully satisfies real-time monitoring requirements. Notably, MCI-YOLOv9 reduces parameter count substantially versus YOLOv9, reflecting exceptional performance balance. Compared to alternative object detection algorithms, MCI-YOLOv9 delivers higher precision, making it better suited for the high-precision demands of poppy detection tasks.

3.6. Visual comparison experiment

This paper selects images from a variety of situations for comparative analysis, encompassing challenging scenarios such as occlusion, shadowing, and scattered planting patterns. These images were processed through both the trained baseline YOLOv9 model and the proposed MCI-YOLOv9 model for visual comparison. The comparative results are presented in Figure 6.



(a)YOLOv9 (b) MCI-YOLOv9

Figure 7 Comparison of false detection of similar plants



(c) YOLOv9 （d) MCI-YOLOv9

Figure 8 Comparison of missed detection of dense targets



(e) YOLOv9 (f) MCI-YOLOv9

Figure 9 Comparison of missed detection of occluded targets

As demonstrated in Figures 7, 8, and 9, the efficacy of the YOLOv9 algorithm and the proposed MCI-YOLOv9 algorithm was compared through the utilization of aerial images containing diminutive targets across a range of backgrounds. Comparative analysis of Figure (a) and (b) reveals that the YOLOv9 model erroneously identifies botanically similar plants as poppies, while MCI-YOLOv9 demonstrates superior identification accuracy. Comparative analysis of Figure (c) and (d) reveals that when poppy targets exhibit incomplete features due to occlusion, the YOLOv9 model fails to detect them. Simultaneously, Figures (e) and (f) demonstrate that against backgrounds containing morphologically and chromatically similar crops, YOLOv9 also suffers from missed detections. In contrast, the MCI-YOLOv9 algorithm achieves effective recognition in both scenarios, validating its superior detection capabilities for small aerial poppy targets. This approach has been shown to significantly reduce false positives and false negatives while achieving higher levels of detection precision.

4. Conclusion

To address the low recognition accuracy in anti-drug UAV surveillance—caused by building occlusions, crop obstructions, challenges in small-target identification, and diverse phenotypic characteristics—this paper proposes the enhanced MCI-YOLOv9 model. MCI-YOLOv9 achieves superior precision in multi-feature target recognition, small-object detection, and occluded plant identification, establishing an optimized equilibrium between recognition accuracy and operational speed. The principal findings are summarized as follows:

(1) Acknowledging the diverse phenotypic characteristics of poppy plants, this paper innovatively implements comprehensive optimization of the backbone network within the original YOLO architecture. By integrating offsets and learnable parameters into convolutional layers, we successfully construct an advanced CAM module. The resulting network structure enhances the model's noise suppression capability and feature extraction capacity for targets with diverse morphologies and scales.

(2) A Content-Guided Attention (CGA) mechanism was incorporated into the algorithm's neck network to amplify channel weights for small-target features, thereby enhancing detection accuracy for low-resolution images and small target. This significantly boosts model performance in highly complex recognition scenarios.

(3) By introducing the Inner-IoU loss function, this study effectively resolves the limitation of weak generalization capability in prior loss functions—caused by significant size variations and uneven distribution of aerial poppies—which adversely impacted recognition accuracy. Inner-IoU adjusts the scale of additional bounding boxes by modifying the scaling factor, thereby improving the model's capacity for generalization. This innovation accelerates convergence speed while boosting detection accuracy.

In summary, this study addresses the complex environmental demands of UAV-based poppy detection through targeted optimization and enhancement of existing algorithmic models, proposing the MCI-YOLOv9 framework. MCI-YOLOv9 demonstrates exceptional performance in small-target aerial recognition, significantly enhancing the identification and classification accuracy of poppy seedlings, flowers, and seed pods for anti-drug inspections. These advancements not only elevate operational efficiency in law enforcement but also provide critical theoretical frameworks and practical methodologies for related technological development.

References

[1] Zhou J, Tian Y, Yuan C, et al. Improved UAV opium poppy detection using an updated YOLOv3 model[J]. Sensors, 2019, 19(22): 4851.

[2] Radovic M, Adarkwa O, Wang Q. Object recognition in aerial images using convolutional neural networks[J]. Journal of Imaging, 2017, 3(2): 21.

[3] Zhang Z, Xia W, Xie G, et al. Fast opium poppy detection in Unmanned Aerial Vehicle (UAV) imagery based on deep neural network[J]. Drones, 2023, 7(9): 559.

[4] Liu X, Tian Y, Yuan C, et al. Opium poppy detection using deep learning[J]. Remote Sensing, 2018, 10(12): 1886.

[5] Rominger K R, Meyer S E. Drones, deep learning, and endangered plants: A method for population-level census using image analysis[J]. Drones, 2021, 5(4): 126.

[6] Wang Q, Wang C, Wu H, et al. A two-stage low-altitude remote sensing papaver somniferum image detection system based on YOLOv5s+ DenseNet121[J]. Remote Sensing, 2022, 14(8): 1834.

[7] Rominger K, Meyer S E. Application of UAV-based methodology for census of an endangered plant species in a fragile habitat[J]. Remote Sensing, 2019, 11(6): 719.

[8] Zou C, Jeon W S, Rhee S Y. Research on the multiple small target detection methodology in remote sensing[J]. Sensors, 2024, 24(10): 3211.

[9] Qin W, Yang X, Wang Y, et al. YOLOPoul: Performance evaluation of a novel YOLO object detectors benchmark for multi-class manure identification to warn about poultry digestive diseases[J]. Smart Agricultural Technology, 2025: 101145.

[10] Ye R, Gao Q, Qian Y, et al. Improved yolov8 and sahi model for the collaborative detection of small targets at the micro scale: A case study of pest detection in tea[J]. Agronomy, 2024, 14(5): 1034.

[11] Song X, Yan L, Liu S, et al. Agricultural Image Processing: Challenges, Advances, and Future Trends[J]. Applied Sciences, 2025.

[12] Sun Z, Yang H, Zhang Z, et al. An improved YOLOv5-based tapping trajectory detection method for natural rubber trees[J]. Agriculture, 2022, 12(9): 1309.

[13] Nedungatt S, KS B, Lal S. MDEANet: modified detail-enhanced convolution and attention-based network for dehazing of remote sensing images[J]. Multimedia Tools and Applications, 2025, 84(18): 18943-18966.

[14] Yuan M, Meng H, Yan T, et al. Fa YOLO: fog-aware instance segmentation of ships with feature refinement in foggy scenes: M. Yuan et al[J]. Multimedia Systems, 2025, 31(4): 291.

[15] Chen Z, He Z, Lu Z M. DEA-Net: Single image dehazing based on detail-enhanced convolution and content-guided attention[J]. IEEE transactions on image processing, 2024, 33: 1002-1015.

[16] Yekeben Y, Cheng S, Du A. CGFTNet: Content-Guided Frequency Domain Transform Network for Face Super-Resolution[J]. Information, 2024, 15(12): 765.

[17] Sun Z, Yang H, Zhang Z, et al. An improved YOLOv5-based tapping trajectory detection method for natural rubber trees[J]. Agriculture, 2022, 12(9): 1309.

[18] Yu Z, Liu Y, Yu S, et al. Automatic detection method of dairy cow feeding behavior based on YOLO improved model and edge computing[J]. Sensors, 2022, 22(9): 3271.

[19] Dong L, Zhu H, Ren H, et al. Developing YOLOv5s model with enhancement mechanisms for precision parts with irregular shapes[J]. Advanced Engineering Informatics, 2025, 65: 103257.

[20] Xue X W, Dai Q, Zhao M. SiamMask RAM for Underwater Target Tracking in Sonar Images[C]Journal of Physics: Conference Series. IOP Publishing, 2024, 2891(3): 032031.

[21] Yin Y, Zhang Z, Wei L, et al. Pedestrian detection algorithm integrating large kernel attention and YOLOV5 lightweight model[J]. PLoS one, 2023, 18(11): e0294865.

[22] Chen B, Ding F, Ma B, et al. A method for real-time recognition of safflower filaments in unstructured environments using the YOLO-SaFi model[J]. Sensors, 2024, 24(13): 4410.