ICJK
Chinese Journal of Information Fusion
http://dx.doi.org/10.62762/CJIF.2026.233582

RESEARCH ARTICLE

Check for updates

A Multi-Scale UAV Cluster Detection and Counting Framework Based on Multi-Source Information Fusion and Enhanced YOLOv9

Xinpeng Lu[1], **Shijie Chen**[1], **Jiashuo Sun**[2], **Zhichao Yin**[3] **and Moufa Hu**[3,*]

[1] College of Electronic Science, National University of Defense Technology, Changsha 410073, China
[2] College of Advanced Interdisciplinary Studies, National University of Defense Technology, Changsha 410073, China
[3] Key Laboratory of Automatic Target Recognition, National University of Defense Technology, Changsha 410073, China

Abstract

To address the safety hazards posed by unauthorized "black flight" of unmanned aerial vehicle (UAV) clusters, as well as the limitations of existing detection methods in tiny target recognition and complex environment adaptability, this paper proposes a real-time detection and counting method based on a feature fusion framework. The core contribution of this work is a novel multi-level fusion architecture that aggregates multi-scale visual features to enhance detection robustness under challenging conditions. First, we construct a multi scene UAV data set containing 7,113 images, and design a dedicated feature fusion module based on improved YOLOv9 backbone. The Haar Wavelet Down-sampling (HWD) module is introduced to fuse high-frequency edge details with low-frequency approximation features, preserving critical texture information that is essential for tiny target detection. The Programmable Gradient Information (PGI) mechanism is adopted as an auxiliary reversible branch, which propagates reliable deep-layer gradient information to shallow layers to effectively mitigate the information bottleneck problem. Second, a decision-level fusion strategy is developed to aggregate confidence scores and localization outputs from multi-scale predictions. The strategy is further optimized with a tailored Non-Maximum Suppression (NMS) algorithm to resolve occlusion issues in dense UAV clusters. Evaluations on the self-built dense UAV dataset show that the proposed method achieves a detection accuracy of 85% with a per-frame processing time of 0.10–0.15 s. Tests across diverse scenarios validate that the method can accurately locate and identify UAV targets. By achieving a favorable balance between detection accuracy and real-time performance, the proposed method effectively alleviates the missed detection and false detection problems in dense UAV cluster detection tasks.

Keywords: UAV cluster, target detection, YOLOv9, deep

Academic Editor:
Qiang Yu

Submitted: 15 January 2026
Revised: 24 May 2026
Accepted: 08 June 2026
Published: 25 September 2026

Vol. 3, **No.** 3, 2026.
10.62762/CJIF.2026.233582

***Corresponding author:**
Moufa Hu
mfhu20080630@nudt.edu.cn

Citation
Lu, X., Chen, S., Sun, J., Yin, Z., & Hu, M. (2026). A Multi-Scale UAV Cluster Detection and Counting Framework Based on Multi-Source Information Fusion and Enhanced YOLOv9. *Chinese Journal of Information Fusion*, 3(3), 166–177.

learning, low altitude safety, feature fusion.

1 Introduction

Nowadays, as the core carrier of low altitude economy, UAV has been employed in applications such as agriculture, search and rescue operations, surveillance systems, power inspection, logistics distribution, economic construction and mission-critical services [1]. With the development of networking technology, UAV application mode is evolving from single machine operation to large-scale and collaborative "cluster" mode. However, a large number of unauthorized and difficult-to-control flights [2] has brought severe challenges to public security, low altitude airspace management and national defense and military defense.

In the face of increasingly complex low altitude security situation, the reliability of traditional detection methods is lacking when facing dense cluster targets. For example, existing optical-based detection methods face significant limitations in complex environments, such as difficulty in suppressing background clutter and low detection accuracy for small targets [3]. Single optical imaging also struggles in harsh environments and cannot distinguish similar targets effectively. To address these challenges, information fusion and image fusion technologies have emerged as key solutions, which can integrate complementary information from different feature levels and decisions stages to enhance target features and suppress background interference. How to solve the mutual occlusion between targets, distinguish between UAVs and birds and other biological targets [4], and realize the real-time capture of high-speed moving micro targets in complex background, has become the technical bottleneck to be solved. Therefore, the research and development of a fusion detection system based on advanced optical and artificial intelligence technology [5] can quickly identify, accurately classify and count dense UAV clusters in real time, which has important practical significance and strategic value for filling the gap in efficient early warning capabilities.

The main contributions of this article are as follows:

(1) High-Quality UAV Dataset Construction: A multi-scenario UAV dataset containing 7,113 images is constructed to address the scarcity of training data for tiny UAVs. Adopting an "open-source collection + independent construction" strategy, we ensure sample diversity across different target attitudes and environmental conditions.

(2) Feature Extraction and Fusion Module: A Haar Wavelet Down-sampling (HWD) module is introduced to fuse high-frequency edge details with low-frequency components, preventing information loss during downsampling. The Programmable Gradient Information (PGI) mechanism acts as an auxiliary branch to fuse reliable gradient information across different depths, alleviating the information bottleneck of deep networks. The GELAN backbone network aggregates multi-scale features via multi-level feature interaction and adaptive weighting, optimizing gradient propagation and computational efficiency for UAV cluster scenarios.

(3) Decision-Level Fusion and Experimental Validation: A decision-level fusion module is developed to aggregate multi-scale prediction confidence and localization information. Extensive experiments demonstrate that the proposed fusion framework achieves 85% detection accuracy with a processing speed of 0.10–0.15 s per frame, effectively addressing core problems such as feature loss and environmental interference.

The rest of this article is organized as follows. Related works are reviewed in Section 2. Section 3 details the proposed multi-source information fusion methodology, including the feature-level fusion (PGI, GELAN, HWD) and decision-level fusion modules. Section 4 presents the experimental setup, dataset preparation, and performance analysis across different scenarios. Section 5 concludes this article and discusses future prospects.

2 Related Work

Many scholars at home and abroad have carried out extensive research on UAV recognition and made significant progress. In terms of UAV optical feature recognition, for example, Zhao et al. [6] proposed an improved lightweight model YOLOv8-E for fast and accurate detection of small UAVs in complex environments. They enhanced the backbone network with multi-scale feature extraction strategies and improved channel attention mechanisms, achieving near-real-time detection performance with competitive mAP on the UAV small target detection task; Jiang et al. [7] proposed a lightweight object detection method based on improved YOLOv8 for drone images, which streamlines the network architecture and reduces computational burden while maintaining competitive detection accuracy on UAV aerial imagery benchmarks;

Han et al. [8] proposed Light-YOLOv5, a lightweight drone detector designed for resource-constrained cameras. By streamlining the YOLOv5 architecture and reducing computational complexity, the model achieves a favorable balance between detection speed and accuracy, enabling effective drone detection on edge devices; Jiang et al. [9] constructed Anti-UAV, a large-scale multi-modal benchmark dataset for UAV tracking that covers diverse scenarios including thermal infrared and RGB modalities, providing a comprehensive evaluation platform for UAV detection and tracking algorithms. Gao et al. [10] constructed a novel UAV dataset and proposed an improved YOLOv9-based detection method. By introducing structural optimizations to the backbone and neck, the method achieves substantial improvements in both precision and recall on the self-built UAV dataset, demonstrating the advantage of YOLOv9 in small target UAV detection scenarios.

In the aspect of UAV other feature recognition, for example, Al-Emadi et al. [11] proposed a deep learning-based method for audio-based drone detection and identification. By training convolutional neural networks on drone acoustic signals, the method demonstrates that sound-based features can effectively complement visual cues for UAV recognition, particularly in scenarios where optical sensors are limited; Medaiyese et al. [12] developed a hierarchical learning framework for UAV detection and identification based on radio frequency (RF) signals. The framework employs a layered classification strategy that progressively distinguishes UAV types from other wireless devices, achieving reliable identification performance across different confidence ranges; Al-Emadi et al. [13] further improved audio-based drone detection by augmenting the training dataset with samples generated through generative adversarial networks (GANs). This data enhancement strategy effectively addresses the scarcity of drone acoustic training samples and improves model robustness under varying environmental noise conditions and distances. McCoy et al. [14] proposed an ensemble deep learning framework for multimodal UAV classification that jointly processes acoustic, image/video, and radio frequency (RF) signals. The experimental results demonstrate that this multimodal fusion approach outperforms single-modality detection methods, offering superior robustness against environmental interference; Based on the framework of machine learning, Alipour et al. [15] used the maximum likelihood estimation (MLE) method to estimate the packet arrival time by identifying the packet size of encrypted WiFi traffic. The recognition accuracy of UAV was 88.5% to 9.82%.

From the above studies, they basically adopted the deep learning model. In terms of optical feature recognition, many scholars used different methods to improve the accuracy of UAV recognition, but did not completely solve the recognition problems of target occlusion, different environments and high density. Ma et al. [16] proposed YOLO-UAV, a multi-scale feature fusion detection framework for UAV imagery based on improved YOLOv5. By incorporating dense connections and efficient feature aggregation in the backbone and neck networks, the method significantly improves the detection of small and densely distributed targets in complex backgrounds, achieving strong performance on UAV imagery benchmarks. In other feature recognition methods, many scholars have made great progress by using UAV signals, voiceprints and other features, but their methods are vulnerable to environmental factors, resulting in unstable recognition accuracy.

In the field of small object detection and complex scenarios, YOLOv9 demonstrates greater advantages. YOLOv5n exhibits relatively lower recall rates, potentially leading to more missed detections. In drone recognition scenarios with complex backgrounds and minute objects, YOLOv8 may underperform compared to YOLOv9, which has been specifically optimized for information flow. YOLOv10's innovation lies in eliminating non-max suppression, enabling fully end-to-end inference and achieving substantial speed improvements. However, this may come at the cost of reduced ability to capture subtle features.

Therefore, based on the YOLOv9 algorithm, this paper proposes to collect UAV images, prepare data sets and build a network algorithm to extract the characteristics of the images, and conduct in-depth analysis of the input images. Through the continuous optimization and improvement of the algorithm and system, the accurate identification of UAVs is finally realized, and the number of UAVs is counted, which provides an efficient and stable early warning means for resisting cluster UAVs in the fields of public security, national defense and military.

3 Methodology

In this paper, an intelligent UAV cluster detection and counting system is developed based on a multi-source information fusion framework centered on an enhanced YOLOv9 architecture. The core innovation is a hierarchical fusion architecture, as illustrated in Figure 1, which systematically integrates multi-level visual information to address the challenges of multi-scale UAV detection. The framework comprises two main fusion stages: feature-level fusion and decision-level fusion.

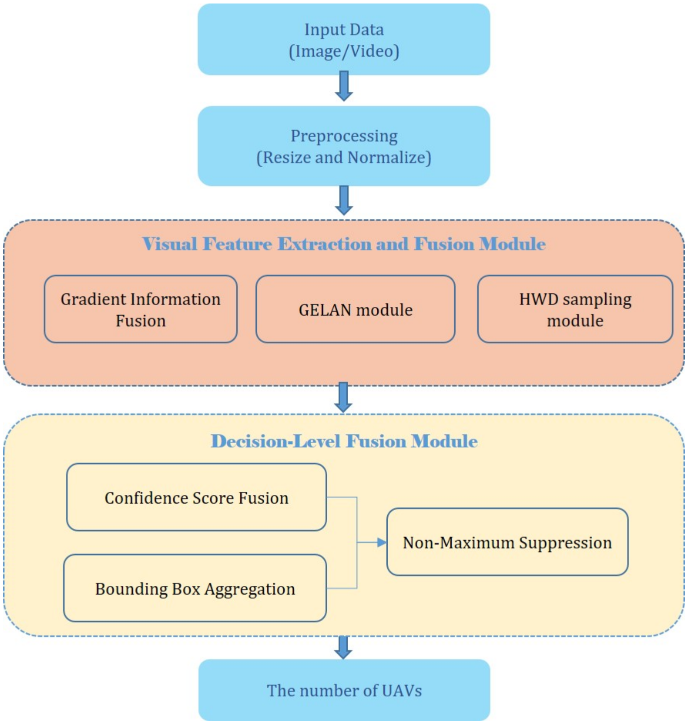

Figure 1. Detection and counting framework based on multi-source information fusion.

The feature extraction and fusion module focuses on integrating multi-scale and multi-frequency features from input images. It consists of three key components: (1) The HWD module fuses high-frequency edge details with low-frequency approximations to avoid information loss during resolution reduction; (2) The PGI mechanism generates reliable gradients to update the main branch via an auxiliary reversible branch during training. This resolves the deep network information bottleneck and enhances tiny target detection without increasing inference cost; (3) The GELAN backbone network aggregates multi-scale features via multi-level feature interaction and adaptive weighting, optimizing gradient propagation and computational efficiency for UAV cluster scenarios.

The decision-level fusion module integrates confidence scores and localization information generated by multiple prediction heads. To eliminate redundant detections and suppress false positives in dense cluster scenarios, we adopt an enhanced NMS algorithm. This algorithm performs holistic fusion of decision-level outputs by adaptively aggregating confidence scores from overlapping proposals, producing cleaner and more reliable detection results. In addition, a confidence aggregation strategy is applied to further refine final predictions by weighting multi-head outputs based on their spatial consistency. This decision-level fusion ensures that only the most salient and non- redundant detections are retained, enabling accurate real-time counting of UAVs in dense clusters.

The overall framework adopts a layered architecture (data, algorithm, application) to realize end-to-end detection. Deployment and training are implemented in Python with the PyTorch framework, loading YOLOv9-c pre-trained weights. A local feature enhancement strategy is applied, and the model is trained with an SGD optimizer for 150 epochs. This hierarchical design ensures that no critical information is lost from the pixel level to the final decision, providing a robust and systematic solution for dense small-target detection in complex environments.

3.1 Visual Feature Extraction and Fusion Module

3.1.1 PGI Mechanism for Gradient Information Fusion

PGI [17] is designed to address the "information bottleneck" phenomenon in deep neural networks [18]. In the context of drone detection [19], let X be the input image data and Y be the ground truth labels of the drones. As the network depth increases, the features F extracted by a traditional network f_θ suffer from a loss of Mutual Information, denoted as $I(X,Y) \geq I(F,Y)$. For miniature drones with minimal pixel coverage, this information attenuation leads to ambiguous gradient update directions.

To mitigate this issue, we introduce an auxiliary reversible branch to generate reliable gradients. According to the reversible neural network theory [20], if there exists a reversible function g_ϕ, it guarantees a lossless mapping of information from the input data X to the feature layers, expressed as:

$$I(X, g_\phi(X)) = I(X, X) \tag{1}$$

This ensures that the edge and texture information (high-frequency features) of the miniature drones is completely preserved within the auxiliary branch. During the training process, the total loss function L_{total} is composed of the main branch loss L_{main} and the auxiliary branch loss L_{aux}:

$$L_{\text{total}} = L_{\text{main}}(f_\theta(X), Y) + \alpha \cdot L_{\text{aux}}(g_\phi(X), Y) \tag{2}$$

where α is a balancing coefficient. Through backpropagation, the reliable gradients $\nabla_\theta L_{\text{aux}}$ generated by the auxiliary branch are aggregated with the main branch gradients to jointly guide the update of the parameters θ of the backbone network (GELAN):

$$\theta_{\text{new}} \leftarrow \theta - \eta \left(\nabla_\theta L_{\text{main}} + \alpha \nabla_\theta L_{\text{aux}}\right) \tag{3}$$

This constitutes a fusion of gradient information from multiple computational paths, forcing the network to learn faint features that are easily overlooked. During the inference phase, the auxiliary branch g_ϕ is removed, retaining only f_θ. This achieves high-precision detection of dense miniature drones without increasing the computational load (FLOPs) during inference.

In deep neural networks, original feature information in input data attenuates layer by layer as the number of layers increases, resulting in gradient update deviations. As an auxiliary supervision framework, PGI solves the deep network training problem through the synergy of the main branch, auxiliary reversible branch, and multi-level auxiliary information. The effectiveness of this multi-branch gradient fusion strategy for UAV small target detection has been validated in recent work [21], where GELAN combined with Sobel-enhanced auxiliary supervision significantly improves feature extraction accuracy in remote sensing scenarios. By introducing auxiliary reversible branches, PGI generates reliable gradient information to guide parameter updates of the main branch. This enables the model to maximally retain complete semantic information of input images when dealing with small targets such as UAVs, ensuring that subtle target features remain "visible" in deep networks.

3.1.2 GELAN module

YOLOv9 uses a sophisticated Generalized Efficient Layer Aggregation Network (GELAN) for feature extraction, which is built upon the Cross Stage Partial Network (CSPNet) [22] to achieve efficient gradient path planning and improved learning capability with reduced computational overhead. GELAN takes gradient path planning as the core design concept and balances the feature expression ability and computational efficiency through the generalization layer aggregation strategy. Its core advantage is to break through the limitations of the traditional fixed connection mode and realize the efficient integration of multi-scale and multi-level features, so as to improve the parameter efficiency and feature aggregation [23]. The architecture adapts to the variable target scale, complex background and real-time requirements in UAV cluster recognition, and its principle mechanism can be divided into two modules: feature level division and dynamic aggregation.

In the feature level division stage, GELAN performs multi-scale decomposition on the feature map extracted from the input image through the backbone network, and sets the input feature tensor as $X \in \mathbb{R}^{C\times H\times W}$ (C as the number of channels and H, W are the spatial dimensions), and L level features $\{F_l\}_{l=1}^{L}$ are generated by convolution operation with step size of 2, where the L-th level feature $F_l \in \mathbb{R}^{C_l\times H_l\times W_l}$ meets the requirements of $H_l = H/(2^l)$ and $W_l = W/(2^l)$, realizing the multi-scale feature extraction requirements of individual and formation.

The dynamic aggregation stage is the core of GELAN. The weighted cross layer aggregation strategy is adopted to dynamically allocate the feature weights of each level through learnable parameters. Let the aggregation feature be F_{agg}, then

$$F_{\text{agg}} = \sum_{l=1}^{L} \alpha_l \cdot \text{Conv}(F_l, \theta_l) \tag{4}$$

where α_l is the feature weight of layer l (if $\sum_{l=1}^{L} \alpha_l = 1$), θ_l is kernel parameters and $\text{Conv}(\cdot)$ is an adaptive convolution operation to achieve feature dimension unification and information fusion. This mechanism dynamically fuses features from different scales and semantic levels, strengthening target characteristics and suppressing background noise [24].

The GELAN module improves the feature extraction ability by combining a variety of convolution operations. This principle of diverse convolutional strategies for tiny UAV target detection has been demonstrated in optimized YOLO-based architectures [25], where combining multiple convolutional designs effectively preserves fine-grained features of miniature UAV targets. GELAN combines the advantages of cross phase local network (CSPNET) and efficient layer aggregation network (ELAN), and simplifies the feature extraction process by optimizing gradient path planning. In order to ensure the accuracy and meet the real-time requirements, this paper uses GELAN as the backbone feature extraction network. It can greatly reduce the amount of parameters and computational complexity without sacrificing the accuracy, minimizing computational overhead while preserving rich feature representations [26], so that the algorithm can be deployed on non-server hardware such as laptop computers for real-time reasoning.

3.1.3 HWD sampling module

Given the extremely low pixel proportion of tiny UAVs in images, traditional downsampling methods (such as max pooling or strided convolution) have significant limitations. While reducing feature map size, they often directly discard high-frequency information containing target textures, resulting in "feature loss". To solve this problem, this paper introduces HWD wavelet downsampling for spatial feature fusion. HWD decomposes the input feature map into low-frequency approximations and high-frequency details, then fuses them to achieve lossless downsampling. HWD is a widely recognized and compact transform, which is widely used in image coding, edge extraction and binary logic design. Therefore, this module can be easily integrated into CNN to enhance the performance of the model [27]. In this paper, we replace certain convolutions in the original downsampling architecture with the HWD module within the backbone to enhance the model's semantic segmentation performance.

The proposed HWD module comprises two patterns: (1) the lossless feature encoding block and (2) the feature representation learning block. We utilize the Haar wavelet transform to decrease the resolution of feature maps while retaining all information. The representation learning block consists of a standard convolution layer, batch normalization, and a ReLU activation layer. It is employed to extract discriminative features.

HWD takes multi-resolution analysis as the core, realizes feature dimension reduction and key information retention through simple average and difference operations, and adapts to the requirements of low computational power deployment and small target feature extraction in UAV cluster recognition. The core principle is to decompose the input characteristic map into low-frequency approximation and high-frequency details, and then compress the data by downsampling. Let the input characteristic tensor be $X \in \mathbb{R}^{C\times H\times W}$ (C is the number of channels, H and W are the spatial dimensions), HWD first decomposes the features through Haar wavelet basis convolution: low frequency approximation coefficient

$$s_{jn} = \frac{1}{\sqrt{2}}(s_{j+1,2n} + s_{j+1,2n+1}) \tag{5}$$

high frequency detail coefficient

$$\omega_{jn} = \frac{1}{\sqrt{2}}(s_{j+1,2n} - s_{j+1,2n+1}) \tag{6}$$

For a 2D feature map, this process is applied along rows and columns, fusing information into four sub-bands (LL, LH, HL, HH), and finally the low-frequency component $F_L \in \mathbb{R}^{C\times H/2\times W/2}$ is retained to complete the downsampling, while the high-frequency components are fused to preserve edge texture. This mechanism is implemented by 2×2 fixed convolution kernel, featuring high computational efficiency. The HWD module acts as a specialized feature-level fusion operator, explicitly retaining key semantic features such as UAV cluster formation contours, and suppresses high-frequency noise from low-altitude backgrounds.

3.2 Decision-Level Fusion Module

After the feature-level fusion stage, the network produces multi-scale prediction tensors. To further enhance robustness, we propose a Decision-Level Fusion Module that aggregates these predictions. This module treats outputs from different detection heads (e.g., for small, medium, and large targets) as independent information sources. Let $P_i = \{C_i, B_i\}$ be the set of predictions (confidences C_i and bounding boxes B_i) from the i-th detection scale. The goal of decision-level fusion is to generate a unified detection set P_{final} that is more reliable than any single scale.

Confidence Score Fusion: For a given predicted object, its final confidence score C_{final} is computed as a weighted sum of its scores from different scales where it appears:

$$C_{\text{final}} = \sum_{i=1}^{N} \omega_i \cdot C_i \tag{7}$$

where ω_i are fusion weights (e.g., learned or set empirically) and $\sum \omega_i = 1$.

Bounding Box Aggregation: For an object detected by multiple scales, the final bounding box coordinates $B_{\text{final}} = (x_{\text{final}}, y_{\text{final}}, w_{\text{final}}, h_{\text{final}})$ are computed as a confidence-weighted average of the individual predictions:

$$B_{\text{final}} = \frac{\sum_{i=1}^{N} C_i \cdot B_i}{\sum_{i=1}^{N} C_i} \tag{8}$$

Non-Maximum Suppression (NMS): After fusing confidences and boxes, an NMS algorithm is applied to remove redundant detections [28, 29]. The NMS algorithm eliminates boxes with a high Intersection-over-Union (IoU) overlap, retaining only the optimal bounding box for each target. This process is formalized as:

$$P_{\text{keep}} = \text{NMS}(P_{\text{final}}, \tau_{\text{IoU}}) \tag{9}$$

where τ_{IoU} is the IoU threshold. After processing each frame, the algorithm counts the number of ultimately retained detection boxes, combined with target tracking logic (such as simple inter-frame matching), to realize real-time statistics and deduplication of UAV cluster counts in the current frame.

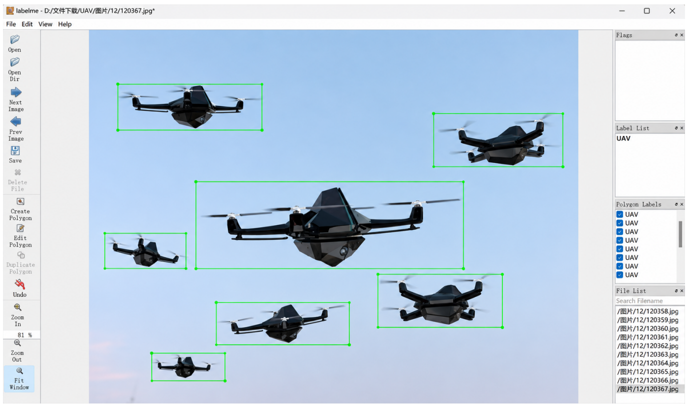

Figure 2. Dataset annotation interface.

4 Experiments

4.1 Dataset Construction and Preprocessing

High quality, multi-scene image dataset is the basis of training high-precision detection model. Due to the scarcity of public micro UAV data sets, we adopt the strategy of "open-source collection + independent construction". The specific implementation steps are as follows:

Data collection: there are few public research results in the field of UAV detection, mainly due to the lack of training data sets. We retrieved and downloaded the existing UAV image data from GitHub, Kaggle and other database websites using Internet resources, and established a basic database. In order to increase the dynamic diversity of data, we extracted images from the collected UAV flight video at the frequency of one frame per second, and obtained a large number of UAV samples under different attitudes.

Data annotation: we use the professional image annotation tool LabelMe to annotate the collected images one by one. In the labeling process, the rectangular box is drawn strictly to the edge of the UAV, and the corresponding label file is generated. The marked data can accurately reflect the position and contour information of UAV in different environments. Aiming at the problem that the characteristics of micro UAV are not obvious in the dark light environment, we adopt a specific image enhancement technology. Some images are highlighted and cropped to enhance the contrast of local features, so that the model can effectively distinguish target details in dark conditions. An example of the annotation interface is shown in Figure 2.

Finally, after several rounds of manual screening, the blurred or mislabeled images were eliminated, and 7113 high-quality images were selected to form the final dataset. Figure 3 and Figure 4 present representative samples from the constructed dataset, including high-altitude circling UAVs (Figure 3) and daytime routine UAV detection scenarios (Figure 4), respectively, demonstrating the diversity of target scales and backgrounds covered in our dataset.

Figure 3. Sample of high-altitude circling UAVs.

Figure 4. Sample of daytime routine UAV detection.

The processed dataset is randomly divided into training set, test set, and validation set according to a ratio of 8:1:1. Before being sent to the network for training, a script is written to uniformly adjust the size of all input images to 640×640 pixels to meet the input tensor requirements of the YOLOv9 model.

4.2 Model Training and Result Analysis

During model training, the HWD module is used to replace part of the traditional convolution downsampling, and Haar wavelet transform is used to retain high-frequency texture information of images. In addition, a local feature-based training method is adopted. In the original training process, target regions with limited recognition effects are cropped into independent images, which are combined with high-resolution single UAV images to generate a new dataset for joint training. The hyper parameter settings are as follows: batch size is set to 4, epochs are set to 150, workers are set to 0, the optimizer selects the SGD optimizer, and the random number seed is fixed to 0 for reproducible experimental results. The results are presented as a confusion matrix, as shown in Table 1.

Table 1. Confusion matrix results.

	UAV	Non UAV
UAV	0.85	0
Non UAV	0.15	1

According to the confusion matrix analysis, the model achieves a precision of 0.85 for the "UAV" category. This indicates that the model can effectively extract UAV features and distinguish them from backgrounds. The confusion matrix also shows that the model has a certain false detection rate in the low-confidence interval. Analysis shows that this is mainly due to interference signals generated by complex backgrounds (such as shaking leaves, birds, etc.). Therefore, the default confidence threshold is set at approximately 0.5, with a dynamic adjustment function to filter out these low-confidence false detection targets. Although there is a small amount of background interference, the model generally achieves high-precision recognition of tiny UAVs, laying a solid model foundation for subsequent real-time detection and counting functions.

4.3 Ablation Study

To evaluate the performance of the proposed algorithm, we conducted ablation experiments. We selected the UAV dataset constructed in Section 4.1 and performed training and testing under identical conditions. Using the standard YOLOv9 model as the baseline, we compared it with the multi-source information fusion framework proposed in this paper. The experimental results are presented in Table 2. The results show that after adding feature fusion and decision-level fusion modules (confidence score fusion + bounding box aggregation), the precision reaches 91.2% and mAP@0.5 reaches 88.2%, which are 19.6% and 34.8% higher than those of baseline YOLOv9, respectively. Although the addition of fusion modules slightly reduces FPS (from 15.2 to 7.14), the system still maintains real-time detection performance.

4.4 Different Scenarios

In order to comprehensively evaluate the performance of the improved algorithm, some typical scenarios are selected for testing.

4.4.1 Scenario 1: Small targets against ground background

UAV images of small targets with the ground as the background were selected, with the UAV fleet flying in formation at medium and long ranges. With minimal interference, this scenario serves as the baseline group to verify the basic performance of the model. As shown in Figure 5, there are 19 UAVs in the scene. When the confidence threshold is set to 0.4, the proposed algorithm can accurately detect every single UAV. When the threshold is set to 0.75, 18 UAVs can be detected. The detection process takes 0.125 s. This shows that in conventional long-range scenarios without extreme interference, the YOLOv9 baseline model combined with our parameter optimization has extremely stable detection capability to meet daily inspection needs.

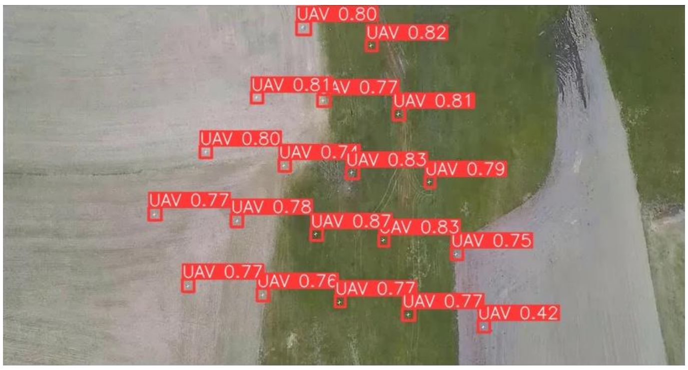

Figure 5. Detection results of Scenario 1.

Table 2. Ablation experiments results.

	Precision (%)	F1 (%)	Recall (%)	mAP@0.5 (%)	mAP@0.5:0.95 (%)	FPS
YOLOv9	71.6	59.2	50.4	53.4	33.7	15.2
Ours	91.2	83.7	77.3	88.2	56.7	7.14

4.4.2 Scenario 2: Tiny targets against sky background

The UAV group ascends to a very high altitude for circling. The targets appear as tiny black spots in the image, making specific models indistinguishable to the naked eye, with a pure sky background. As shown in Figure 6, there are 3 UAVs in the scene. Although the target pixel ratio is extremely low, the improved model still achieves a 100% detection rate, with confidence for each target exceeding 0.8, and the detection process takes 0.120 s. This performance benefits from the HWD module. Unlike traditional pooling layers that directly discard high-frequency information, the HWD module retains edge texture features of small targets through wavelet transform. When the feature map size is halved, subtle texture features such as UAV edges and rotors are not lost, but encoded into the channel dimension. This "lossless" downsampling strategy significantly improves the model's perception ability for small and weak targets at long ranges, so that the model can still capture small target points in deep networks.

Figure 6. Detection results of Scenario 2.

4.5 Scenario 3: Large targets against sky background

A daytime environment with good lighting conditions is selected, with UAVs flying in close formation. At this time, target outlines are clear, rotor and fuselage structures are visible, and the background includes the sky and horizon. As shown in Figure 7, under clear feature conditions, the confidence of test results is 0.75 or above, and the detection process takes 0.129 s. Detection boxes perfectly fit the edges of UAVs without redundant frames, and can accurately capture UAVs at edge positions. This proves that the model has high feature extraction integrity under ideal conditions and can make full use of rich texture information at close range for accurate classification.

Figure 7. Detection results of Scenario 3.

The quantitative results for all three scenarios are summarized in Table 3. Under the same confidence threshold, the model achieves 100% detection accuracy across all scenarios, with processing times consistently around 0.12–0.13 s per frame, demonstrating the robustness and real-time capability of the proposed method.

Table 3. Scenario test results.

Scenario	Actual number of t argets in the scene	Number of detected targets	Time consuming/s
Scenario 1	19	19	0.125
Scenario 2	3	3	0.120
Scenario 3	5	5	0.129

5 Conclusion

UAVs have been widely used in civil, military, and other fields, and play an important role in improving efficiency and enabling innovative applications. This study focuses on the low-altitude security challenges brought by unauthorized and random flights of tiny UAV clusters.

By constructing a multi-scale, multi-scenario UAV-specific dataset, introducing the HWD wavelet downsampling module, designing a local feature enhancement training strategy, and developing GUI software with real-time inference and counting functions, high-precision detection and counting of UAV clusters are realized. To address the issues of feature loss in multi-scale UAV targets and interference from complex environments, this paper proposes a real-time detection method based on a multi-source information fusion framework, rather than a simple modification of the YOLOv9 baseline. The framework includes a feature-level fusion stage (integrating high-frequency spatial details via the HWD module and deep gradient information via the PGI mechanism) and a decision-level fusion stage (aggregating multi-scale confidence scores and bounding boxes). Experimental verification shows that the proposed method achieves a detection accuracy of 85% with a per-frame processing speed of 0.10–0.15 s. With a confidence threshold above 0.4, it effectively solves core problems such as small target feature loss, complex environmental interference, and difficulty in counting occluded clusters. This research enriches the dense small target detection algorithm system, integrates HWD wavelet transform and local feature enhancement strategies into the YOLOv9 architecture, and provides a new technical approach for cluster detection of tiny UAVs.

However, this study still has limitations. Although the dataset covers day and night, long-range and short-range scenes, the proportion of extreme weather samples (such as heavy rain and dense fog) and special target types (such as subminiature and stealth UAVs) is relatively low. The counting stability of the model in dense target scenarios still has room for improvement. In future research, we will further expand extreme scenarios and special target types to improve the generalization ability of the model. We will focus on advanced feature fusion technologies as the key research direction, including knowledge-guided multi-scale feature fusion and fusion of physical features and digital features. These strategies will effectively enhance the feature representation capability of UAV targets of different sizes, especially small and occluded targets, and further improve the overall performance and practical applicability of the detection algorithm.

Data Availability Statement

Data will be made available on request.

Funding

This work was supported without any funding.

Conflicts of Interest

The authors declare no conflicts of interest.

AI Use Statement

The authors declare that DeepSeek-R1 was used for language editing, and DeepL Translator was used for translation of manuscript from Chinese into English. The authors have carefully reviewed, revised, and verified the AI-assisted output and take full responsibility for the content of the manuscript.

Ethical Approval and Consent to Participate

Not applicable.

References

[1] Ahmad, T., Morel, A., Cheng, N., Palaniappan, K., Calyam, P., Sun, K., & Pan, J. (2025). Future UAV/drone systems for intelligent active surveillance and monitoring. *ACM Computing Surveys*, *58*(2), 1-37. [CrossRef]

[2] Liao, X., Xu, C., & Yue, H. (2019, June). Enable UAVs safely flight in low-altitude: A Preliminary research of the public air route network of UAVs. In *2019 International Conference on Unmanned Aircraft Systems (ICUAS)* (pp. 959-964). IEEE. [CrossRef]

[3] Wang, F., Wang, H., Qin, Z., & Tang, J. (2023). UAV target detection algorithm based on improved YOLOv8. *IEEE Access*, *11*, 116534–116544. [CrossRef]

[4] Jin, W. C., Oh, D., Lee, S. C., & Choi, J. W. (2025, November). Motion-Based Bird-UAV Classification Using 3D-CNN for Long-Range Anti-UAV Systems. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management* (pp. 6867-6868). [CrossRef]

[5] Norbelt, M., Luo, X., Sun, J., Wang, J., Claude, U., & Wisal, M. (2024, July). Detecting UAV Target with Deep Convolutional Neural Network. In *Proceedings of the 2024 6th International Electronics Communication Conference* (pp. 11-18). [CrossRef]

[6] Zhao, Y., Wang, L., Lei, G., Guo, C., & Ma, Q. (2024). Lightweight UAV small target detection and perception based on improved YOLOv8-E. *Drones*, *8*(11), 681. [CrossRef]

[7] Jiang, M., Si, Z., Yang, K., & Zhang, Y. (2024, July). Improved YOLOv8-Based Lightweight Object Detection on Drone Images. In *International Conference on Intelligent Computing* (pp. 426-434). Singapore: Springer Nature Singapore. [CrossRef]

[8] Han, J., Cao, R., Brighente, A., & Conti, M. (2023). Light-YOLOv5: A lightweight drone detector for resource-constrained cameras. *IEEE Internet of Things Journal*, *11*(6), 11046-11057. [CrossRef]

[9] Jiang, N., Wang, K., Peng, X., Yu, X., Wang, Q., Xing, J., ... & Han, Z. (2023). Anti-UAV: A large multi-modal benchmark for UAV tracking. *IEEE Transactions on Multimedia, 25*, 486–500. [CrossRef]

[10] Gao, D., Tang, J., Li, H., Wang, B., Qiu, J., Chen, S., & Mei, X. (2024). A novel dataset and detection method for unmanned aerial vehicles using an improved YOLOv9 algorithm. *Sensors, 24*(23), 7512. [CrossRef]

[11] Al-Emadi, S., Al-Ali, A., Mohammad, A., & Al-Ali, A. (2019, June). Audio based drone detection and identification using deep learning. In *2019 15th International Wireless Communications & Mobile Computing Conference (IWCMC)* (pp. 459-464). IEEE. [CrossRef]

[12] Medaiyese, O. O., Ezuma, M., Lauf, A. P., & Adeniran, A. A. (2022). Hierarchical learning framework for UAV detection and identification. *IEEE Journal of Radio Frequency Identification, 6*, 176-188. [CrossRef]

[13] Al-Emadi, S., Al-Ali, A., & Al-Ali, A. (2021). Audio-based drone detection and identification using deep learning techniques with dataset enhancement through generative adversarial networks. *Sensors, 21*(15), 4953. [CrossRef]

[14] McCoy, J., Rawal, A., Rawat, D. B., & Sadler, B. M. (2022). Ensemble deep learning for sustainable multimodal UAV classification. *IEEE Transactions on Intelligent Transportation Systems, 24*(12), 15425-15434. [CrossRef]

[15] Alipour-Fanid, A., Dabaghchian, M., Wang, N., Wang, P., Zhao, L., & Zeng, K. (2019). Machine learning-based delay-aware UAV detection and operation mode identification over encrypted Wi-Fi traffic. *IEEE Transactions on Information Forensics and Security, 15*, 2346-2360. [CrossRef]

[16] Ma, C., Fu, Y., Wang, D., Guo, R., Zhao, X., & Fang, J. (2023). YOLO-UAV: Object detection method of unmanned aerial vehicle imagery based on efficient multi-scale feature fusion. *IEEE Access, 11*, 126857-126878. [CrossRef]

[17] Wang, C. Y., Yeh, I. H., & Mark Liao, H. Y. (2024, September). Yolov9: Learning what you want to learn using programmable gradient information. In *European conference on computer vision* (pp. 1-21). Cham: Springer Nature Switzerland. [CrossRef]

[18] Tishby, N., & Zaslavsky, N. (2015, April). Deep learning and the information bottleneck principle. In *2015 IEEE Information Theory Workshop (ITW)* (pp. 1-5). IEEE. [CrossRef]

[19] Zhai, X., Huang, Z., Li, T., Liu, H., & Wang, S. (2023). YOLO-Drone: an optimized YOLOv8 network for tiny UAV object detection. *Electronics, 12*(17), 3664. [CrossRef]

[20] Gomez, A. N., Ren, M., Urtasun, R., & Grosse, R. B. (2017). The reversible residual network: Backpropagation without storing activations. *Advances in neural information processing systems, 30*.

[21] Xu, S., Cui, K., Chen, G., & Zhang, M. (2025). SGA-YOLOv8: Sobel-GELAN and AIFI-dilated-attention fusion network for UAV remote-sensing small target detection. *The Journal of Supercomputing, 81*(14), 1291. [CrossRef]

[22] Wang, C. Y., Liao, H. Y. M., Wu, Y. H., Chen, P. Y., Hsieh, J. W., & Yeh, I. H. (2020, June). CSPNet: A new backbone that can enhance learning capability of CNN. In *2020 IEEE/CVF conference on computer vision and pattern recognition workshops (CVPRW)* (pp. 1571-1580). IEEE. [CrossRef]

[23] Sakib Bin Islam, M., Chowdhury, M. E., Hasan-Zia, M., Kashem, S. B. A., Majid, M. E., Ansaruddin Kunju, A. K., ... & Nashbat, M. (2025). VisioDECT: a novel approach to drone detection using CBAM-integrated YOLO and GELAN-E models. *Neural Computing and Applications, 37*(24), 20181-20204. [CrossRef]

[24] Hu, K., Lu, J., Zheng, C., Xiang, Q., & Miao, L. (2025). MFF-YOLOv8: Small Object Detection Based on Multi-Scale Feature Fusion for UAV Remote Sensing Images. *IET Image Processing, 19*(1), e70066. [CrossRef]

[25] Bai, C., Zhang, K., Jin, H., Qian, P., Zhai, R., & Lu, K. (2025). SFFEF-YOLO: Small object detection network based on fine-grained feature extraction and fusion for unmanned aerial images. *Image and Vision Computing, 156*, 105469. [CrossRef]

[26] Zhang, H., Peng, Y., & Liu, Y. L. (2025). UAV aerial photography target detection based on improved YOLOv9. *The Journal of Supercomputing, 81*(3), 492. [CrossRef]

[27] Xu, G., Liao, W., Zhang, X., Li, C., He, X., & Wu, X. (2023). Haar wavelet downsampling: A simple but effective downsampling module for semantic segmentation. *Pattern recognition, 143*, 109819. [CrossRef]

[28] Bodla, N., Singh, B., Chellappa, R., & Davis, L. S. (2017, October). Soft-NMS—Improving object detection with one line of code. In *2017 IEEE international conference on computer vision (ICCV)* (pp. 5562-5570). IEEE. [CrossRef]

[29] Huang, J. (2025, June). Small target detection algorithm based on improved YOLOv8s. In *2025 5th International Conference on Artificial Intelligence, Big Data and Algorithms (CAIBDA)* (pp. 59-62). IEEE. [CrossRef]

Xinpeng Lu, undergraduate majoring in electronic information engineering, School of Electronic Science, National University of Defense Technology, mainly engaged in research on edge intelligent computing. Published 1 paper and obtained 1 Software Copyright. (Email: 1636953696@qq.com)

Shijie Chen, undergraduate majoring in electronic information engineering, School of Electronic Science, National University of Defense Technology, mainly engaged in embedded system design and other research. Published 1 paper, won 1 software copyright, and won the second prize of military modeling. (Email: 3144727498@qq.com)

Jiashuo Sun, undergraduate majoring in Optoelectronic Information Science and Engineering in the College of Frontier interdisciplinary sciences, National University of Defense Technology, and is mainly engaged in the research of target recognition. Published 1 paper and obtained 1 software copyright. (Email: 3424001040@qq.com)

Zhichao Yin received the MS degree in Geomatics Science and Technology from Chongqing Jiaotong University, China in 2024. Currently serves as an engineer in Key Laboratory of Automatic Target Recognition, National University of Defense Technology, P. R. China. His research interests indoor positioning, target detection and recognition, intelligent image real-time information processing, and Pedestrian Activity Recognition. (Email: 15200430340@163.com)

Moufa Hu received the BS, MS and PhD degree from National University of Defense Technology, Changsha, P. R. China, in 2001, 2003 and 2008. He is now an associate researcher in Key Laboratory of Automatic Target Recognition, National University of Defense Technology, P. R. China. He has authored or coauthored more than 40 journal and conference proceedings papers. He has been the principal investigator of 20 research projects. His current research interests include optical imaging target detection and recognition, intelligent image real-time information processing, and algorithm performance evaluation. (Email: mfhu20080630@nudt.edu.cn)