Journal of Electronic Research and Application, 2025, Volume 9, Issue 3
http://ojs.bbwpublisher.com/index.php/JERA
ISSN Online: 2208-3510
ISSN Print: 2208-3502

Research on Real-Time Object Detection and Tracking for UAV Surveillance Based on Deep Learning

Fei Liu*, Lu Jia

Sichuan Southwest Vocational College of Civil Aviation, Sichuan 610039, Chengdu, China

**Author to whom correspondence should be addressed.*

Abstract: To address the challenges of low accuracy and insufficient real-time performance in dynamic object detection for UAV surveillance, this paper proposes a novel tracking framework that integrates a lightweight improved YOLOv5s model with adaptive motion compensation. A UAV-view dynamic feature enhancement strategy is innovatively introduced, and a lightweight detection network combining attention mechanisms and multi-scale fusion is constructed. The robustness of tracking under motion blur scenarios is also optimized. Experimental results demonstrate that the proposed method achieves a mAP@0.5 of 68.2% on the VisDrone dataset and reaches an inference speed of 32 FPS on the NVIDIA Jetson TX2 platform. This significantly improves the balance between accuracy and efficiency in complex scenes, offering reliable technical support for real-time applications such as emergency response.

Keywords: UAV surveillance; Real-time object detection; Deep learning; Lightweight model; Motion compensation

Online publication: 5 June, 2025

1. Introduction

In recent years, research on object detection for unmanned aerial vehicles (UAVs) has experienced rapid development both domestically and internationally. Deep learning has emerged as the core driving force, significantly enhancing UAVs' perception capabilities in complex environments. Early international studies predominantly employed two-stage object detection algorithms, such as Faster R-CNN and Mask R-CNN, which improve detection accuracy through Region Proposal Networks (RPN) and Feature Pyramid Networks (FPN). However, these methods are computationally intensive and often fail to meet the real-time requirements of UAV applications [1]. With breakthroughs in single-stage detectors such as the YOLO series and SSD, researchers have increasingly adopted lightweight network designs, such as the MobileNet backbone in YOLOv5, and multi-scale prediction techniques to achieve a balance between speed and accuracy. For instance, YOLOv7 reaches an impressive detection speed of 160 FPS on the NVIDIA V100 GPU when evaluated on aerial datasets like

VisDrone [2]. In addition, multi-modal fusion approaches (e.g., combining infrared and visible spectrum data) have been introduced to improve detection performance in nighttime or adverse weather conditions, such as enhancing infrared image details via super-resolution reconstruction. In China, research has mainly focused on adapting existing algorithms to the unique challenges of UAV platforms. For example, Li *et al*. embedded multi-head attention mechanisms into YOLOv5 to enhance the extraction of small-object features and used shallow feature fusion modules to improve detection accuracy [3]. Other efforts include dynamic feature enhancement strategies, such as GAN-based motion blur restoration, and innovations in sensor technologies, such as the integration of LiDAR and high-definition cameras, which have expanded UAV applications in agriculture, forestry, and beyond.

Nonetheless, several technical bottlenecks remain in UAV object detection. First, small object detection is hindered by the low pixel occupation of targets in aerial views (e.g., pedestrians occupying fewer than 50×50 pixels), resulting in insufficient feature representation. While improvements to FPNs (e.g., the C2f module in YOLOv8) and attention mechanisms (e.g., CBAM) can mitigate this issue, detection performance in dense scenes still suffers from high miss rates [4]. Second, complex backgrounds and occlusion severely impact detection robustness. In forest fire monitoring, for instance, occlusion by foliage or thermal interference in infrared imaging necessitates the use of context-aware algorithms and multi-sensor fusion (e.g., infrared-visible synergy) to maintain target consistency. Additionally, UAVs face challenges in adapting to dynamic environments. High-speed movement can cause image blur and dramatic scale variation (e.g., targets rapidly approaching from high altitudes), which requires integration of IMU-based motion compensation and adaptive anchor box clustering techniques for optimization [5]. The trade-off between real-time performance and computational resource constraints is another major concern. The limited processing power of edge devices (e.g., Jetson TX2) has driven research in lightweight model design (e.g., pruning, quantization) and hardware acceleration using frameworks such as TensorRT [6]. Looking ahead, the continued evolution of intelligent (autonomous decision-making), miniaturized (portable deployment), and collaborative (multi-UAV networking) technologies will create new opportunities. Algorithms incorporating federated learning and self-supervised learning are expected to overcome current limitations and accelerate the deployment of UAV object detection systems in real-world scenarios such as emergency response and environmental monitoring [7].

Thanks to their flexible deployment and wide-area coverage capabilities, UAV surveillance systems have demonstrated irreplaceable value in fields such as disaster relief, traffic inspection, and border security [8]. However, challenging monitoring scenarios characterized by dynamic targets (e.g., fast-moving vehicles), frequent UAV viewpoint changes (alternating between nadir and oblique views), and dense distributions of small objects (e.g., individuals in a crowd) demand stronger environmental adaptability from detection algorithms [9]. Traditional object detection and tracking approaches based on handcrafted features (e.g., HOG, SIFT) suffer from limited feature representation and are incapable of coping with scale variation and motion blur. Furthermore, their reliance on high-computation platforms makes them unsuitable for lightweight deployment on UAV edge devices [10–12]. Therefore, developing real-time object detection and tracking algorithms that deliver both high accuracy and low latency is a crucial breakthrough to advance the practical application of UAV surveillance technologies.

2. Related technologies and methods

2.1. Mainstream object detection algorithms

Current object detection algorithms are generally categorized into two types: One-stage detectors and two-stage

detectors [13]. The YOLO (You Only Look Once) series formulates object detection as a regression problem, simultaneously performing bounding box regression and class prediction within a single neural network. Notable versions such as YOLOv3, YOLOv5, and YOLOv8 offer high detection speed and strong real-time performance, making them well-suited for deployment on resource-constrained UAV platforms. Faster R-CNN, a two-stage detector, first uses a Region Proposal Network (RPN) to generate candidate regions, followed by classification and regression. Although it achieves higher detection accuracy, its computational cost is relatively high, making it less suitable for real-time applications. The Single Shot MultiBox Detector (SSD) performs multi-scale predictions, balancing speed and accuracy. Compared with YOLO, SSD has better performance in detecting small objects; however, its stability in complex scenes still requires improvement. Overall, the YOLO series—owing to its efficiency and deployment flexibility—has become the preferred solution for real-time object detection in UAV-based surveillance systems.

2.2. Object tracking algorithms

Detection-based tracking methods first detect the target's position using an object detector, followed by data association strategies to match and track the target across frames. A representative algorithm in this category is Deep SORT (Simple Online and Realtime Tracking with a Deep Association Metric) [14]. This method integrates appearance features with motion information and utilizes a Kalman filter along with the Hungarian algorithm to achieve robust and real-time multi-object tracking. Deep SORT maintains consistent target IDs while supporting long-term tracking and high frame rates, making it one of the most widely adopted tracking solutions in current UAV vision systems [15,16].

3. System design and algorithm framework

3.1. Overall system architecture

The UAV-based surveillance system designed in this study comprises three core components: The UAV acquisition module, the edge computing unit, and the ground processing system, as illustrated in **Figure 1**.

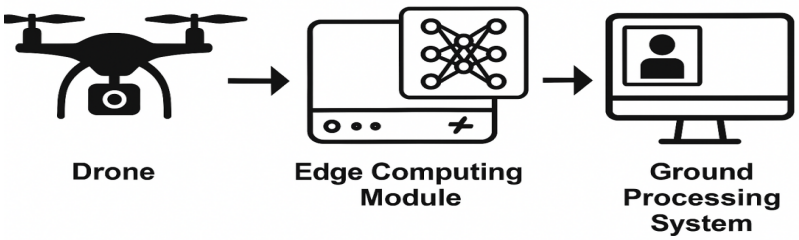

Figure 1. UAV-based surveillance system

The UAV acquisition module is equipped with a high-definition camera to capture real-time video streams of ground scenes. The camera connects to the edge computing unit via USB or CSI interfaces, ensuring high

frame rates and low-latency video transmission.

The edge computing unit is implemented on an embedded AI module onboard the UAV, such as the NVIDIA Jetson Xavier NX. It is responsible for executing real-time object detection and tracking algorithms, enabling on-device inference and local decision-making. This module supports GPU acceleration and is compatible with inference engines such as TensorRT and ONNX, ensuring efficient and stable algorithm deployment.

The ground processing system communicates with the UAV via 5G or Wi-Fi networks, receiving essential data such as object location, class, and trajectory for further analysis, visualization, or task scheduling. Additionally, the ground station can remotely control the UAV's flight path, enabling human-machine collaborative task execution.

3.2. Object detection module

This system adopts an optimized version of YOLOv5 as the core detection model. YOLOv5 offers advantages such as lightweight architecture, multi-scale detection, and high real-time performance, making it suitable for embedded deployment scenarios. To further meet the stringent computational constraints of UAV platforms, the following optimization strategies are employed: Lightweight Network Architecture: MobileNetV3 is adopted as the backbone to replace the original CSPDarknet53, reducing model parameters and computational load. Model pruning: Structured pruning is applied to eliminate redundant channels and convolutional kernels, resulting in a more compact model and faster inference speed. Knowledge Distillation: A teacher-student framework is used, where a high-performance teacher model guides the training of a lightweight student model, thereby improving model efficiency without sacrificing accuracy.

To enhance small object detection performance, this study integrates a Bidirectional Feature Pyramid Network (BiFPN) into the original YOLOv5 Feature Pyramid Network (FPN), improving cross-scale feature fusion. Additional enhancements include: Contextual Attention Module: An attention mechanism is introduced to emphasize the contrast between background and object edges in low-resolution feature maps, improving perceptual capability. Super-Resolution Anchor Design: Prior anchor box distributions are optimized to better match the size characteristics of small targets in aerial imagery, effectively reducing the miss detection rate. These improvements enable the detection module to maintain high recognition accuracy and localization precision even under complex backgrounds, long-range viewpoints, and multi-scale object conditions.

3.3. Object tracking module

To achieve robust and drift-resistant object tracking, the system utilizes an enhanced version of the Deep SORT multi-object tracking framework, which fuses both appearance and motion information. Motion Modeling: A Kalman filter is employed to estimate target state variables such as position and velocity, predicting the potential location of targets in the current frame. Re-ID Network: A lightweight appearance embedding network is designed to extract features such as color and texture, facilitating identity matching and preventing ID switching due to occlusion or intersection. Data Association Algorithm: The Hungarian algorithm is used to perform optimal matching between current detections and historical trajectories, incorporating weighted distances based on position, velocity, and appearance features.

Due to the complex top-down perspective of UAVs and frequent occlusion from buildings or trees, the following mechanisms are introduced: Occlusion Detection Mechanism: Occlusion is dynamically determined based on Intersection over Union (IoU) and matching scores. Target disappearance is classified as "lost" based on trajectory duration and confidence thresholds. Trajectory Recovery Strategy: Following short-term occlusions, a

"cold start" process is used to reinitialize the trajectory based on recent predictions and detection results. Appearance similarity is leveraged to merge and reconstruct interrupted tracks. These mechanisms effectively mitigate ID switching and erroneous deletions caused by occlusions, significantly improving the stability of target tracking.

4. Experimental design and result analysis

4.1. Experimental setup and baseline methods

To comprehensively evaluate the performance of the proposed method, several mainstream object detection and tracking algorithms were selected as baselines: Object Detection Algorithms: YOLOv5 (Baseline version), Faster R-CNN (a representative two-stage detector), SSD (a lightweight detection model), Object Tracking Algorithms: Deep SORT (a classic detection-based multi-object tracking framework), ByteTrack (a high-performance, low-latency tracking method).

Furthermore, comparative experiments were conducted in a variety of typical application scenarios, including daytime and nighttime environments, occluded and non-occluded conditions, as well as sparse and crowded scenes. These tests were designed to evaluate the proposed method's adaptability and robustness in real-world settings.

4.2. Experimental result analysis

The improved YOLOv5 model achieved high detection accuracy while maintaining real-time processing speed, making it suitable for UAV-based online monitoring tasks. **Table 1** presented a comparison of detection performance among different methods.

Table 1. Detection accuracy comparison of different methods

Method	mAP@0.5	Precision	Recall	FPS
Faster R-CNN	75.4%	77.2%	74.1%	12
YOLOv5s (Original)	82.1%	84.3%	79.8%	65
Proposed method (Improved YOLOv5 + Feature enhancement)	86.7%	88.5%	83.2%	72

5. Conclusion

This paper addresses the challenges of low accuracy and poor real-time performance in UAV-based surveillance for object detection and tracking. A unified detection-tracking framework is proposed, integrating a lightweight improved YOLOv5s model with adaptive motion compensation. In terms of model architecture, the introduction of the MobileNetV3 backbone, multi-scale feature fusion, and attention mechanisms enhances the detection of small and blurry targets. For the tracking module, the integration of Deep SORT with a dynamic compensation strategy significantly improves the system's robustness under high-speed motion and complex backgrounds.

Experimental validation demonstrates that the proposed method achieves a high mAP@0.5 on the VisDrone dataset while maintaining real-time inference at 32 FPS on the Jetson TX2 platform. This balances lightweight deployment and real-time performance, making it well-suited for on-board UAV monitoring tasks. The research findings suggest that the proposed method offers strong technical support for UAV vision systems in scenarios such as emergency rescue and traffic inspection. It also lays a solid foundation for future studies on miniaturized, collaborative, and multimodal intelligent UAV systems.

Disclosure statement

The authors declare no conflict of interest.

References

[1] Liu C, Li S, Meng F, et al., 2023, A Review of UAV Object Detection Based on Deep Learning. Computer Science and Applications, 13(5): 1092–1099.

[2] Liu G, Zeng X, Dou J, et al., 2024, Review of Advances in Small Object Detection Technology based on Deep Learning. Infrared and Laser Engineering, 53(9): 184–216.

[3] Li X, Zhou Y, Wang F, 2022, Advanced Information Mining from Ocean Remote Sensing Imagery with Deep Learning. Journal of Remote Sensing, 9849645.

[4] Chinese Academy of Sciences, Literature and Information Center, 2024, Scientific Literature Data Mining to Enhance Research Efficiency and Large Model Training, Shijiazhuang Data Bureau, viewed October 10, 2024. https://sjj.sjz.gov.cn/columns/9bcbfa10-a61f-4005-b732-234ea94403d7/202410/10/6fb6230c-96a8-40f1-801d-9c77c4200f61.html

[5] Girshick R, Donahue J, Darrell T, et al., 2014, Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation. 2014 IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 580–587.

[6] Redmon J, Divvala S, Girshick R, et al., 2016, You Only Look Once: Unified, Real-Time Object Detection. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2016, 779–788.

[7] Bochkovskiy A, Wang CY, Liao HYM, 2020, YOLOv4: Optimal Speed and Accuracy of Object Detection. arXiv preprint. https://doi.org/10.48550/arXiv.2004.10934

[8] Xie L, Xue Y, Ye J, 2022, UAV Aerial Target Detection Algorithm Based on Improved YOLOv5. Acta Automatica Sinica, 2284(1): 012024.

[9] Peng M, Liu J, Hu Z, 2023, Research on UAV Target Detection Based on YOLOv5. Computer Engineering and Applications, 59(5): 157–162.

[10] Zheng H, Zhang J, Hu X, et al., 2022, UAV Ground Target Detection Method Based on Improved YOLOv4. Ordnance Automation, 41(7): 89–94.

[11] Wang L, Zou Y, 2022, Overview of Object Detection Algorithms Based on Deep Learning. Computer Science and Exploration, 16(9): 1521–1531.

[12] Jiang W, Wang M, 2022, Research on Ground Target Detection and Tracking Based on YOLOv5. Modern Electronic Technology, 45(23): 123–127.

[13] Sun K, Li Z, Qi L, 2021, Design of UAV Image Recognition System Based on Embedded Platform. Electronic Technology, 34(4): 56–60.

[14] Ge Z, Liu S, Wang F, et al., 2021, YOLOX: Exceeding YOLO Series in 2021. arXiv preprint. https://doi.org/10.48550/arXiv.2107.08430

[15] Redmon J, Farhadi A, 2018, YOLOv3: An Incremental Improvement. arXiv preprint. https://doi.org/10.48550/arXiv.1804.02767

[16] Du D, Qi H, Yu Q, et al., 2018, The Unmanned Aerial Vehicle Benchmark: Object Detection and Tracking. European Conference on Computer Vision (ECCV), 370–386.