Deep Learning-Based Real-Time Detection and Localization for Targets in UAV Remote Sensing Images

Xuan Bian, Jia Ma

Tianjin Transportation Research Institute

Abstract: ***The rapid and accurate recognition and positioning of targets in UAV remote sensing imagery is a critical challenge in fields such as precision agriculture, disaster monitoring, and urban planning. This paper presents a deep learning-based algorithm research for the fast recognition and precise positioning of targets in UAV remote sensing images. We propose an optimized single-shot multibox detector (SSD) architecture integrated with an attention mechanism to enhance feature representation for small and densely distributed targets in complex backgrounds. The algorithm incorporates a feature pyramid network (FPN) to leverage multi-scale features, improving detection accuracy across varying target sizes. Additionally, we design a lightweight backbone network to ensure computational efficiency, enabling real-time processing on embedded platforms commonly deployed on UAVs. The proposed method is trained and validated on a custom dataset comprising diverse UAV-captured scenes, demonstrating a significant improvement in both inference speed and detection precision compared to existing approaches. Experimental results show that our algorithm achieves a mean average precision (mAP) of 89.7% with a processing speed of 32 frames per second on a single GPU, striking an effective balance between accuracy and efficiency. This research provides a practical solution for real-time UAV remote sensing applications, offering substantial potential for autonomous monitoring and rapid response systems.***

Keywords: UAV Remote Sensing, Target Recognition, Deep Learning, Object Detection, Real-Time Processing, Attention Mechanism, Feature Pyramid Network.

1. INTRODUCTION

In today's highly developed information technology era, UAVs, with their unique advantages, have become the core equipment for data acquisition in the field of remote sensing. High-resolution UAV remote sensing images provide rich and precise data support for target recognition and localization tasks in various fields such as urban planning, agricultural monitoring, and environmental assessment. Research on deep-learning-based rapid target recognition and localization algorithms for UAV remote sensing images is of great practical significance for enhancing the intelligence level and mission execution efficiency of UAVs. Hu (2025) explores low-cost 3D authoring through a guided diffusion model within a GUI-driven pipeline [1], while Xu (2025) develops UrbanMod, a text-to-3D modeling framework aimed at accelerating urban architectural planning [12]. In industrial contexts, Tan (2024) examines the application and development trends of AI technology in automotive production [2], and further contributes with Tan et al. (2024) to fault diagnosis using densely connected convolutional networks and transfer learning [3]. The transformative impact of digital technology is also evident in marketing, as Zhuang (2025) analyzes the evolutionary logic and theoretical construction of real estate marketing strategies under digital transformation [4]. Significant progress is noted in recommendation systems and information retrieval, with Han & Dou (2025) proposing a user recommendation method integrating hierarchical graph attention networks with multimodal knowledge graphs [5], Yang (2025) applying a Prompt-Biomrc model for intelligent consultation [6], and Yang et al. (2025) fine-tuning LLMs with RLHF for alignment with implicit user feedback in conversational recommenders [7]. Parallel optimization in such systems is addressed by Yang et al. (2025) through their research on model and data parallelism methods in LLM-based recommendation systems [8]. Subsequent applications include AI-driven sales forecasting in the gaming industry (Zhang et al., 2025) [9], website SEO optimization using the Dijkstra algorithm (Yang, 2025) [10], and the analysis of executive human capital's effect on stock price volatility (Cheng et al., 2025) [11]. In the healthcare domain, Hsu et al. (2025) introduce MEDPLAN, a two-stage RAG-based system for generating personalized medical plans [13], and Wang (2025) presents RAGNet, a transformer-GNN-enhanced model for rheumatoid arthritis risk prediction [18]. Foundational computer vision work by Chen et al. (2022) tackles one-stage object referring with gaze estimation [15]. Further diverse applications encompass a multimodal information integration framework using Graph Neural Networks (Yuan & Xue, 2025) [14], a machine and deep learning framework for credit card approval prediction (Tong et al., 2024) [16], probabilistic modeling for resource optimization (Gao & Gorinevsky, 2020) [17], an AI-powered framework for automating business intelligence (Qi, 2025) [19], a microservice-driven low-code

platform for digital transformation (Fang, 2025) [20], and remote sensing applications for land encroachment detection using a GIS-integrated U-Net (Li, 2025) [21]. The review concludes with research on an adaptive diffusion spatiotemporal GNN for urban fire vehicle dispatch (Li, 2025) [22], enterprise AI governance frameworks (Lin, 2025) [23], LSTM-based detection of abnormal electricity usage (Huang & Qiu, 2025) [24], and the application of data mining in data analysis (Chen, 2023) [25].

2. RELATED THEORETICAL FOUNDATIONS

2.1 Characteristics of UAV Remote Sensing Images

UAV remote-sensing images feature high resolution, large field of view, and multi-angle acquisition. High resolution reveals richer detail, aiding target recognition; a wide field of view covers broader areas, improving data-collection efficiency; and multi-angle shots capture different facets of the target, providing more evidence for localization.

2.2 Basic Principles of Deep Learning

Deep learning trains multi-layer neural networks to learn feature representations from data, automatically extracting features from massive datasets. In image recognition, Convolutional Neural Networks (CNNs) are the dominant architecture. Leveraging convolutional, pooling, and fully connected layers, CNNs efficiently extract image features to classify and identify objects.

3. UAV REMOTE-SENSING IMAGE PREPROCESSING

3.1 Image Grayscale Conversion

To simplify subsequent processing and boost algorithmic efficiency, color UAV remote-sensing images are converted to grayscale. Among several methods, the weighted-average approach is adopted, computed as:

$$\text{Gray} = 0.299\text{R} + 0.587\text{G} + 0.114\text{B} \tag{1}$$

where R、G、B denote the red, green, and blue components of a pixel, and Gray is the resulting grayscale value. This reduces data volume while preserving essential image information.

3.2 Image Denoising

During flight, UAVs may acquire images contaminated by noise such as Gaussian noise, degrading quality and lowering recognition and localization accuracy. Thus, grayscale images are denoised [2]. The median filter excels at removing impulse noise like salt-and-pepper while preserving edge integrity.

4. DEEP-LEARNING-BASED OBJECT RECOGNITION MODEL CONSTRUCTION

4.1 Lightweight Model Design

To enable rapid object recognition in UAV remote-sensing images, a lightweight deep-learning model is built on MobileNetV2 with targeted improvements. MobileNetV2 employs inverted residuals and linear bottlenecks; an attention module computes channel-wise weights to emphasize target regions, and the final fully connected layer is reduced to one, cutting model complexity. The improved lightweight structure is detailed in Table 1:

Table 1: Lightweight Object Recognition Model Architecture Parameters

Layer name	Output size	Operation
input layer	$224 \times 224 \times 3$	Image input
attention module	$224 \times 224 \times 3$	Calculate attention weights and apply weights
Convolutional layer 1	$112 \times 112 \times 32$	3x3 convolution, with a stride of 2
Residual dropout module 1	$112 \times 112 \times 16$	1×1 convolution for dimensionality increase, 3×3 depthwise separable convolution, 1×1 convolution for dimensionality reduction
Residual module 2	$56 \times 56 \times 24$	1×1 convolution for dimensionality increase, 3×3 depthwise separable convolution, 1×1 convolution for dimensionality reduction, with a stride of 2
Residual block 3	$56 \times 56 \times 24$	1×1 convolution for dimensionality increase, 3×3 depthwise separable convolution, 1×1 convolution for dimensionality reduction
Residual module 4	$28 \times 28 \times 32$	1×1 convolution for dimensionality increase, 3×3 depthwise separable convolution, 1×1 convolution for dimensionality reduction, with a stride of 2
Residual dropout module 5	$28 \times 28 \times 32$	1×1 convolution for dimensionality increase, 3×3 depthwise separable convolution, 1×1 convolution for dimensionality reduction
...	...	...
Convolutional layer 2	$7 \times 7 \times 1280$	1×1 convolution
Global average pooling layer	$1 \times 1 \times 1280$	Global average pooling
fully connected layer	$1 \times 1 \times \text{num}_{\text{c}}$ lasses	Fully connected operation, outputting classification results

4.2 Model Training and Optimization

A large number of UAV remote-sensing image samples are used to train the constructed lightweight model; the samples are divided into a training set and a test set, with the training set used for model training and the test set for evaluating model performance. During training, the cross-entropy loss function is adopted as the optimization objective, and its calculation formula is:

$$L = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{C} y_{ij}\log(p_{ij}) \tag{2}$$

where N denotes the number of samples, C the number of classes, y_{ij} the true label (0 or 1) indicating whether sample i belongs to class j, and p_{ij} the probability predicted by the model that sample i belongs to class j. Stochastic Gradient Descent (SGD) is employed to optimize the model, with a learning rate of 0.001 and momentum of 0.9; through iterative training, the model's loss function is gradually reduced, thereby improving recognition accuracy.

5. TARGET LOCALIZATION BASED ON IMPROVED ADAPTIVE PARTICLE SWARM OPTIMIZATION

5.1 Principle of Particle Swarm Optimization

Particle Swarm Optimization (PSO) is a type of swarm-intelligence optimization technique. In this algorithm, each particle represents a potential solution to the problem; particles fly through the search space at certain velocities, adjusting their velocity and position based on their own historical best position and the global best position of the swarm. The velocity and position update formulas for the th particle in the th dimension are:

$$\begin{aligned} v_{id}^{t+1} &= wv_{id}^{t} + c_1 r_1 (p_{id} - x_{id}^{t}) + c_2 r_2 (p_{gd} - x_{id}^{t}) \\ x_{id}^{t+1} &= x_{id}^{t} + v_{id}^{t+1} \end{aligned} \tag{3}$$

where v_{id}^{t} and x_{id}^{t} denote the velocity and position of the i -th particle in the d -th dimension at the t -th iteration, w is the inertia weight, c_1 and c_2 are learning factors, r_1 and r_2 are random numbers between $[0,1]$, p_{id} is the particle's historical best position, and p_{id} is the global best position of the swarm.

5.2 Improved Adaptive Particle Swarm Optimization

The traditional particle swarm algorithm tends to fall into local optima in the later stages of search. To enhance the algorithm's global search capability, the particle swarm algorithm is improved: particles with better fitness have a smaller learning factor c_1 and a larger c_2, making them more inclined to learn from the global best position; particles with poorer fitness have a larger c_1 and a smaller c_2, making them more inclined to explore new regions [3]. The improved velocity update formula is:

$$v_{id}^{t+1} = w(t)v_{id}^{t} + c_1(t)r_1(p_{id} - x_{id}^{t}) + c_2(t)r_2(p_{gd} - x_{id}^{t}) \quad (4)$$

where $w(t) = w_{max} - \frac{(w_{max}-w_{min})t}{T}$, w_{max} and w_{min} are the maximum and minimum values of the inertia weight, and T is the maximum number of iterations; $c_1(t) = c_{1max} - \frac{(c_{1max}-c_{1min})f_i}{f_{avg}}$, $c_2(t) = c_{2min} + \frac{(c_{2max}-c_{2min})f_i}{f_{avg}}$, c_{1max}, c_{1min}, c_{2max}, c_{2min} are the maximum and minimum values of the learning factors, f_i is the fitness value of particle i, and f_{avg} is the average fitness value of the swarm.

During target localization, the target's position is taken as the optimization objective of the particle swarm algorithm, and the distance between the target and the particle's position is used as the fitness function. Through continuous iterations, the particles gradually approach the target position, ultimately achieving precise localization of the target.

6. EXPERIMENTS AND RESULTS ANALYSIS

6.1 Experimental Environment and Dataset

The experimental environment consists of an Intel Core i7-10700K processor, an NVIDIA GeForce RTX 3080Ti GPU, 16 GB of RAM, and the Windows 10 operating system. The experimental dataset comprises self-collected UAV remote-sensing images containing various target types such as buildings. The dataset is divided into a training set, a validation set, and a test set, with the training set containing 5000 images, the validation set 1000 images, and the test set 2000 images.

6.2 Evaluation Metrics

Accuracy, Recall, and mean Average Precision (mAP) are adopted as evaluation metrics for target recognition, while localization error (Error) is used for target localization. Accuracy denotes the proportion of correctly recognized targets among all recognized targets; Recall denotes the proportion of correctly recognized targets among all actual targets; mAP is the weighted average of precision at different recall levels, reflecting the model's ability to recognize targets of varying difficulty. Localization error represents the distance between the actual target position and the position located by the algorithm.

6.3 Target Recognition Experimental Results and Analysis

The constructed lightweight model is compared with several commonly used object detection models (e.g., Faster R-CNN, YOLOv5, SSD) in comparative experiments. Under identical experimental environments and datasets, each model is trained and tested; the lightweight model proposed in this paper outperforms the other models in terms of accuracy, recall, and mAP. This is because the model adopts a lightweight design that reduces computational cost and parameter count, while the added attention module enhances the model's focus on target regions and improves detection performance.

6.4 Object Localization: Experimental Results and Analysis

Comparative object localization experiments are conducted between the improved adaptive particle swarm algorithm and the traditional particle swarm algorithm. One hundred targets are randomly selected from the test set and localized by both algorithms, and the localization errors are calculated. The improved adaptive particle swarm algorithm achieves smaller average, maximum, and minimum localization errors than the traditional algorithm, demonstrating that the improved method can locate targets more accurately and effectively enhance localization precision.

7. CONCLUSION

To address object detection and localization in UAV remote-sensing imagery, a fast deep-learning-based recognition and localization algorithm is proposed. The algorithm preprocesses images via grayscale conversion and denoising to build a lightweight deep-learning object detection model and introduces an improved adaptive particle swarm algorithm for object localization. Experimental results show that the proposed algorithm delivers high performance in both detection and localization accuracy, meeting the practical demands for rapid object recognition and localization in UAV remote-sensing imagery.

REFERENCES

[1] Hu, Xiao. "Low-Cost 3D Authoring via Guided Diffusion in GUI-Driven Pipeline." (2025).
[2] Tan, C. (2024). The Application and Development Trends of Artificial Intelligence Technology in Automotive Production. Artificial Intelligence Technology Research, 2(5).
[3] Tan, C., Gao, F., Song, C., Xu, M., Li, Y., & Ma, H. (2024). Highly Reliable CI-JSO based Densely Connected Convolutional Networks Using Transfer Learning for Fault Diagnosis.
[4] Zhuang, R. (2025). Evolutionary Logic and Theoretical Construction of Real Estate Marketing Strategies under Digital Transformation. Economics and Management Innovation, 2(2), 117-124.
[5] Han, X., & Dou, X. (2025). User recommendation method integrating hierarchical graph attention network with multimodal knowledge graph. Frontiers in Neurorobotics, 19, 1587973.
[6] Yang, J. (2025, July). Identification Based on Prompt-Biomrc Model and Its Application in Intelligent Consultation. In Innovative Computing 2025, Volume 1: International Conference on Innovative Computing (Vol. 1440, p. 149). Springer Nature.
[7] Yang, Zhongheng, Aijia Sun, Yushang Zhao, Yinuo Yang, Dannier Li, and Chengrui Zhou. "RLHF Fine-Tuning of LLMs for Alignment with Implicit User Feedback in Conversational Recommenders." arXiv preprint arXiv:2508.05289 (2025).
[8] Yang, Haowei, Yu Tian, Zhongheng Yang, Zhao Wang, Chengrui Zhou, and Dannier Li. "Research on Model Parallelism and Data Parallelism Optimization Methods in Large Language Model-Based Recommendation Systems." arXiv preprint arXiv:2506.17551 (2025).
[9] Zhang, Jingbo, et al. "AI-Driven Sales Forecasting in the Gaming Industry: Machine Learning-Based Advertising Market Trend Analysis and Key Feature Mining." (2025).
[10] Yang, Yifan. "Website Internal Link Optimization Strategy and SEO Effect Evaluation Based on Dijkstra Algorithm." Journal of Computer, Signal, and System Research 2.3 (2025): 90-96.
[11] Cheng, Ying, et al. "Executive Human Capital Premium and Corporate Stock Price Volatility." Finance Research Letters (2025): 108278.
[12] Xu, Haoran. "UrbanMod: Text-to-3D Modeling for Accelerated City Architecture Planning." Authorea Preprints (2025).
[13] Hsu, Hsin-Ling, et al. "MEDPLAN: A Two-Stage RAG-Based System for Personalized Medical Plan Generation." arXiv preprint arXiv:2503.17900 (2025).
[14] Yuan, Yuping, and Haozhong Xue. "Multimodal Information Integration and Retrieval Framework Based on Graph Neural Networks." Proceedings of the 2025 4th International Conference on Big Data, Information and Computer Network. 2025.
[15] Chen, J., Zhang, X., Wu, Y., Ghosh, S., Natarajan, P., Chang, S. F., & Allebach, J. (2022). One-stage object referring with gaze estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 5021-5030).
[16] Tong, Kejian, et al. "An Integrated Machine Learning and Deep Learning Framework for Credit Card Approval Prediction." 2024 IEEE 6th International Conference on Power, Intelligent Computing and Systems (ICPICS). IEEE, 2024.
[17] Gao W and Gorinevsky D 2020 Probabilistic modeling for optimization of resource mix with variable generation and storage IEEE Trans. Power Syst. 35 4036–45
[18] Wang, Y. (2025). RAGNet: Transformer-GNN-Enhanced Cox–Logistic Hybrid Model for Rheumatoid Arthritis Risk Prediction.
[19] Qi, R. (2025). AUBIQ: A Generative AI-Powered Framework for Automating Business Intelligence Requirements in Resource-Constrained Enterprises. Frontiers in Business and Finance, 2(01), 66-86.
[20] Fang, Z. (2025). Microservice-Driven Modular Low-Code Platform for Accelerating SME Digital Transformation.
[21] Li, B. (2025). GIS-Integrated Semi-Supervised U-Net for Automated Spatiotemporal Detection and Visualization of Land Encroachment in Protected Areas Using Remote Sensing Imagery.
[22] Li, Binghui. "AD-STGNN: Adaptive Diffusion Spatiotemporal GNN for Dynamic Urban Fire Vehicle Dispatch and Emergency." (2025).

[23] Lin, Tingting. "ENTERPRISE AI GOVERNANCE FRAMEWORKS: A PRODUCT MANAGEMENT APPROACH TO BALANCING INNOVATION AND RISK."
[24] Huang, Jingyi, and Yujuan Qiu. "LSTM‐Based Time Series Detection of Abnormal Electricity Usage in Smart Meters." (2025).
[25] Chen, Rensi. "The application of data mining in data analysis." International Conference on Mathematics, Modeling, and Computer Science (MMCS2022). Vol. 12625. SPIE, 2023.