Adaptive Transformer-Based Multi-Object Detection Using Hybrid Metaheuristic Optimization

R. Mallikka[1]

[1] Post Doctoral Fellow

Lincoln University College, Petaling Jaya, Selangor Darul Ehsan-47301, Malaysia.
Email: pdf.mallikka@lincoln.edu.my

[1] Assistant Professor, Department of Computer Science and Engineering,
SRM Institute of Science and Technology Tiruchirappalli, Tamil Nadu, India
Email mallikka2002@gmail.com

Abstract

Detecting multiple objects in surveillance scenes is challenging due to scale variations, occlusions, background clutter, and changing illumination, which limit the performance of existing object detection models. To address these challenges, this paper proposes an Adaptive Transformer-Based Multi-Object Detection Framework optimized using a hybrid Fire Hawk Optimization (FHO) and Arithmetic Optimization Algorithm (AOA). The proposed framework integrates Contrast Limited Adaptive Histogram Equalization (CLAHE)-based image enhancement, Gaussian noise normalization, a Vision Transformer with multi-scale feature fusion, and channel-spatial attention to extract discriminative features and achieve precise object localization. The hybrid optimization strategy employs FHO for effective global exploration and AOA for fine-grained parameter refinement during the exploitation stage, resulting in improved convergence and detection performance. Experimental evaluation on the VisDrone dataset demonstrates superior performance over state-of-the-art CNN- and transformer-based baseline models, achieving higher Precision, Recall, F1-Score, mAP@0.5, and Intersection over Union (IoU). These results indicate that the proposed framework provides a reliable and efficient solution for real-time surveillance applications, including intelligent traffic monitoring, public safety, smart city infrastructure, autonomous surveillance, and security management.

Keywords: Multi-object detection; Vision Transformer; Metaheuristic optimization; Arithmetic Optimization Algorithm; Fire Hawk Optimization; VisDrone

Introduction

Surveillance systems rely on accurate object detection, but performance degrades under occlusion, scale variation, background clutter, and illumination changes [1]. CNN-based detectors such as YOLO and Faster R-CNN perform well in standard conditions, yet their limited receptive fields restrict modeling of long-range object relationships, particularly in drone surveillance where objects are small, dense, and frequently occluded [2][3].

Vision Transformers (ViTs) address this limitation using self-attention to model dependencies between image patches regardless of spatial distance. Studies such as UAV-DETR [5] and TPH-YOLOv5 [7] demonstrate that transformer-based detectors outperform CNN models on VisDrone. Integrating ViT with multi-scale feature fusion further improves detection across varying object sizes [4].

However, transformer hyperparameter tuning remains computationally challenging. Gradient-based optimizers often converge to local optima in high-dimensional search spaces. To address this, the proposed framework combines Fire Hawk Optimization (FHO) for broad early exploration with

Arithmetic Optimization Algorithm (AOA) [8] for precise local refinement. This hybrid strategy forms the core optimization contribution of the work.

The main contributions are:

• CLAHE-based contrast enhancement and Gaussian noise normalization for adaptive preprocessing.

• Channel and spatial attention mechanisms to suppress background noise and improve feature discrimination.

• A hybrid FHO-AOA optimizer that enhances detection accuracy and convergence compared with conventional optimization methods.

The remainder of the paper is organized as follows: Section 2 reviews related work, Section 3 presents the methodology, Section 4 discusses experimental results, Section 5 outlines limitations, and Section 6 concludes the study.

Related work

Previous studies on aerial and surveillance object detection mainly focus on transformer-based detection, metaheuristic optimization, and UAV-specific models.

Transformer-Based Object Detection

DHQ-DETR [4] improved object detection on COCO and VisDrone using distribution-based bounding box prediction, while UAV-DETR [5] enhanced small-object detection through multi-scale feature fusion and frequency-based downsampling. Although both methods demonstrate the effectiveness of transformers for drone imagery, they do not address hyperparameter optimization.

Metaheuristic Optimization in Deep Learning

Masaleno et al. [1] used a genetic algorithm to optimize YOLOv8 hyperparameters for baggage detection, achieving 91.3% mAP@0.5. Khadidos and Yafoz [2] applied the Dandelion Optimizer to an LSTM-AE model for real-time object detection with high accuracy and reduced execution time. These studies show the effectiveness of metaheuristic optimization, but they were not evaluated with ViT-based aerial detection models.

Drone and UAV Object Detection

TPH-YOLOv5 [7] integrated transformer heads and CBAM attention into YOLOv5, improving VisDrone2021 mAP by nearly 7%. Li et al. [3] proposed a lightweight ViT for pedestrian detection on constrained hardware. However, existing works do not combine ViT-based detection with hybrid metaheuristic optimization, which is the main focus of the proposed framework.

Proposed methodology

The proposed AOA-FHO ViT framework consists of five stages: preprocessing, ViT-based detection with multi-scale fusion, attention-based refinement, hybrid FHO-AOA optimization, and detection output generation. Figure 1 shows the overall architecture of the proposed AOA-FHO ViT framework showing the five-stage detection pipeline from preprocessed input to NMS-filtered detection output

Figure 1. Proposed AOA-FHO ViT framework for multi-object detection.

Data acquisition and preprocessing

The framework uses COCO 2017 [9] and VisDrone-2019-DET [10] datasets. CLAHE enhances local contrast while preserving details in bright and dark regions. Gaussian noise normalization with adaptive kernel estimation removes sensor noise without affecting object edges.

Vision transformer with multi-scale feature fusion

The ViT backbone divides images into patches, converts them into token embeddings, and processes them using multi-head self-attention to capture long-range dependencies. A Multi-Scale Feature Fusion (MSFF) module combines features from multiple transformer stages using an FPN-style pyramid, improving detection of both small and large objects.

Channel and spatial attention mechanisms

A Squeeze-and-Excitation channel attention block recalibrates feature importance using global statistics, while spatial attention identifies important spatial regions through 2D convolutions. These lightweight modules improve focus on true object regions and suppress background interference.

Hybrid FHO-AOA parameter optimization

The hybrid optimizer tunes learning rate, attention head count, patch size, and FPN fusion weights. FHO performs broad early exploration to avoid local optima, while AOA refines promising solutions during later stages. A linearly decaying exploration factor controls the transition between exploration and exploitation. Optimization is guided by a fitness function combining mAP@0.5 and F1-Score.

Detection output

The detection head generates bounding boxes and class probabilities for candidate objects. Non-Maximum Suppression (NMS) removes redundant detections. Performance is evaluated using Precision, Recall, F1-Score, mAP@0.5, mAP@0.5:0.95, and IoU.

Experimental results

Experimental setup

Experiments were conducted in Python using PyTorch on an NVIDIA RTX 3090 GPU, Intel Core i9 CPU, and 64 GB RAM. The ViT backbone used ImageNet-pretrained weights, while the hybrid FHO-AOA optimizer was executed for 50 iterations with a population size of 30, using 100 training epochs and batch size 16. Results were averaged over three runs with different random seeds.

VisDrone 2017 contains over 118K training images across 80 categories, while VisDrone-2019-DET includes drone-captured images with more than 540K annotated objects across 10 categories, characterized by dense distributions, scale variation, occlusion, and illumination changes. Figure 2 shows sample dataset from VisDrone showing the diversity of object scales, densities, and conditions encountered during training.

Figure 2. sample dataset VisDrone

Discussions

Table 1. Performance comparison on VisDrone dataset

Method	Precision (%)	Recall (%)	F1-Score (%)	mAP@0.5 (%)	IoU (%)
YOLOv5 [6]	78.4	74.2	76.2	72.1	68.3
TPH-YOLOv5 [7]	82.6	79.1	80.8	77.4	72.5
UAV-DETR [5]	84.3	81.0	82.6	79.8	74.2
DHQ-DETR [4]	85.1	82.4	83.7	81.2	75.6
Proposed (AOA-FHO ViT)	91.7	88.9	90.3	87.6	82.4

Table 2. Ablation study on VisDrone dataset

Configuration	Precision (%)	Recall (%)	F1-Score (%)	mAP@0.5 (%)
ViT only (no MSFF)	82.1	79.4	80.7	77.2
ViT + MSFF (no attention)	85.3	82.6	83.9	80.8
ViT + MSFF + Attention (Adam)	87.4	84.9	86.1	83.1
ViT + MSFF + Attention (AOA)	89.2	86.7	87.9	85.3
ViT + MSFF + Attention (FHO)	88.6	86.1	87.3	84.7
Proposed Full (AOA-FHO ViT)	91.7	88.9	90.3	87.6

On VisDrone (Table 1), the proposed method reaches 91.7% Precision, 88.9% Recall, 90.3% F1-Score, and 87.6% mAP@0.5. The nearest competitor, DHQ-DETR [4], scores 81.2% mAP@0.5, a gap of 6.4 points. YOLOv5 [6] trails by 15.5 mAP@0.5 points, which quantifies how much ViT-based detection with metaheuristic optimization gains over a standard CNN baseline on drone data.

Removing the MSFF module drops mAP@0.5 by 10.4 points from the full model. That is the largest single-component effect in the study, which makes sense: without multi-scale fusion, small objects in VisDrone frames are consistently missed because only one feature scale is available. Adding attention back raises mAP@0.5 by 2.3 points over the MSFF-only model. Switching from Adam to the hybrid FHO-AOA optimizer adds another 4.5 points, which is the second-largest gain. Running AOA or FHO alone scores 85.3% and 84.7% mAP@0.5 respectively; the hybrid reaches 87.6%, confirming the two algorithms genuinely complement each other. Figure 3 shows sample detection outputs from the proposed model on VisDrone scenes. Each bounding box is labeled with the predicted class and confidence score.

Limitations

The hybrid FHO-AOA optimizer increases training cost due to population-based iterations and has only been evaluated on static image datasets. Future work will focus on lightweight optimization strategies and real-time video-based surveillance with temporal modeling.

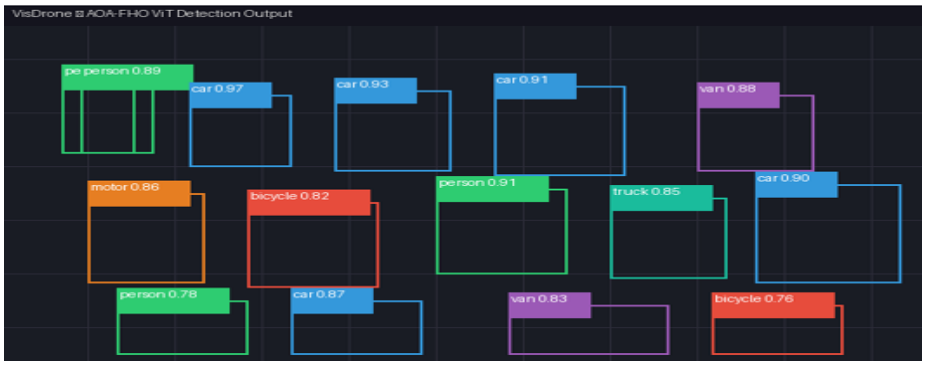

Fig. 3 Sample detection outputs on VisDrone

Conclusion

- Existing object detection methods struggle with scale variation, occlusion, background clutter, and illumination changes in surveillance scenes.
- The study aims to improve detection accuracy and robustness in complex environments.
- Proposed an Adaptive Transformer-Based Multi-Object Detection framework using CLAHE preprocessing, Vision Transformer with multi-scale feature fusion, channel-spatial attention, and hybrid FHO-AOA optimization.
- The proposed framework improved detection performance, achieving a 6.4% higher mAP@0.5 than DHQ-DETR on the VisDrone dataset.
- Ablation studies confirmed the effectiveness of multi-scale feature fusion and hybrid optimization.
- Future Work is to extend evaluation to more datasets and develop lightweight models for real-time edge deployment.

References

1. A. Masaleno, M. Huda, A. Fudholi, and C. Ratanamahatana, "Metaheuristic Hyperparameter Optimization and Explainable Deep Learning for Baggage Threat Detection," Emerging Science Journal, vol. 10, no. 1, Feb. 2026.
2. A. O. Khadidos and A. Yafoz, "Leveraging RetinaNet Based Object Detection Model for Assisting Visually Impaired Individuals with Metaheuristic Optimization Algorithm," Scientific Reports, May 2025. https://doi.org/10.1038/s41598-025-96068-8
3. J. Dai, J. Pan, F. Xu, and Z. Shen, "Lightweight ViT-Based Pedestrian Multi-Object Detection in Building Construction Scenarios," in Proc. 2nd Int. Conf. Machine Learning, IEEE, Sep. 2024.
4. C. Li, J. Zhang, B. Huo, and Y. Xue, "DHQ-DETR: Distributed and High-Quality Object Query for Enhanced Dense Detection in Remote Sensing," Remote Sensing, vol. 17, Feb. 2025. https://doi.org/10.3390/rs17040618
5. H. Zhang, K. Liu, Z. Gan, and G. Zhu, "UAV-DETR: Efficient End-to-End Object Detection for Unmanned Aerial Vehicle Imagery," arXiv:2501.01855, Jan. 2025.
6. G. Jocher et al., "YOLOv5 by Ultralytics," Zenodo, 2020. https://doi.org/10.5281/zenodo.3908559
7. X. Zhu, S. Lyu, X. Wang, and Q. Zhao, "TPH-YOLOv5: Improved YOLOv5 Based on Transformer Prediction Head for Object Detection on Drone-captured Scenarios," arXiv:2108.11539, Aug. 2021.
8. L. Abualigah et al., "The Arithmetic Optimization Algorithm," Computer Methods in Applied Mechanics and Engineering, vol. 376, Apr. 2021. https://doi.org/10.1016/j.cma.2020.113609
9. T. Lin et al., "Microsoft COCO: Common Objects in Context," in Proc. ECCV, 2014. https://doi.org/10.1007/978-3-319-10602-1_48
10. P. Zhu et al., "Vision Meets Drones: A Challenge," arXiv:1804.07437, 2018.