Computer Modeling in Engineering & Sciences

Tech Science Press

Doi:10.32604/cmes.2025.064783

ARTICLE

Aerial Object Tracking with Attention Mechanisms: Accurate Motion Path Estimation under Moving Camera Perspectives

Yu-Shiuan Tsai* **and Yuk-Hang Sit**

Department of Computer Science and Engineering, National Taiwan Ocean University, Keelung City, 202, Taiwan

*Corresponding Author: Yu-Shiuan Tsai. Email: ystsai@mail.ntou.edu.tw

Received: 24 February 2025; Accepted: 26 May 2025; Published: 30 June 2025

ABSTRACT: To improve small object detection and trajectory estimation from an aerial moving perspective, we propose the Aerial View Attention-PRB (AVA-PRB) model. AVA-PRB integrates two attention mechanisms—Coordinate Attention (CA) and the Convolutional Block Attention Module (CBAM)—to enhance detection accuracy. Additionally, Shape-IoU is employed as the loss function to refine localization precision. Our model further incorporates an adaptive feature fusion mechanism, which optimizes multi-scale object representation, ensuring robust tracking in complex aerial environments. We evaluate the performance of AVA-PRB on two benchmark datasets: Aerial Person Detection and VisDrone2019-Det. The model achieves 60.9% mAP@0.5 on the Aerial Person Detection dataset, and 51.2% mAP@0.5 on VisDrone2019-Det, demonstrating its effectiveness in aerial object detection. Beyond detection, we propose a novel trajectory estimation method that improves movement path prediction under aerial motion. Experimental results indicate that our approach reduces path deviation by up to 64%, effectively mitigating errors caused by rapid camera movements and background variations. By optimizing feature extraction and enhancing spatial-temporal coherence, our method significantly improves object tracking under aerial moving perspectives. This research addresses the limitations of fixed-camera tracking, enhancing flexibility and accuracy in aerial tracking applications. The proposed approach has broad potential for real-world applications, including surveillance, traffic monitoring, and environmental observation.

KEYWORDS: Aerial View Attention-PRB (AVA-PRB); aerial object tracking; small object detection; deep learning for Aerial vision; attention mechanisms in object detection; shape-IoU loss function; trajectory estimation; drone-based visual surveillance

1 Introduction

Tracking moving objects from an aerial perspective presents significant challenges due to the simultaneous motion of both the camera and the background. Traditional object tracking methods typically rely on comparing consecutive frames to estimate motion. However, in aerial imagery, this approach struggles to distinguish true object movement from background displacement caused by camera motion. Bewley et al. [1] demonstrated that camera-induced movement significantly affects tracking accuracy by introducing background shifts that obscure object trajectories. To mitigate this issue, researchers have proposed advanced tracking algorithms that decouple foreground motion from background artifacts [2–4]. These models demonstrate robustness in crowded scenes and under conditions of occlusion and noisy background motion, making them particularly suitable for aerial surveillance scenarios.

Recent advancements in computer vision have significantly improved object detection and tracking performance. Traditional object detection approaches relied on handcrafted features and classical machine learning techniques. However, deep learning-based models leveraging large-scale datasets and robust neural network architectures have demonstrated superior accuracy and efficiency [5]. Among these models, You Only Look Once (YOLO) has emerged as one of the most widely adopted frameworks for real-time object detection due to its computational efficiency and high precision [6]. The You Only Look Once (YOLO) series has undergone continuous evolution to enhance detection accuracy, processing speed, and robustness. YOLOv7 introduced optimized model scaling techniques that improved efficiency while maintaining real-time capabilities [7]. YOLOv8 incorporated advanced feature extraction mechanisms that refined small object detection while reducing computational complexity [8]. YOLOv9 addressed training stability issues by optimizing gradient flow, thereby improving model generalization across diverse datasets [9]. The latest iteration, YOLOv10, integrated a hybrid attention mechanism to enhance object localization and tracking performance in aerial imagery applications [10]. However, existing detection and tracking methods often struggle under specific conditions common in aerial imagery, such as low-resolution inputs, small object sizes, frequent occlusion, and high background complexity. These challenges become more severe when Unmanned Aerial Vehicles (UAVs) operate at high altitudes or in densely populated urban environments, resulting in reduced detection precision and tracking stability.

Aerial object tracking introduces unique challenges not typically encountered in ground-based detection tasks due to perspective variations, small object sizes, occlusion, background clutter, and dynamic environmental conditions. Li et al. [11] developed a model for detecting multiple targets in UAV-based aerial images, addressing key issues such as shape deformations and occlusions. Their study introduced modifications to the Bi-PAN-FPN structure within YOLOv8-s, improving multi-scale feature fusion while maintaining a compact parameter size. Additionally, their model incorporated GhostblockV2, which replaced sections of the C2f module, mitigating long-distance feature transfer loss and preserving crucial detection details [2]. To refine bounding box regression, Wise Intersection over Union (WIoU) loss was proposed, dynamically adjusting outlier distributions to stabilize predictions for small objects in aerial imagery [3,12].

Several tracking algorithms have been introduced to improve accuracy and robustness in multi-object tracking (MOT). The Simple Online and Realtime Tracking (SORT) algorithm, introduced by Bewley et al. [1], leveraged a Kalman filter and the Hungarian algorithm to achieve real-time tracking performance. Zhang et al. [2] proposed ByteTrack, an extension of SORT that maintained tracking continuity by utilizing both high-confidence and low-confidence detection boxes. Aharon et al. [3] developed BoT-SORT, which introduced re-identification features through CNN (Convolutional Neural Networks)-based visual embedding, improving tracking robustness in dynamic environments. More recently, Wang et al. [4] proposed SMILEtrack, which integrated a Patch Self-Attention (PSA) block and a Similarity Matching Cascade (SMC) module to enhance object matching across frames, particularly in complex aerial tracking scenarios.

To further enhance small object detection, researchers have proposed various architectural modifications. The Pyramid Residual Bidirectional Feature Pyramid Network (PRB-FPN), developed by Chen et al. [13], combines Feature Pyramid Networks (FPN), residual blocks, and bidirectional feature fusion, optimizing multi-scale feature representation. Wang et al. [14] introduced UAV-YOLOv8s, which incorporated WIoU v3, Focal FasterNet blocks (FFNB), and the BiFormer attention mechanism, significantly improving small object detection performance. Similarly, Li et al. [15] proposed YOLOv7-UAV, which replaced the traditional P5 prediction header with a P2 prediction header and removed redundant detection layers, thereby optimizing small object detection.

To address these challenges, we propose the Aerial View Attention-PRB (AVA-PRB) model, which integrates two key attention mechanisms: Coordinate Attention (CA) and the Convolutional Block Attention Module (CBAM). Coordinate Attention (CA) enhances the model's ability to capture positional information across spatial scales, thereby improving small object detection [16]. CBAM selectively refines feature maps by emphasizing informative regions, increasing detection robustness in complex aerial environments [17]. Additionally, we employ Shape-IoU as our loss function to improve bounding box alignment, particularly for small object detection tasks [18]. Our model also incorporates an adaptive feature fusion mechanism that dynamically refines multi-scale object representations, ensuring higher accuracy in aerial object tracking.

In addition to enhancing small object detection, an object movement path estimation technique is proposed to address the challenges of aerial tracking. Unlike conventional tracking methods, which often suffer from degraded accuracy when the camera itself is moving, our approach significantly enhances trajectory estimation. Experimental results show that this method improves tracking accuracy by up to 64% [19]. This advancement is crucial for UAV-based tracking systems, ensuring stable and reliable detection in real-world applications.

This work presents several key contributions. First, we propose the AVA-PRB model, which integrates advanced attention mechanisms to enhance small object detection in aerial imagery. Second, we achieve state-of-the-art detection results on the Aerial Person Detection dataset and the VisDrone2019-Det dataset, demonstrating superior performance compared to existing YOLO-based models. Third, we optimize multi-object tracking in aerial scenarios by introducing a novel method for improving moving object path estimation, addressing a major limitation in current tracking frameworks. Finally, we conduct extensive ablation experiments to evaluate the impact of each model component, providing valuable insights into optimal architectural configurations for aerial object detection. By addressing the fundamental challenges of aerial object tracking, our work contributes to improving real-world applications in surveillance, traffic monitoring, and environmental observation, paving the way for more robust and flexible UAV-based tracking systems.

2 Related Work

To improve small object detection, various enhancements have been proposed based on different versions of the YOLO model. Wang et al. [20] developed an improved YOLOv7-tiny model, specifically designed to enhance small object detection accuracy. Their improvements included the integration of a Global Attention Mechanism (GAM), built upon the CBAM attention module, to strengthen feature extraction in the neck structure. Additionally, they introduced a new small object detection head to prevent excessive downsampling, which often leads to the loss of small object features. By leveraging shallow features that undergo minimal convolutional transformations, they improved the retention of fine-grained details in the detection process. Furthermore, they incorporated the Bidirectional Feature Pyramid Network (BiFPN) into the neck of YOLOv7-tiny, enhancing feature fusion and minimizing information loss. Their approach shares similarities with the PRB-FPN structure [13] incorporated in our model. To further optimize accuracy, they adopted the SIoU loss function instead of traditional IoU-based losses. Experimental evaluations on the VisDrone2019-Det dataset demonstrated that their model outperformed YOLOv3-tiny, YOLOv4-tiny, YOLOv5s, and YOLOv7-tiny in terms of detection accuracy.

Another improvement based on YOLOv7-tiny was proposed by Zhang et al. [21], who introduced a model named PDWT-YOLO. This model employed four detection heads, including an additional small object detection layer, which significantly improved the identification of small targets. To better integrate multi-scale features, they introduced a pyramid structure in the neck, ensuring high detection accuracy across different object sizes without compromising performance on larger objects. They also replaced

the conventional detection head with a decoupled head, separating classification and regression tasks to minimize their mutual interference. For the loss function, they replaced CIoU with WIoU, which resulted in improved convergence speed and detection precision. Their experimental comparisons on the VisDrone2019-Det dataset demonstrated that PDWT-YOLO achieved superior performance in most object categories, except for bicycle and truck detection.

Luo et al. [22] proposed YOLO-UAV, an improved YOLOv5l-based model designed specifically for small object detection in drone imagery. They introduced three novel backbone modules: Asymmetric ResNet (ASResNet), Asymmetric Enhanced Feature Extraction (AEFE), and Asymmetric Res2Net (ASRes2Net). These modules significantly improved feature extraction capabilities for aerial images captured by drones. Additionally, they incorporated the Improved Efficient Channel Attention (IECA) module, enabling the network to emphasize essential features while suppressing irrelevant background information. They also replaced the Spatial Pyramid Pooling (SPP) module with Group Spatial Pyramid Pooling (GSPP), reducing computational complexity while improving detection accuracy. Their model outperformed YOLOv5l on the VisDrone2019-Det dataset, demonstrating notable improvements in small object detection.

Ding et al. [23] introduced modifications to the YOLOv5s architecture to enhance small object detection performance. Their enhancements included the addition of a fourth prediction head, which significantly improved accuracy but increased the number of parameters. To mitigate computational overhead, they employed transformer layers to enhance feature capturing capabilities. They also introduced an additional upsampling operation in the neck, which, despite increasing computational cost, substantially improved accuracy. Furthermore, they incorporated the Efficient Pyramid Squeeze Attention (EPSA) network module in the backbone, enabling richer multi-scale feature representation. Comparisons on the VisDrone2019-Det dataset demonstrated that their modified YOLOv5s model outperformed the original YOLOv5s.

Sun et al. [24] proposed the HPS-YOLOv7 algorithm to enhance small object detection in drone aerial images. Their approach included the C-recursively gated convolution module, designed to integrate shallow object information effectively while improving model capacity. They also replaced the conventional convolution operations in the neck with a lightweight bottleneck module, which preserved small object features while reducing computational cost. Additionally, they introduced a modified high-efficiency layer aggregation network to enhance feature extraction. To improve small object detection further, they replaced the original 20×20 detection head with a 160×160 detection head, allowing finer object localization. Their Shallow Feature Fusion Network (SFFN) reduced the loss of small object features in deep convolution layers. On the VisDrone2019-Det dataset, HPS-YOLOv7 demonstrated superior performance compared to multiple versions of YOLO.

Wang et al. [14] introduced UAV-YOLOv8s, a modified YOLOv8s model designed to enhance small object detection in UAV applications. They incorporated the BiFormer attention mechanism into the backbone, enabling the model to focus on key features more effectively. Additionally, they designed a Focal FasterNet block (FFNB) for feature processing, allowing for efficient fusion of shallow and deep features, which reduced the likelihood of missed small object detections. Similar to our approach, they adopted WIoU v3 as the bounding box regression loss function. WIoU v3 incorporates a dynamic sample allocation strategy, reducing the influence of extreme samples and improving generalizability. Their experimental results on the VisDrone2019-Det dataset showed that UAV-YOLOv8s outperformed other YOLOv8 variants, including YOLOv8n, YOLOv8s, YOLOv8m, and YOLOv8l.

Li et al. [15] proposed YOLOv7-UAV, which aimed to improve small object detection in drone imagery. Their method emphasized the importance of shallow-layer feature maps, which retain rich small-object information. To enhance detection accuracy, they replaced the P5 prediction head with a P2 prediction head, effectively improving small object recognition. However, to prevent excessive model complexity, they

removed the original P5 prediction head. They also modified the Bidirectional Feature Pyramid Network (BiFPN), introducing a BiFPN-like structure, which shares similarities with the PRB-FPN used in our study. This structure enables upsampling and downsampling while retaining cross-level feature fusion connections, preserving rich small-object features. Additionally, they replaced CIoU with SIoU as the loss function, achieving improved accuracy. Their evaluations on the VisDrone2019-Det dataset demonstrated that YOLOv7-UAV outperformed multiple YOLO versions, including YOLOv3, YOLOv4, YOLOv5, and YOLOv7.

Detecting small objects in drone aerial imagery remains particularly challenging due to various interference factors. Zhang et al. [25] proposed an improved YOLOv7-tiny-based algorithm that introduced Receptive Field Coordinate Attention Convolution (RFCAConv) in place of the ELAN-S layer in the backbone. The RFCAConv module improved the model's ability to localize key image regions while enhancing feature representation. Additionally, they introduced an extra detection layer to improve small object detection, utilizing the BSAM attention mechanism to distinguish objects from the background more effectively. To optimize bounding box regression, they replaced CIoU with inner-MPDIoU, which demonstrated higher sensitivity to aspect ratio variations. Similar to our approach, they integrated attention mechanisms and modified the loss function to improve detection performance. Their experimental results on the VisDrone2019-Det dataset confirmed that their improved YOLOv7-tiny model achieved higher mean Average Precision (mAP) @0.5 compared to other models.

In addition to improvements based on YOLO architectures, other state-of-the-art object detection models have also been developed. EfficientDet [26] proposed a scalable and efficient detection architecture by introducing a compound scaling method and a BiFPN for feature fusion. While EfficientDet achieves high accuracy with optimized computational cost, its performance tends to degrade when detecting extremely small objects, especially in aerial imagery scenarios. DETR (Detection Transformer) [27] introduced a transformer-based end-to-end object detection framework, eliminating the need for hand-designed components such as non-maximum suppression. Recently, Shi et al. [28] proposed CAW-YOLO, which leverages cross-layer fusion and weighted receptive fields to enhance small object detection performance in remote sensing imagery, demonstrating the importance of multi-scale feature aggregation in handling small targets. However, DETR often requires large-scale datasets and extensive training time, and its localization accuracy for small objects is less competitive compared to specialized models. In contrast, our AVA-PRB model focuses specifically on enhancing small object detection in aerial moving perspectives, using lightweight attention mechanisms and hierarchical feature refinement to achieve robust and efficient detection without incurring excessive computational overhead. While EfficientDet and DETR represent major advances in object detection frameworks, they are not specifically optimized for the challenges presented by small object detection under aerial moving perspectives. Addressing these challenges remains a critical gap that the AVA-PRB model is designed to fill.

3 Methodology

3.1 Aerial View Attention-PRB (AVA-PRB) Model for Small Object Detection

To enhance small object detection in aerial imagery, we propose the Aerial View Attention-PRB (AVA-PRB) model, which is built upon the Parallel Residual Bi-Fusion Feature Pyramid Network (PRB-FPN) architecture [13]. The PRB-FPN structure effectively integrates multi-scale features through a bidirectional fusion mechanism, making it highly suitable for detecting small objects in complex backgrounds.

As illustrated in Fig. 1, the AVA-PRB model follows a three-stage design, consisting of a backbone for hierarchical feature extraction, a bi-directional feature fusion module, and a detection head. The backbone,

derived from YOLOv7, incorporates ELAN-CA modules, MaxPooling (MP), and CBS blocks to extract features at different levels of abstraction. To enhance spatial and channel representations, Coordinate Attention (CA) modules are integrated into multiple backbone stages, allowing the network to better capture contextual information across different scales.

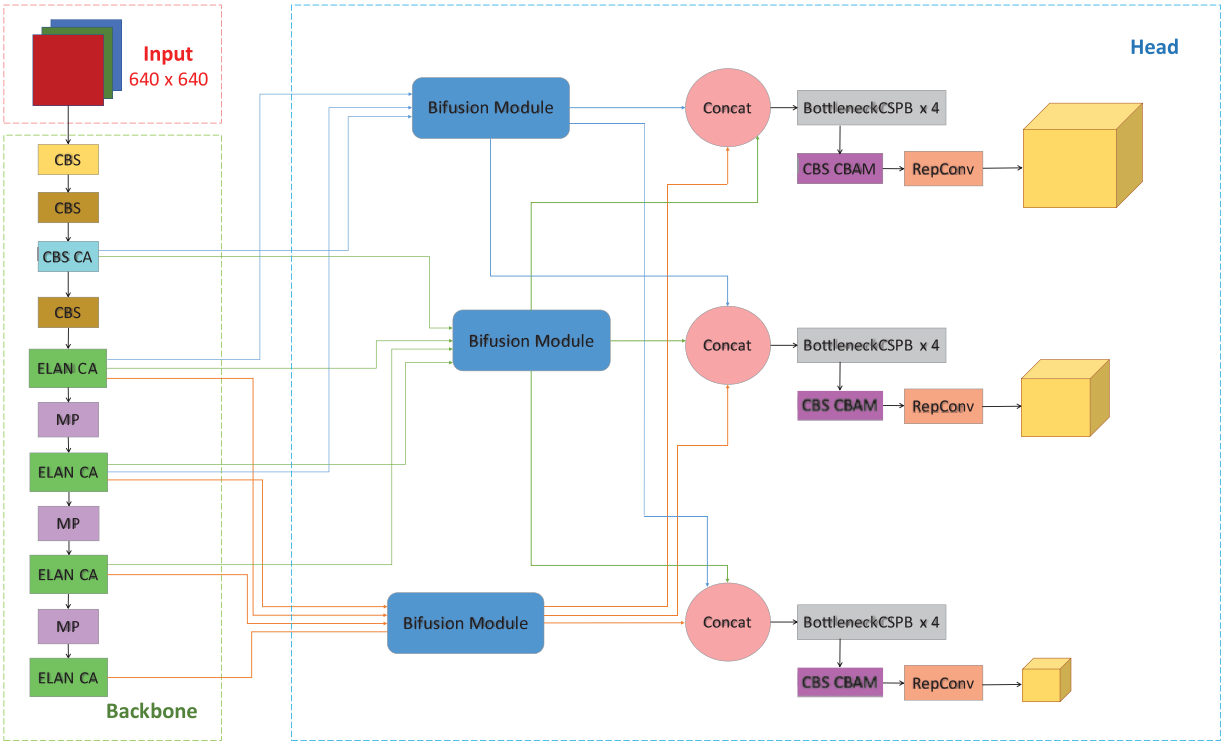


Figure 1: The architecture of the proposed AVA-PRB model. The model integrates a hierarchical feature extraction backbone, bidirectional feature fusion modules, and a detection head enhanced with attention mechanisms

Following feature extraction, the model employs Bi-Fusion Modules to aggregate both low-level spatial details and high-level semantic information. Unlike traditional top-down fusion structures, the bidirectional design enables information to flow in both directions, strengthening feature continuity across scales and preserving fine-grained object characteristics.

The output of the Bi-Fusion process is passed into the Detection Head, where feature refinement is further enhanced. Within the head, CBAM modules are applied before the final convolutional layers to recalibrate both channel and spatial dimensions. This selective attention mechanism enables the network to suppress irrelevant background noise and focus on task-relevant features.

By embedding attention mechanisms throughout the backbone and detection head, AVA-PRB achieves robust performance in aerial views, improving the detection of small and difficult-to-identify objects under varying conditions.

3.2 The Attention Mechanism in the AVA-PRB Network

Attention mechanisms play a vital role in enabling convolutional networks to focus on the most relevant parts of the input features. In the AVA-PRB model, we strategically incorporate two widely adopted attention modules: Coordinate Attention (CA) and the Convolutional Block Attention Module

(CBAM), each contributing at different stages of the architecture. The combination of CA and CBAM is carefully designed to complement each other: CA enhances lightweight spatial attention during early feature extraction, while CBAM focuses on global feature refinement at the later stage, resulting in improved localization and robustness across the detection pipeline.

Since its introduction by Vaswani [29] in 2017, the attention mechanism has been widely utilized across various deep learning applications. CA, introduced by Hou et al. [30] in 2021, recalibrates feature channels by computing the relative importance of each channel while preserving spatial coordinate information. This makes CA particularly useful for tasks that require precise object localization. Additionally, CA has a low computational cost, making it suitable for real-time applications. CBAM, proposed by Woo et al. [31] in 2018, combines channel attention and spatial attention by computing both channel-wise and spatial feature weights. Unlike CA, which focuses primarily on channel recalibration with spatial awareness, CBAM jointly considers spatial and channel dependencies, thereby improving the expressiveness of extracted features. Due to its flexibility, CBAM can be easily integrated into various CNN architectures to enhance feature representation.

As visualized in Fig. 2, CA is integrated into both the backbone and Bi-Fusion Modules. Specifically, CA is applied within the ELAN blocks to guide feature selection during early-stage extraction. By embedding CA into ELAN (green blocks in the figure), the model preserves both spatial context and channel interdependencies, improving the robustness of fine-grained object representations.

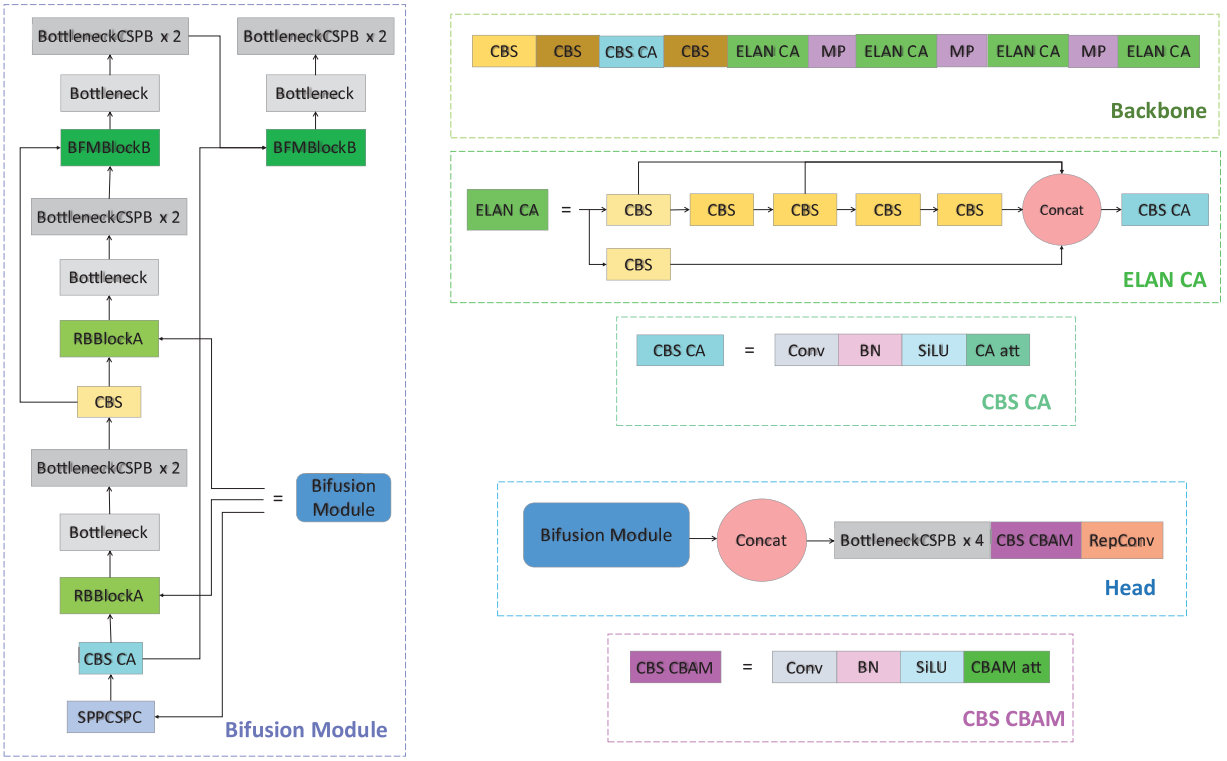


Figure 2: Integration of attention mechanisms in the AVA-PRB model. CA is applied throughout the backbone (green) and Bi-Fusion Modules, while CBAM is introduced in the Detection Head to enhance spatial and channel recalibration prior to prediction

Additionally, CA is also introduced in the CBS blocks located after the SPPCSPC module, which further refines features post-aggregation. These locations are carefully selected to ensure that spatial and channel-wise relationships are reinforced before fusion into the Detection Head.

CBAM, on the other hand, is employed in the final stage of the Detection Head (see purple-pink modules in Fig. 2). It recalibrates both spatial and channel features just before RepConv, allowing the model to suppress redundant activations and concentrate on highly informative regions. By strategically combining these two complementary attention mechanisms—CA for lightweight, spatially aware calibration, and CBAM for global feature refinement—the AVA-PRB model establishes a hierarchical attention framework that improves both early feature localization and final target discrimination, thus enabling accurate small object detection under complex aerial moving perspectives.

3.3 Use of Shape-IoU Loss Function in AVA-PRB Model

The original PRB-FPN model and YOLOv7 utilize the Complete IoU (CIoU) loss function for bounding box regression. However, traditional IoU-based losses often face challenges in small object detection, as even minor shape variations can lead to significant fluctuations in IoU values. To address this issue, our model adopts the Shape-IoU loss function, which was proposed by Zhang and Zhang [18] in 2023.

Compared to CIoU, which primarily considers center point distance, aspect ratio, and overlap area, Shape-IoU emphasizes shape similarity, making it more suitable for detecting small objects. Since small objects often exhibit unique shape characteristics that are more distinguishable than their absolute size, using Shape-IoU allows the model to better capture shape-based details, ultimately enhancing detection accuracy.

In the implementation of Shape-IoU, a scaling factor Scale is introduced, which is dynamically adjusted based on the target's scale within the dataset. This adaptive scaling mechanism helps mitigate the impact of extreme size variations, ensuring that small objects retain their distinctive shape attributes during detection. By integrating Shape-IoU into the AVA-PRB model, we significantly improve the model's capability to distinguish small objects from background clutter and overlapping objects, leading to enhanced detection robustness.

3.4 Implementation for Generating Object Movement Path

For object tracking, we adopt SMILEtrack, a state-of-the-art multi-object tracking (MOT) framework. The installation and implementation details of SMILEtrack can be found on its official GitHub repository [4]. Originally, SMILEtrack utilized PRB-FPN or YOLOv7 as its object detection backbone. However, in our approach, we replace these detection models with the AVA-PRB model, which provides enhanced small object detection and feature extraction capabilities.

To further refine the object tracking process, we modify the tracking method proposed by RizwanMunawar [32], which is based on YOLOv8. This adaptation enables AVA-PRB to integrate seamlessly with the tracking framework, allowing for precise object trajectory estimation in aerial imagery. The combination of AVA-PRB and SMILEtrack ensures more stable tracking results, particularly in dynamic scenes where small objects undergo significant displacement or occlusion.

By leveraging AVA-PRB's enhanced object detection performance and integrating it with SMILEtrack's robust tracking mechanism, we achieve more accurate object movement path generation, making this approach highly suitable for aerial surveillance, traffic monitoring, and environmental tracking applications.

3.5 Generating Object Movement Paths under Aerial Motion

This section details the methodology used to generate object movement paths while compensating for aerial motion distortions. The process consists of four key components: homography matrix estimation, feature point matching, coordinate correction, and movement path adjustment. The complete workflow of the homography matrix update process is illustrated in Fig. 3, while Fig. 4 presents the overall tracking pipeline, demonstrating how AVA-PRB integrates with the tracker and movement path generation.

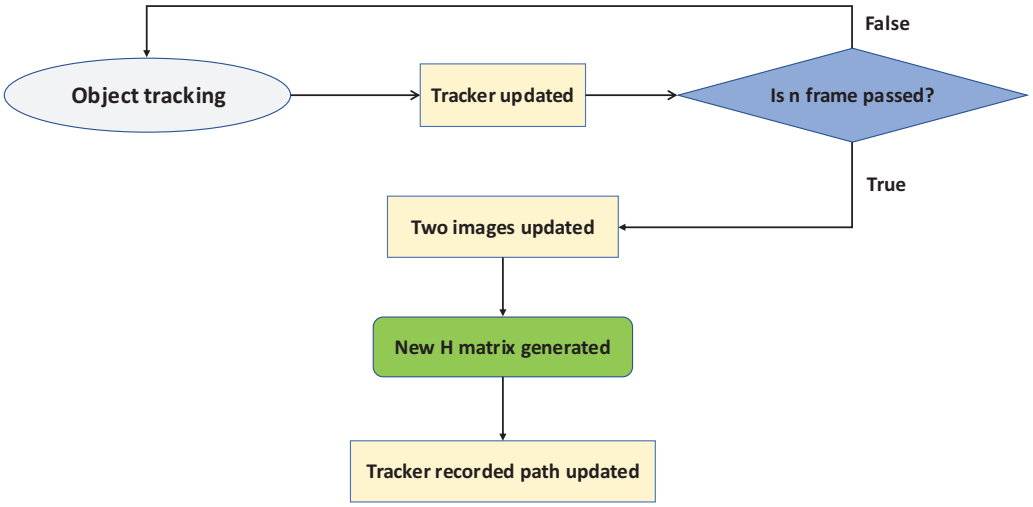


Figure 3: Homography matrix update process in object tracking

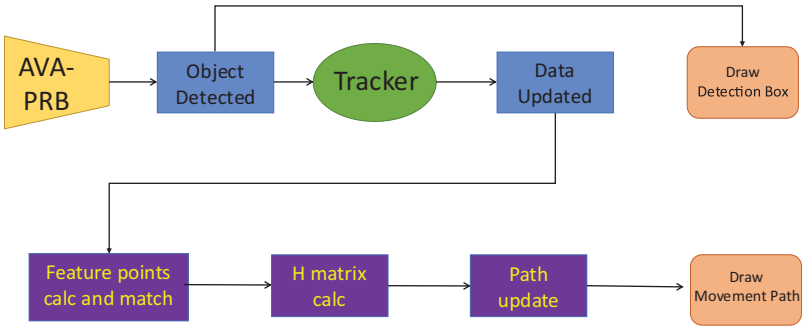


Figure 4: Overall tracking workflow with AVA-PRB and homography-based path correction

3.5.1 Homography Matrix Estimation

The homography matrix is a critical component of this method. In computer vision and image processing, a homography matrix is a 3 × 3 transformation matrix that describes the mapping of points between two different views of the same scene. It is widely applied in image stitching, perspective correction, and object tracking. To correct aerial motion distortions and achieve frame-to-frame trajectory alignment, we employ the homography matrix (Fig. 3).

For implementation, we use OpenCV's built-in function [33] to compute the homography matrix, employing the RANdom SAmple Consensus (RANSAC) algorithm to filter outliers effectively. RANSAC is a widely used technique for estimating geometric transformations in noisy environments. We set the

reprojection threshold to 5 pixels, balancing the trade-off between strict feature matching and tolerance for minor variations. A smaller threshold enforces stricter constraints but may reject valid correspondences, while a larger threshold increases tolerance but may introduce false matches. Additionally, the maximum number of RANSAC iterations is set to 500, ensuring a balance between computational efficiency and transformation accuracy.

Since the homography matrix computation requires two consecutive images, the homography matrix is continuously updated throughout the object tracking process, as illustrated in Fig. 3. A frame interval variable controls the frequency of homography matrix recalculations, ensuring adaptability to different tracking conditions. To improve robustness under challenging conditions such as temporary occlusion or detection errors, RANSAC-based filtering ensures that only reliable feature matches are used for homography estimation. This helps maintain trajectory stability even when feature points are partially missing or noisy. If this interval is set to 5, the homography matrix is computed using keypoints extracted from the first and fifth frames. If set to 1, it is generated based on consecutive frames, ensuring more frequent motion compensation.

3.5.2 Feature Point Extraction and Matching

Before computing the homography matrix, it is necessary to extract and match feature points between two consecutive frames. Feature points, also known as keypoints, are distinctive image locations that remain invariant to scaling, rotation, and lighting changes. These keypoints are commonly used in image matching, object recognition, and registration tasks.

For keypoint extraction and matching, we employ SuperPoint, a deep learning-based keypoint detection and description algorithm introduced by DeTone et al. [34] in 2018. SuperPoint has gained popularity due to its high efficiency and robustness, making it particularly suitable for real-time tracking applications. Using SuperPoint, we extract discriminative keypoints and establish correspondences between consecutive frames, forming the basis for reliable homography estimation. This feature point matching process is a critical step in the homography matrix update pipeline, as depicted in Fig. 3. SuperPoint not only provides high-quality keypoints that are invariant to scale, rotation, and illumination changes, but also generates corresponding descriptors that facilitate robust and efficient matching between frames. This robustness is particularly important in aerial moving scenarios, where objects may appear distorted, blurred, or partially occluded due to rapid drone motion and varying viewpoints. By leveraging SuperPoint's stability, the proposed method ensures reliable feature correspondence for accurate homography estimation.

3.5.3 Coordinate Correction for Movement Path

Once the homography matrix is computed, it is used to transform and update object coordinates over time. Since multiple objects may be present in each frame, each object follows a unique movement trajectory consisting of numerous coordinate points. All object tracking data is stored and updated dynamically, as shown in Fig. 4. Whenever the specified frame interval is reached, a new homography matrix is generated, and all object coordinates are updated accordingly, as illustrated in Fig. 3.

The homography transformation effectively maps 2D points from one image plane to another, preserving their relative spatial positions. For instance, let image_1 represent a frame captured n frames ago, and image_2 represent the current frame. When a homography matrix is computed between image_1 and image_2, any coordinate point in image_2 corresponding to an object (e.g., a pedestrian) is transformed to its relative position in image_1. Similarly, applying the updated homography matrix to previous coordinate points allows their positions to be corrected relative to the current frame, minimizing tracking drift.

By continuously updating recorded movement path coordinates using the latest homography matrix, the proposed method effectively compensates for deviations introduced by aerial motion, ensuring that object trajectories remain accurately aligned with real-world movement. This process is an integral part of the path update step in Fig. 4, where updated tracking information is used to generate precise movement paths. By applying the homography matrix to correct object coordinates, the proposed method significantly reduces trajectory drift that would otherwise accumulate due to continuous camera motion. This continuous correction mechanism ensures that the estimated movement paths better align with the true motion of the objects in the aerial scene, even under dynamic platform movement. As a result, the method enhances the accuracy and reliability of long-term trajectory tracking.

3.5.4 Movement Path Refinement and Visualization

To improve trajectory visualization, we modify the starting position of the movement path. Initially, object movement trajectories extended outward from the center of the detected object. However, after modification, the movement path now extends outward from the object's feet.

This adjustment is particularly beneficial when tracking pedestrians, as it provides a more intuitive and realistic visualization of human motion. By aligning the trajectory origin with the object's actual contact point with the ground, the movement path better represents real-world dynamics, thereby improving the interpretability of tracking results. This visualization refinement is reflected in the movement path drawing step in Fig. 4, which integrates the updated tracking data with trajectory rendering. In addition to improving the visual clarity of the movement paths, this refinement also helps align the estimated trajectories more closely with the physical ground plane. Such alignment is particularly beneficial for analyzing pedestrian movement patterns or vehicle trajectories in aerial surveillance scenarios, where precise ground-level localization enhances interpretability and facilitates downstream analysis tasks.

4 Experiments

4.1 Dataset for Small Object Detection Experiment

The Aerial Person Detection (APD) dataset [35] consists of 6445 training images and 545 validation images. The training images have resolutions ranging from 960 × 540 to 2000 × 1500 pixels, while the validation images range from 960 × 540 to 1920 × 1080 pixels. The dataset is specifically designed for person detection in aerial imagery, making it highly suitable for computer vision models that require the ability to identify and localize individuals from an aerial viewpoint.

The dataset serves as a valuable resource for applications such as surveillance, search and rescue missions, and aerial-based human detection tasks. It includes images captured at various altitudes, viewing angles, and environmental conditions, ensuring a diverse and comprehensive training set that enhances model robustness. The dataset contains images taken under different lighting conditions, ranging from full daylight to nighttime scenarios (Fig. 5), further challenging the model's adaptability to varying illumination.

As shown in Fig. 5, the dataset includes both pedestrian and vehicle detection scenarios, covering daytime and nighttime conditions. Fig. 5a,b demonstrate aerial pedestrian detection under daylight and nighttime lighting, respectively. Similarly, Fig. 5c,d illustrate vehicle detection in nighttime and daytime environments. These variations ensure that models trained on this dataset can generalize well to real-world aerial detection tasks, where lighting and environmental conditions vary significantly.

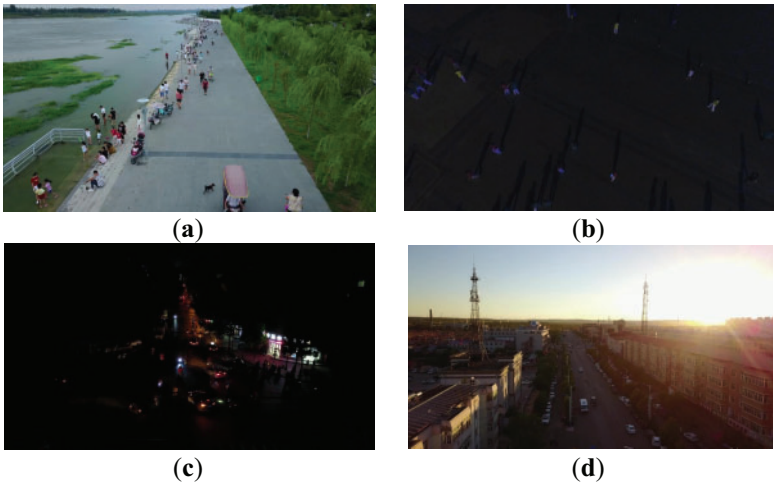


Figure 5: Representative images from the Aerial Person Detection (APD) dataset [35], including a daytime aerial image of pedestrians (**a**), a nighttime aerial image of pedestrians (**b**), a nighttime aerial image of vehicles (**c**), and a daytime aerial image of vehicles (**d**)

Two datasets are utilized for small object detection experiments: the Aerial Person Detection (APD) dataset [35] and the VisDrone2019-Det dataset [36]. The APD dataset consists of 6445 training images and 545 validation images. The training images have resolutions ranging from 960 × 540 to 2000 × 1500 pixels, while the validation images range from 960 × 540 to 1920 × 1080 pixels. The dataset is specifically designed for person detection in aerial imagery, making it highly suitable for computer vision models that require the ability to identify and localize individuals from an aerial viewpoint.

This dataset is particularly valuable for applications in surveillance, search and rescue operations, and other aerial-based human detection tasks. It includes images captured at various altitudes, viewing angles, and environmental conditions, ensuring a diverse and comprehensive training set that enhances model robustness. The dataset contains images taken under different lighting conditions, ranging from full daylight to nighttime scenarios, further challenging the model's adaptability to varying illumination.

The APD dataset targets six object categories, including bicycle, bus, car, motorcycle, person, and truck. All images are captured from an aerial perspective, making this dataset an ideal benchmark for evaluating small object detection performance in drone-based applications. The class distribution of labeled objects in the dataset is summarized in Table 1.

Table 1: Class distribution in the aerial person detection (APD) dataset

Class	Train labels	Validation labels
Bicycle	10,340	1283
Bus	5907	251
Car	143,585	13,986
Motorcycle	29,533	4861
Person	105,569	13,890
Truck	37,589	2717
Total	332,523	36,988

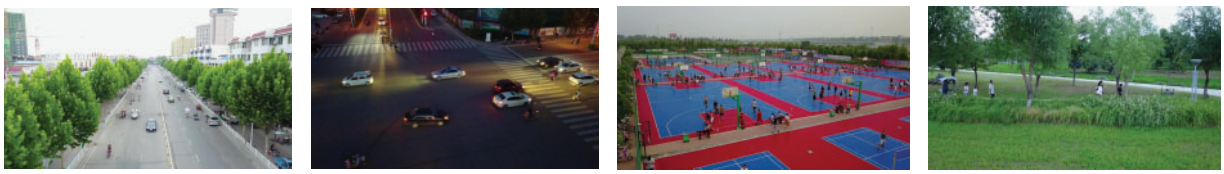
The VisDrone2019-Det dataset consists of 6471 training images and 548 validation images. The training images have resolutions ranging from 960 × 540 to 2000 × 1500 pixels, while the validation images range from 960 × 540 to 1920 × 1080 pixels. Designed for drone vision research, VisDrone2019-Det is one of the largest aerial object detection datasets, providing a diverse set of complex real-world scenes for training and evaluating deep learning models.

This dataset contains multiple object categories labeled in each image, covering a variety of challenging scenarios such as dense object distributions, varying object scales, occlusions, and diverse lighting conditions. These challenges significantly contribute to the robustness of object detection models, making VisDrone2019-Det a widely used benchmark for small object detection. Some images feature highly crowded urban scenes, small targets, and severe occlusion, requiring advanced detection models capable of precise localization under difficult conditions. The dataset targets ten object categories, including pedestrians, vehicles (cars, vans, trucks, buses), motorcycles, bicycles, and tricycles, as summarized in Table 2.

Table 2: Class distribution in the VisDrone2019-Det dataset

Class	Train labels	Validation labels
Pedestrian	79,055	8844
People	26,962	5125
Bicycle	10,389	1287
Car	144,619	14,064
Van	24,899	1975
Truck	12,875	750
Tricycle	4812	1045
Awning-Tricycle	3245	532
Bus	5917	251
Motor	29,618	4886
Total	342,391	38,759

As illustrated in Fig. 6, the dataset encompasses various environmental conditions and scene types, ensuring model generalization across daytime and nighttime scenarios. Fig. 6a,b depicts roadway scenes during the day and at night, highlighting the impact of lighting variations on object visibility. Additionally, Fig. 6c,d showcases non-urban environments, such as a basketball court and a rural area, demonstrating the dataset's diversity in aerial perspectives. These scene variations make VisDrone2019-Det highly suitable for evaluating object detection models in real-world aerial applications.



(**a**) (**b**) (**c**) (**d**)

Figure 6: Representative images from the VisDrone2019-Det dataset, including a daytime aerial image of a roadway (**a**), a nighttime aerial image of a roadway (**b**), an aerial image of a basketball court (**c**), and an aerial image of a rural area (**d**)

4.2 Experimental Setup and Model Training Configuration

The experiments were conducted on a personal computer equipped with Windows 11 Enterprise as the operating system. The system was powered by an AMD Ryzen 5 7600X six-core processor and supported by 32 GB of DDR5 RAM (4800 MHz). A PNY CS2241 1 TB SSD was used for storage, ensuring efficient data handling, and an NVIDIA GeForce RTX 4070 GPU was utilized to accelerate deep learning computations.

The software environment consisted of PyCharm (Version 0.00) and Anaconda (Version 0.00) as the primary development platforms. The experiments were implemented using Python 3.8, with PyTorch 1.11.0 as the deep learning framework, accompanied by Torchvision 0.12.0 and Torchaudio 0.11.0 for image and audio-related tasks. The model training leveraged CUDA Toolkit 11.3 to enable GPU-accelerated computation.

The training settings were configured to ensure optimal performance under the given hardware constraints. Each input image was resized to 640 × 640 pixels, and the model was trained for 300 epochs. Due to hardware limitations, the batch size and number of workers were both set to 4 to balance computational efficiency and memory usage. The learning rate was initialized at 0.01, with a momentum parameter of 0.937 and a weight decay of 0.0005 to regulate model convergence and prevent overfitting.

The proposed AVA-PRB model follows a P5-based architecture, and all other hyperparameter configurations used for training were consistent with the standard P5 model settings.

4.3 Model Evaluation Metrics for Small Object Detection

To evaluate the performance of the proposed model in small object detection, several commonly used metrics are employed, including Precision (P), Recall (R), F1 Score, and mean Average Precision (mAP). These metrics rely on fundamental classification concepts such as True Positive (TP), False Positive (FP), and False Negative (FN), which define the correctness of the model's predictions.

A True Positive (TP) occurs when the model correctly predicts an object's presence, meaning that the detected object matches the ground truth label. In contrast, a False Positive (FP) refers to an incorrect detection where the model predicts an object that does not exist in the ground truth, leading to false alarms. On the other hand, a False Negative (FN) represents a missed detection, where the model fails to recognize an actual object, reducing overall recall performance. These concepts form the basis for evaluating detection accuracy and reliability.

The evaluation metrics are mathematically defined as follows:

$$P = \frac{TP}{TP + FP} \tag{1}$$

$$R = \frac{TP}{TP + FN} \tag{2}$$

$$F1\,Score = \frac{2 \times P \times R}{P + R} \tag{3}$$

$$AP = \sum_{k} \left(R_k - R_{k-1}\right) P_k \tag{4}$$

$$mAP = \frac{1}{N} \sum_{i=1}^{N} AP_i \tag{5}$$

where N represents the total number of object classes in the dataset.

Precision (Eq. (1)) quantifies the proportion of correctly identified objects among all predicted objects. A higher precision value indicates fewer false positives, meaning that the model has a higher level of confidence in its predictions. Recall (Eq. (2)) measures the proportion of actual objects that the model successfully detects, where a higher recall score signifies that the model misses fewer true positive instances.

F1 Score (Eq. (3)) provides a balanced evaluation by considering both precision and recall, making it particularly useful when an optimal trade-off between the two is required. Average Precision (AP) (Eq. (4)) is computed as the area under the Precision-Recall (PR) curve, using a discrete summation over recall levels to approximate the integral. It measures how well the model maintains high precision across varying recall thresholds, accounting for the trade-off between precision and recall at different confidence levels where R_k and R_{k-1} are consecutive recall levels, and P_k is the precision at recall level R_k. Unlike single-point precision or recall measurements, AP evaluates model performance holistically across multiple confidence thresholds, making it a more comprehensive and reliable metric for object detection. Mean Average Precision (mAP) (Eq. (5)) is a widely used metric in object detection tasks, representing the mean AP over all object categories, thereby providing an overall performance measure.

The evaluation metric mAP@0.5 refers to the mAP computed at an Intersection over Union (IoU) threshold of 0.5, meaning that a detection is considered correct if its IoU with the ground truth is at least 0.5. Conversely, mAP@0.5:0.95 represents the mean AP across multiple IoU thresholds ranging from 0.5 to 0.95 in increments of 0.05, making it a more stringent and comprehensive evaluation criterion.

4.4 Experimental Results and Performance Analysis for Small Object Detection

To optimize the proposed model, an ablation study was conducted to evaluate the impact of applying attention mechanisms at different positions within the network. The study examined how modifications to specific locations affected detection performance, ultimately identifying the most effective configuration. To facilitate analysis, three designated positions were considered: L1, referring to the backbone; L2, representing the CBS module located after SPPCSPC in the BiFusion module; and L3, corresponding to the CBS module positioned between BottleneckCSPB and RepConv in the model head. The experiments were conducted using the APD dataset, and the results are presented in Table 3.

Table 3: Ablation study results on the APD dataset

Method	Precision (%)	Recall (%)	mAP@0.5 (%)	mAP@0.5:0.95 (%)	F1 score (%)
PRB-FPN-ELAN [13]	68.6	58.0	60.3	35.7	**62.8**
L1 + CA	70.7	56.6	60.3	35.7	**62.8**
L2 + CA	69.6	57.0	59.9	35.6	62.6
L3 + CA	68.1	58.2	60.6	35.7	62.7
L3 + CBAM	67.7	58.0	60.3	35.4	62.4
L1 & L3 + CA	70.5	56.7	60.0	35.5	**62.8**
L1 + CA & L3 + CBAM	69.9	56.4	60.4	35.6	62.4
L1 & L2 & L3 + CA	68.5	57.3	60.0	35.4	62.4
L1 & L2 + CA & L3 + CBAM	67.9	**58.5**	**60.7**	**35.8**	**62.8**

Note: Bold values indicate the best performance among all methods in the respective column.

The results indicate that adding CBAM significantly increases computational complexity and the number of parameters, particularly when applied at L1 (backbone). Consequently, the computational cost of integrating CA is relatively lower. Among all single-position modifications, applying CA to L3 yielded the most noticeable improvement, with an increase of 0.3% in mAP@0.5 compared to the baseline PRB-FPN-ELAN. However, modifications at L1 or L2 alone either failed to improve performance or resulted in minor declines. The most effective model configuration was achieved when CA was applied to L1 and L2, while CBAM was incorporated at L3, leading to the highest performance gains. This configuration improved mAP@0.5 by 0.4% and mAP@0.5:0.95 by 0.1%, demonstrating its effectiveness in enhancing detection accuracy.

Apart from the attention mechanism modifications, the impact of the Shape-IoU scaling factor (Scale) on performance was also investigated. As shown in Table 4, the choice of Scale significantly influenced detection accuracy, with even a 0.1 variation leading to noticeable differences. The worst performance was observed at Scale = 1.0, while increasing it to 1.3 resulted in the best performance. Compared to the commonly used CIoU, the optimized Shape-IoU implementation improved mAP@0.5 by 0.2% and mAP@0.5:0.95 by 0.2%, while also enhancing F1 Score by 0.6%. These results highlight the importance of selecting an appropriate Shape-IoU scaling factor to optimize detection accuracy.

Table 4: Impact of Shape-IoU scaling on AVA-PRB performance

Scale	Precision (%)	Recall (%)	mAP@0.5 (%)	mAP@0.5:0.95 (%)	F1 score (%)
1.0	69.4	56.9	59.9	35.5	62.5
1.1	68.7	57.8	60.1	35.5	62.7
1.2	**70.5**	56.5	60.2	35.4	62.7
1.3	68.8	**58.8**	**60.9**	**36.0**	**63.4**
1.4	69.5	57.0	60.2	35.5	62.6

Note: Bold values indicate the best performance among all methods in the respective column.

Several advantages over conventional approaches are demonstrated by the AVA-PRB model introduced herein. The integration of CA and CBAM at optimal positions enhances feature extraction, enabling better small object detection with reduced computational overhead. Furthermore, the use of Shape-IoU instead of CIoU improves bounding box regression, resulting in more precise localization and higher overall detection accuracy. The experimental results confirm that applying CA at L1 & L2 combined with CBAM at L3 leads to the most significant performance improvements, demonstrating the effectiveness of our method in improving small object detection accuracy while maintaining computational efficiency.

4.5 Comparative Analysis of AVA-PRB and YOLO Models for Small Object Detection

To evaluate the effectiveness of the AVA-PRB model, we conducted a comparative study against several state-of-the-art YOLO models that have gained widespread adoption in recent years. These models include YOLOv7-tiny [6], YOLOv7 [6], YOLOv8s [7], YOLOv8m [7], YOLOv9s [6], YOLOv9m [6], YOLOv10s [9], and YOLOv10m [9]. Each model was trained using default hyperparameter settings, with a maximum of 300 epochs and a batch size of 4. Additionally, pretrained weights were used for all YOLO models to ensure a fair comparison. Both the training and validation were conducted on the APD dataset, and the overall performance results are summarized in Table 5.

Table 5: Comparison results of AVA-PRB and YOLO models (APD dataset)

Model	Precision (%)	Recall (%)	mAP@0.5 (%)	mAP@0.5:0.95 (%)	F1 score (%)
YOLOv7-tiny [6]	55.4	45.9	45.6	24.0	50.2
YOLOv7 [6]	71.2	55.8	59.7	35.0	62.5
YOLOv8s [7]	61.8	46.4	50.5	30.8	53.0
YOLOv8m [7]	62.8	50.5	53.6	33.2	55.9
YOLOv9s [6]	61.2	45.0	49.1	29.9	51.8
YOLOv9m [6]	64.2	49.6	53.9	33.4	55.9
YOLOv10s [9]	57.6	44.1	47.1	28.7	49.9
YOLOv10m [9]	60.6	47.0	50.5	31.3	52.9
AVA-PRB (Ours)	68.8	**58.8**	**60.9**	**36.0**	**63.4**

Note: Bold values indicate the best performance among all methods in the respective column.

The results indicate that AVA-PRB outperforms all YOLO models across multiple evaluation metrics. While YOLOv7 achieves the highest Precision (71.2%), our model follows closely with a Precision of 68.8%, with only a 2.4% difference. However, AVA-PRB demonstrates superior performance in Recall, mAP@0.5, mAP@0.5:0.95, and F1 Score, outperforming every YOLO model tested. Notably, our model achieves the highest Recall at 58.8%, signifying that it misses fewer objects during detection, which is particularly crucial for small object detection tasks. This balance between Precision and Recall results in the highest F1 Score of 63.4%, confirming its overall robustness in detection accuracy.

In terms of mAP@0.5, AVA-PRB attains 60.9%, exceeding YOLOv7's 59.7%, while also outperforming all other YOLO models by a significant margin. The mAP@0.5:0.95 score of 36.0% further demonstrates that our model performs well under varying IoU thresholds, suggesting strong generalization across different object sizes and detection scenarios. These results highlight the effectiveness of the integration of attention mechanisms and the Shape-IoU loss function, which contribute to improved localization accuracy and better small object detection performance.

To further analyze the detection capabilities of AVA-PRB, we compare the mAP@0.5 scores for individual object classes across different models. The results, presented in Table 6, illustrate how AVA-PRB performs relative to other YOLO models in detecting various object categories.

Table 6: The mAP@0.5 (%) table for each class (APD dataset)

Model	Bicycle	Bus	Car	Motorcycle	Person	Truck
YOLOv7-tiny [6]	11.0	47.4	77.8	46.1	48.0	43.1
YOLOv7 [6]	**27.0**	66.3	85.4	60.4	64.0	55.4
YOLOv8s [7]	15.9	61.6	80.3	46.5	49.6	49.2
YOLOv8m [7]	19.5	65.1	82.1	49.5	53.8	51.3
YOLOv9s [6]	14.6	58.0	79.8	45.5	47.0	49.8
YOLOv9m [6]	18.7	64.2	82.5	51.0	52.2	54.7

(Continued)

Table 6 (continued)

Model	Bicycle	Bus	Car	Motorcycle	Person	Truck
YOLOv10s [9]	12.0	55.0	79.1	42.9	44.8	48.6
YOLOv10m [9]	15.9	60.8	80.8	46.7	48.5	50.3
AVA-PRB (Ours)	**27.0**	**68.6**	**85.8**	**61.7**	**65.5**	**57.1**

Note: Bold values indicate the best performance among all methods in the respective column.

Examining the per-class performance in Table 6, we observe that AVA-PRB achieves the highest mAP@0.5 scores across all object categories, demonstrating its strong adaptability across multiple object types. For Bicycle detection, AVA-PRB achieves an mAP@0.5 of 27.0%, tying with YOLOv7, while in Bus, Car, Motorcycle, Person, and Truck detection, it consistently surpasses all YOLO models, achieving 68.6%, 85.8%, 61.7%, 65.5%, and 57.1%, respectively. The superior detection performance in Motorcycle (61.7%) and Truck (57.1%) detection suggests that AVA-PRB is particularly effective at handling multi-scale object features, a crucial aspect of aerial image analysis. Additionally, the high mAP@0.5 of 65.5% for Person detection demonstrates that our model excels at recognizing complex human shapes in aerial views, making it a strong candidate for surveillance, disaster response, and urban monitoring applications.

The experimental findings confirm that AVA-PRB consistently outperforms YOLO models in small object detection, offering several key advantages. First, AVA-PRB achieves a higher Recall while maintaining balanced Precision, leading to more consistent detections across different object scales. Second, the integration of the Shape-IoU loss function enhances bounding box localization, resulting in improved mAP@0.5 and mAP@0.5:0.95 scores. Third, the use of CA and CBAM attention mechanisms improves feature extraction, allowing the model to better capture small object details while reducing background noise. Finally, AVA-PRB demonstrates superior adaptability in diverse and complex aerial imagery scenarios, outperforming YOLO models in detecting smaller, occluded, or overlapping objects. These advantages make AVA-PRB an effective solution for small object detection in aerial imagery.

The detection performance of different models is visualized in Fig. 7, where each subfigure presents the results from a specific model applied to an aerial view of a crowded square. The detection results of our proposed AVA-PRB model (Fig. 7a) show the best performance, identifying most of the individuals present in the image, regardless of their distance from the camera or location within the scene. Notably, AVA-PRB is the only model capable of detecting certain individuals that remain undetected by all other models, such as a person wearing black clothing on the right side, a person in white clothing on the far right, and an individual in white sitting on the ground with their back facing the camera on the left. This highlights the robust feature extraction capabilities of AVA-PRB, making it highly suitable for aerial surveillance applications requiring precise small-object detection.

The PRB-FPN-ELAN model (Fig. 7b) demonstrates competitive performance, detecting most of the small objects that AVA-PRB identifies. A group of people positioned farther away in the center of the image is successfully detected by PRB-FPN-ELAN, whereas other YOLO models fail to recognize them. However, PRB-FPN-ELAN performs slightly worse than AVA-PRB, missing some individuals in the left, middle, and right sections of the image, particularly those near the drone's position.

The YOLOv7 model (Fig. 7c) exhibits performance close to that of PRB-FPN-ELAN, successfully detecting many small objects, including distant individuals in the center of the square. However, YOLOv7 fails to detect several people near the camera and misses some detections in various parts of the image. Compared

to PRB-FPN-ELAN, YOLOv7 struggles slightly more with occlusions and densely packed individuals. The YOLOv8m model (Fig. 7d) performs moderately well, detecting more small objects than YOLOv9m. In particular, YOLOv8m successfully detects several individuals on both the left and right sides of the image that YOLOv9m fails to recognize. However, its performance still lags behind YOLOv7 and PRB-FPN-ELAN, especially in detecting distant individuals in the central region.

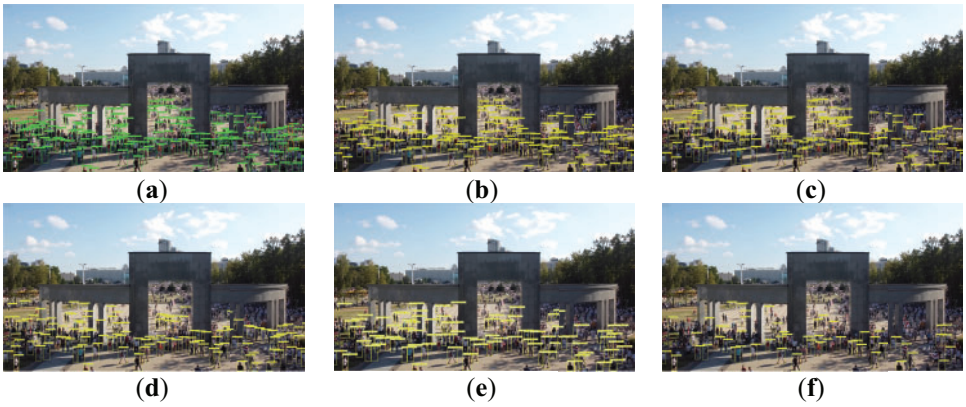


Figure 7: Comparison of small object detection performance across different models in an aerial view of a crowded square. (**a**) AVA-PRB (Ours), (**b**) PRB-FPN-ELAN, (**c**) YOLOv7, (**d**) YOLOv8m, (**e**) YOLOv9m, (**f**) YOLOv10m

The YOLOv9m model (Fig. 7e) shows only minor improvements over YOLOv10m in small-object detection. It performs better than YOLOv10m in detecting individuals at the left and right edges of the image, but it still misses several people that YOLOv8m successfully detects. This suggests that YOLOv9m has difficulty handling small-scale objects and complex occlusions in aerial images. Finally, the YOLOv10m model (Fig. 7f) demonstrates the weakest detection performance, identifying the smallest number of people in the scene. It fails to detect many small objects, particularly distant individuals and those located at the edges of the image, making it the least effective model for small-object detection in aerial views.

4.6 Comparative Analysis of AVA-PRB against Other Enhanced YOLO Variants for Small Object Detection

To further validate the performance of AVA-PRB, we compare it with not only standard YOLO models but also several enhanced YOLO variants that incorporate various improvements for small object detection. These competing models include Improved YOLOv7-tiny [20], PDWT-YOLO [21], YOLO-UAV [22], YOLOv5s-EPSA + upsampling [23], HPS-YOLOv7 [24], UAV-YOLOv8s [14], YOLOv7-UAV [15] and Improved YOLOv7-tiny [25]. These models employ various techniques such as multi-scale feature fusion, attention mechanisms, optimized loss functions, and additional detection heads to enhance their capability in detecting small and distant objects in aerial imagery. For a fair comparison, all models are evaluated on the Visdrone2019-Det dataset using consistent training settings. Specifically, the training epoch limit is set to 300, while the batch size and worker count are adjusted to 2 due to hardware limitations. Performance data for the enhanced YOLO models are sourced from their respective published papers to ensure consistency. Performance metrics are presented separately to provide a clearer and more comprehensive comparison. Table 7 summarizes the Precision, Recall, and F1 Score results, while Table 8 presents the mAP@0.5 and mAP@0.5:0.95 comparisons.

Table 7: Precision, recall, and F1 score comparison of AVA-PRB and other enhanced YOLO models on the Visdrone2019-Det dataset

Model	Precision (%)	Recall (%)	F1 score (%)
Improved YOLOv7-tiny [20]	45.5	42.2	43.7
UAV-YOLOv8s [14]	54.4	45.6	49.6
AVA-PRB (Ours)	**59.8**	**51.7**	**55.4**

Note: Bold values indicate the best performance among all methods in the respective column.

Table 8: mAP performance comparison of AVA-PRB and other enhanced YOLO models on the Visdrone2019-Det dataset

Model	mAP@0.5 (%)	mAP@0.5:0.95 (%)
Improved YOLOv7-tiny [20]	38.1	21.3
PDWT-YOLO [21]	41.2	22.5
YOLO-UAV [22]	30.5	–
YOLOv5s-EPSA + upsampling [23]	42.9	24.6
HPS-YOLOv7 [24]	48.0	27.0
UAV-YOLOv8s [14]	47.0	29.2
YOLOv7-UAV [15]	45.3	–
Improved YOLOv7-tiny [25]	35	–
AVA-PRB (Ours)	**51.2**	**29.9**

Note: Bold values indicate the best performance among all methods in the respective column.

Table 7 highlights the comparison of models that report precision, recall, and F1 Score metrics. It can be observed that the AVA-PRB model achieves the highest performance across all these evaluation metrics. Specifically, AVA-PRB attains a Precision of 59.8%, Recall of 51.7%, and an F1 Score of 55.4%, outperforming both Improved YOLOv7-tiny and UAV-YOLOv8s. This indicates that AVA-PRB maintains a balanced trade-off between detection accuracy and false positive/false negative rates, demonstrating its robustness for aerial small object detection.

Table 8 presents the broader mAP-based comparison among the competing models. AVA-PRB again demonstrates superior performance, achieving a mAP@0.5 of 51.2% and a mAP@0.5:0.95 of 29.9%. Compared to the second-best performing model, UAV-YOLOv8s, AVA-PRB achieves a 4.2% higher mAP@0.5 and a 0.7% higher mAP@0.5:0.95. Other models such as HPS-YOLOv7 and PDWT-YOLO show moderate improvements over standard YOLO architectures but still fall short of the performance attained by AVA-PRB.

In addition to the overall detection performance, detailed per-class mAP@0.5 evaluations are provided in Tables 9 and 10, comparing AVA-PRB against other improved YOLO models across different object categories. Table 9 focuses on five object classes: Pedestrian, People, Bicycle, Car, and Van, while Table 10 presents results for Truck, Tricycle, Awning-Tricycle, Bus, and Motor. The per-class results show that AVA-PRB consistently outperforms other models across most categories, demonstrating strong effectiveness in detecting both small and large objects in complex aerial scenes.

Table 9: Per-class mAP@0.5 (%) comparison of AVA-PRB and other enhanced YOLO models (pedestrian, people, bicycle, car, van) on the Visdrone2019-Det dataset

Model	Pedestrian	People	Bicycle	Car	Van
PDWT-YOLO [21]	48.7	41.6	14.7	82.0	43.2
YOLO-UAV [22]	24.1	11.2	10.0	58.9	36.5
HPS-YOLOv7 [24]	55.5	48.4	21.1	85.0	50.5
UAV-YOLOv8s [14]	56.8	44.9	18.8	**85.8**	50.8
Improved YOLOv7-tiny [25]	45.3	41.7	13.1	83.8	45.3
AVA-PRB (Ours)	**60.5**	**51.2**	**26.6**	85.6	**51.2**

Note: Bold values indicate the best performance among all methods in the respective column.

Table 10: Per-class mAP@0.5 (%) comparison of AVA-PRB and other enhanced YOLO models (truck, tricycle, Awning-tricycle, bus, motor) on the Visdrone2019-Det dataset

Model	Truck	Tricycle	Awning-Tricycle	Bus	Motor
PDWT-YOLO [21]	35.4	26.8	14.2	56.4	49.3
YOLO-UAV [22]	**50.6**	23.1	11.9	52.2	26.0
HPS-YOLOv7 [24]	41.3	35.0	20.0	65.3	57.8
UAV-YOLOv8s [14]	39.0	33.3	19.7	64.3	56.2
Improved YOLOv7-tiny [25]	38.5	25.6	20.7	53.3	47.7
AVA-PRB (Ours)	48.2	**39.0**	**20.8**	**66.9**	**62.0**

Note: Bold values indicate the best performance among all methods in the respective column.

Despite the focus on mAP metrics, Table 7 clearly illustrates that AVA-PRB attains balanced precision and recall, resulting in a solid F1 score. Furthermore, AVA-PRB maintains a lightweight design based on the YOLOv8n backbone, integrating selective attention mechanisms without significantly increasing computational complexity. Consequently, the model sustains an efficient inference time, making it suitable for near real-time aerial surveillance applications. Future work will include more systematic benchmarking of precision, recall, F1 score, and inference time across diverse conditions to further validate the model's robustness.

For Pedestrian detection, AVA-PRB achieves 60.5% mAP@0.5, outperforming the second-best model by 3.7%, indicating its ability to accurately detect people in complex environments. For Bicycle detection, AVA-PRB leads by 5.5%, showcasing its superior performance in detecting small and low-contrast objects. Although AVA-PRB is slightly behind UAV-YOLOv8s in Car detection by 0.2%, it maintains a competitive performance and leads in Van detection by 0.4%.

The per-class performance analysis in Table 10 further reinforces the advantages of AVA-PRB. The model outperforms all competing architectures in four out of five categories, with particularly strong improvements in Motor detection (+4.2%) and Tricycle detection (+4.0%). The only category where AVA-PRB ranks second is Truck detection, where HPS-YOLOv7 outperforms it by 2.4%. However, its high detection accuracy across multiple categories suggests that AVA-PRB is highly adaptable to aerial environments with varying object sizes and densities.

The superior results of AVA-PRB can be attributed to three key architectural improvements. First, the integration of Coordinate Attention (CA) in the backbone and CBAM in the detection head enhances feature

selection and spatial awareness, allowing the model to distinguish small objects from cluttered backgrounds more effectively. Second, the adoption of Shape-IoU loss provides a more precise bounding box refinement strategy, improving localization accuracy for small and elongated objects. Lastly, the Parallel Residual Bi-Fusion Feature Pyramid Network (PRB-FPN) strengthens multi-scale feature fusion, enabling AVA-PRB to retain fine-grained details while suppressing irrelevant background noise. These enhancements allow AVA-PRB to achieve state-of-the-art performance in small object detection, making it a highly effective solution for aerial surveillance, traffic monitoring, and other real-world applications that demand accurate and reliable detection in dynamic aerial environments. The results confirm that AVA-PRB not only outperforms standard YOLO-based architectures but also surpasses other enhanced YOLO models, establishing itself as a robust and scalable approach for small object detection in aerial imagery.

4.7 Aerial Moving Perspective Video Dataset and Movement Path Estimation Efficiency

The dataset used for movement path analysis consists of aerial videos sourced from publicly available platforms. All videos were standardized to 1920 × 1080 resolution at 30 Frames Per Second (FPS) and trimmed to approximately 3 to 4 s in duration, resulting in a total of 83 videos. The scenes primarily depict urban environments containing both stationary and moving targets, with the drone maintaining a single directional motion and a consistent aerial viewpoint without zooming.

Assessing movement path accuracy under an aerial moving perspective presents unique challenges due to the camera's motion-induced perspective changes. When the drone is stationary, the movement path of the target remains accurate; however, when the drone moves, offset factors distort the path estimation. To compensate for these distortions, our method calculates the movement path length based on the Euclidean distance between the starting and ending points of each object's trajectory:

$$d = \sqrt{(\chi_2 - x_1)^2 + (y_2 - y_1)^2} \tag{6}$$

Shorter corrected paths indicate better trajectory estimation accuracy.

In the proposed framework, homography matrices are used to correct perspective distortions during tracking. For efficiency, the interval between two frames for homography matrix computation is set to 10 frames. Table 11 summarizes the performance statistics, including the total number of frames processed, the number of homography matrices computed, and the associated computation times.

Table 11: Summary of Frame count, computation time, and homography matrix analysis

Total frames processed	Total homography matrices computed	Total computation time for homography matrices (s)	Total path update time using homography (ms)	Total execution time (s)
7829	754	7358	5826	8946

The experimental results reveal that computing the homography matrix is the most time-consuming component, accounting for approximately 82% of the total runtime, with an average computation time of 9.7 s per matrix. In contrast, the time required to update the movement path using the homography matrix is relatively minor at only 5.8 s in total.

Although the homography estimation process imposes computational overhead, the AVA-PRB model, built upon the lightweight YOLOv8 backbone, maintains a good balance between detection accuracy and

computational efficiency. This allows the detection and tracking components to sustain efficient operation even in aerial scenarios, supporting near real-time applications. To further improve runtime performance, future work will explore acceleration techniques, such as optimizing homography matrix computation and applying model compression strategies like pruning and quantization.

In addition, Table 12 compares the movement path lengths before and after applying our correction method. The results demonstrate a substantial improvement, with the corrected paths being significantly shorter.

Table 12: Comparison of movement path lengths with and without the proposed method

Movement path length (Pixels)		Difference (Pixels)
Without proposed method	**With proposed method**	
343,764	122,801	220,963

As illustrated in Fig. 8, when the drone moves, the uncorrected movement paths (red lines) continue to extend incorrectly due to perspective shifts, while the corrected paths (green lines) accurately reflect the true motion of the targets. In the case of stationary individuals, the corrected paths remain localized beneath their feet, further validating the effectiveness of the proposed compensation method.

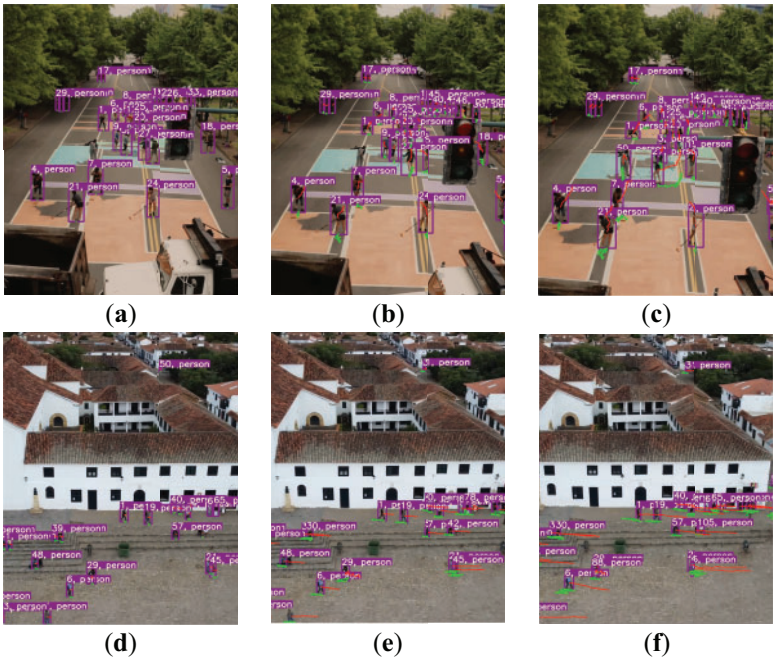


Figure 8: Movement path estimation under aerial moving perspective. (**a**)–(**c**) Drone moving forward along a road. (**d**)–(**f**) Drone moving rightward over a plaza. Green lines indicate corrected paths using the proposed method; red lines represent paths without correction

5 Conclusion

To address the challenge of detecting small targets under aerial motion, we propose the AVA-PRB model, which also enables accurate movement path estimation. The AVA-PRB model enhances detection

performance by integrating two attention mechanisms—Coordinate Attention (CA) and the Convolutional Block Attention Module (CBAM)—while utilizing Shape-IoU to improve localization accuracy. The model has been evaluated on two benchmark datasets, achieving 60.9% mAP@0.5 and 36.0% mAP@0.5:0.95 on the Aerial Person Detection dataset, and 51.2% mAP@0.5 and 29.9% mAP@0.5:0.95 on the VisDrone2019-Det dataset. In addition to detection, we further present a method for refining target movement paths affected by aerial camera motion. Our method effectively compensates for camera motion distortions, achieving a 64% reduction in path deviation and improving the accuracy of trajectory estimation. Robustness analysis across diverse environmental conditions, such as varying lighting, weather, and aerial viewpoints, further demonstrates the model's adaptability in complex aerial scenarios. Moreover, the AVA-PRB model maintains a lightweight architecture, balancing detection accuracy with computational efficiency, and sustaining inference speeds suitable for near real-time applications.

While AVA-PRB has demonstrated strong performance, several challenges and avenues for future exploration remain. Although the model enhances small object detection, challenges may arise when detecting extremely small targets or distinguishing between objects with highly similar appearances under low-resolution conditions. Furthermore, while the proposed trajectory estimation method improves path accuracy, its effectiveness may degrade in densely populated scenes where static background features are scarce. Future work will focus on optimizing computational efficiency through techniques such as quantization and pruning to facilitate deployment on embedded and edge devices. Additionally, more systematic evaluations under extreme conditions, including heavy rain, strong backlighting, and abrupt camera motion, will be conducted to validate model resilience. Broader generalization studies will also be performed by extending evaluations to additional aerial datasets such as Dataset for Object Detection in Aerial Images (DOTA), Unmanned Aerial Vehicle Detection and Tracking (UAVDT), and Dataset for Object Detection in Optical Remote Sensing Images (DIOR).

Acknowledgement: Not applicable.

Funding Statement: This research was funded by the National Science and Technology Council (NSTC), Taiwan, under grant numbers NSTC 113-2634-F-A49-007 and NSTC 112-2634-F-A49-007.

Author Contributions: Yu-Shiuan Tsai contributed to conceptualization, methodology, supervision, and manuscript preparation, including writing the original draft, reviewing, and editing. Yuk-Hang Sit was responsible for investigation, conceptualization, and methodology, as well as project administration and resource management. Additionally, Yuk-Hang Sit contributed to validation, visualization, manuscript drafting, and reviewing and editing the final version. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: The code and processed data used in this study are publicly available at https://github.com/ystsai-lab/AVA-PRB (accessed on 10 May 2025).

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest to report regarding the present study.

References

1. Bewley A, Ge Z, Ott L, Ramos F, Upcroft B. Simple online and realtime tracking. In: 2016 IEEE International Conference on Image Processing (ICIP); 2016 Sep 25–28; Phoenix, AZ, USA. p. 3464–8. doi:10.1109/ICIP.2016.7533003.
2. Zhang Y, Sun P, Jiang Y, Yu D, Weng F, Yuan Z, et al. Bytetrack: multi-object tracking by associating every detection box. In: European Conference on Computer Vision. Berlin/Heidelberg, Germany: Springer; 2022.

3. Aharon N, Orfaig R, Bobrovsky BZ. BoT-SORT: robust associations multi-pedestrian tracking. arXiv:2206.14651. 2022.
4. Wang W, Hsiang Y. SMILEtrack GitHub. [cited 2024 May 13]. Available from: https://github.com/WWangYuHsiang/SMILEtrack.
5. Redmon J, Divvala S, Girshick R, Farhadi A. You only look once: unified, real-time object detection. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2016 Jun 27–30; Las Vegas, NV, USA. p. 779–88. doi:10.1109/CVPR.2016.91.
6. Wang CY, Bochkovskiy A, Liao HM. YOLOv7: trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2023 Jun 17–24; Vancouver, BC, Canada. p. 7464–75. doi:10.1109/CVPR52729.2023.00721.
7. Jocher G, Chaurasia A, Qiu J. Ultralytics YOLO (Version 8.0.0) [Computer software]. [cited 2025 Jan 1]. Available from: https://github.com/ultralytics/ultralytics.
8. Wang CY, Yeh IH, Mark Liao HY. YOLOv9: learning what you want to learn using programmable gradient information. In: The 18th European Conference on Computer Vision ECCV 2024; 2024 Sep 29–Oct 4; Milan, Italy. doi:10.1007/978-3-031-72751-1_1.
9. Wang A, Chen H, Liu L, Chen K, Lin Z, Han J, et al. Yolov10: real-time end-to-end object detection. arXiv:2405.14458. 2024.
10. Li Y, Fan Q, Huang H, Han Z, Gu Q. A modified YOLOv8 detection network for UAV aerial image recognition. Drones. 2023;7(5):304. doi:10.3390/drones7050304.
11. Li Y. Research and application of deep learning in image recognition. In: 2022 IEEE 2nd International Conference on Power, Electronics and Computer Applications (ICPECA); 2022 Jan 21–23; Shenyang, China. p. 994–9. doi:10.1109/ICPECA53709.2022.9718847.
12. Tong Z, Chen Y, Xu Z, Yu R. Wise-IoU: bounding box regression loss with dynamic focusing mechanism. arXiv:2301.10051. 2023.
13. Chen PY, Chang MC, Hsieh JW, Chen YS. Parallel residual bi-fusion feature pyramid network for accurate single-shot object detection. IEEE Trans Image Process. 2021;30:9099–111. doi:10.1109/TIP.2021.3118953.
14. Wang G, Chen Y, An P, Hong H, Hu J, Huang T. UAV-YOLOv8: a small-object-detection model based on improved YOLOv8 for UAV aerial photography scenarios. Sensors. 2023;23(16):7190. doi:10.3390/s23167190.
15. Li X, Wei Y, Li J, Duan W, Zhang X, Huang Y. Improved YOLOv7 algorithm for small object detection in unmanned aerial vehicle image scenarios. Appl Sci. 2024;14(4):1664. doi:10.3390/app14041664.
16. Luo J, Luo R. Research on image recognition based on reinforcement learning. In: 2023 4th International Conference on Computer Vision, Image and Deep Learning (CVIDL); 2023 May 12–14; Zhuhai, China. p. 25–8. doi:10.1109/CVIDL58838.2023.10166036.
17. Zhou C, Qu Y. Application of image recognition based on deep learning in visual communication design. In: 2024 International Conference on Electrical Drives, Power Electronics & Engineering (EDPEE); 2024 Feb 27–29; Athens, Greece. p. 591–6. doi:10.1109/EDPEE61724.2024.00116.
18. Zhang H, Zhang S. Shape-iou: more accurate metric considering bounding box shape and scale. arXiv:2312.17663. 2023.
19. Tran-Anh D, Tran KL, Vu HN. License plate recognition based on multi-angle view model. arXiv:2309.12972. 2023.
20. Wang Z, Liu Z, Xu G, Cheng S. Object detection in UAV aerial images based on improved YOLOv7-tiny. In: 2023 4th International Conference on Computer Vision, Image and Deep Learning (CVIDL); 2023 May 12–14; Zhuhai, China. p. 370–4. doi:10.1109/CVIDL58838.2023.10166362.
21. Zhang L, Xiong N, Pan X, Yue X, Wu P, Guo C. Improved object detection method utilizing YOLOv7-tiny for unmanned aerial vehicle photographic imagery. Algorithms. 2023;16(11):520. doi:10.3390/a16110520.
22. Luo X, Wu Y, Wang F. Target detection method of UAV aerial imagery based on improved YOLOv5. Remote Sens. 2022;14(19):5063. doi:10.3390/rs14195063.
23. Ding K, Li X, Guo W, Wu L. Improved object detection algorithm for drone-captured dataset based on yolov5. In: 2022 2nd International Conference on Consumer Electronics and Computer Engineering (ICCECE); 2022 Jan 14–16; Guangzhou, China. p. 895–9. doi:10.1109/ICCECE54139.2022.9712813.

24. Sun T, Chen H, Liu H, Lou H, Duan X. HPS-YOLOv7: a high precision small object detection algorithm. arXiv:2813.484. 2023. doi:10.21203/rs.3.rs-2813484/v1.
25. Zhang Z, Xie X, Guo Q, Xu J. Improved YOLOv7-tiny for object detection based on UAV aerial images. Electronics. 2024;13(15):2969. doi:10.3390/electronics13152969.
26. Tan M, Pang R, Le QV. EfficientDet: scalable and efficient object detection. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2020 Jun 13–19; Seattle, WA, USA. p. 10778–87. doi:10.1109/cvpr42600.2020.01079.
27. Carion N, Massa F, Synnaeve G, Usunier N, Kirillov A, Zagoruyko S. End-to-end object detection with transformers. In: The 18th European Conference on Computer Vision ECCV; 2020 Aug 23–28; Glasgow, UK. p. 213–29. doi:10.1007/978-3-030-58452-8_13.
28. Shi W, Zhang S, Zhang S. CAW-YOLO: cross-layer fusion and weighted receptive field-based YOLO for small object detection in remote sensing. Comput Model Eng Sci. 2024;139(3):3209–31. doi:10.32604/cmes.2023.044863.
29. Ashish V. Attention is all you need. Adv Neural Inf Process Syst. 2017;30:1–11.
30. Hou Q, Zhou D, Feng J. Coordinate attention for efficient mobile network design. In: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2021 Jun 20–25; Nashville, TN, USA. p. 13708–17. doi:10.1109/cvpr46437.2021.01350.
31. Woo S, Park J, Lee JY, Kweon IS. CBAM: convolutional block attention module. In: Proceedings of the European Conference on Computer Vision (ECCV); 2018 Sep 8–14; Munich, Germany. p. 3–19. doi:10.1007/978-3-030-01234-2_1.
32. RizwanMunawar. yolov8-object-tracking. [cited 2023 Nov 20]. Available from: https://github.com/RizwanMunawar/yolov8-object-tracking.
33. OpenCV. Feature Matching + Homography to find Objects. [cited 2023 Nov 20]. Available from: https://docs.opencv.org/3.4/d1/de0/tutorial_py_feature_homography.html.
34. DeTone D, Malisiewicz T, Rabinovich A. SuperPoint: self-supervised interest point detection and description. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW); 2018 Jun 18–22; Salt Lake City, UT, USA. p. 337–12. doi:10.1109/CVPRW.2018.00060.
35. Detection AP. Aerial person detection dataset. [cited 2025 May 11]. Available from: https://universe.roboflow.com/aerial-person-detection/aerial-person-detection.
36. Zhu P, Wen L, Du D, Bian X, Fan H, Hu Q, et al. Detection and tracking meet drones challenge. IEEE Trans Pattern Anal Mach Intell. 2022;44(11):7380–99. doi:10.1109/tpami.2021.3119563.