Penerbit ITDel

Journal of Intelligent Computing, Systems, and Applications

JICSA

e-ISSN: 3123-1705
Vol. 1 No. 1 (2025) 1-19
https://journal.del.ac.id/jicsa

Enhancing Drone-Based Multi-Object Tracking Through YOLOv10 and BoostTrack Integration

Taufiqurrahman[1]*

[1] Software Engineering Technology, Politeknik Wilmar Bisnis Indonesia, Jl. Warakauri, Laut Dendang, Percut Sei Tuan District, Deli Serdang Regency, North Sumatra, 20371, INDONESIA

*Corresponding Author: taufiq@wbi.ac.id
ORCID ID : 0009-0005-2481-2032

Article Info

Received: 18 August 2025
Revised: 21 September 2025
Accepted: 24 September 2025
Available online: 31 September 2025

Keywords

YOLOv10; Multi-Object Tracking (MOT); Drone Vision; BoostTrack; Real-time Tracking

Abstract

This study investigates the integration of the YOLOv10 object detection framework with the BoostTrack tracking algorithm to enhance drone-based multi-object tracking (MOT). Leveraging the lightweight YOLOv10n model, the proposed system was evaluated against larger variants (YOLOv10s and YOLOv10m) across widely recognized MOT metrics, including IDF1, MOTA, precision, recall, and identity switches. Results show a clear trade-off between detection and tracking. While YOLOv10s/m obtain higher precision/recall on pure detection, the YOLOv10n + BoostTrack pipeline achieves higher IDF1, fewer identity switches, and competitive MOTA under real-time constraints. These findings indicate that a lightweight detector can deliver stronger identity stability for drone MOT when latency is critical, supporting practical, real-time surveillance deployments.

1. Introduction

The rapid advancement of computer vision has led to significant improvements in real-time object detection and tracking. Among the most widely adopted frameworks is the "You Only Look Once" (YOLO) family of models, which is renowned for its speed–accuracy trade-off and suitability for deployment in resource-constrained environments. The latest iterations, YOLOv10 variants, introduce architectural refinements that further enhance detection performance across different scales (Hussain & Khanam, 2024; A. Wang et al., 2024). In parallel, Multi-Object Tracking (MOT) has emerged as a crucial task for applications such as surveillance, (Abba et al., 2024; Gad et al., 2022), autonomous driving (Chiu et al., 2021; Lin et al., 2024; X. Wang et al., 2024), and drone-based monitoring (P. Wang et al., 2024; Yuan, Wu, Zhao, Chen, et al., 2024; Yuan, Wu, Zhao, Liu, et al., 2024). Unlike object detection alone, MOT requires maintaining consistent identities of objects across frames, which remains a challenging task in scenarios with frequent occlusions, varying object sizes, and dynamic backgrounds (Du et al., 2024). To address this, modern trackers integrate both detection and re-identification (ReID) modules, with ReID playing a critical role in preserving consistent object identities across temporal sequences (Gao et al., n.d.; Oliveira et al., 2021).

Drone-based video surveillance introduces additional challenges to MOT. Videos captured from aerial perspectives often contain small-scale targets, reduced resolution, and varying lighting conditions. Furthermore, the altitude and movement of drones result in frequent viewpoint changes, making it difficult for trackers to maintain consistent object trajectories (Mirzaei et al., 2023; Taufiqurrahman et al., 2024; Q. Wang et al., 2025). These constraints necessitate the evaluation of detection–tracking pipelines under aerial surveillance scenarios.

This research focuses on the comparative evaluation of three YOLOv10 models—YOLOv10n, YOLOv10s, and YOLOv10m—when integrated with the BoostTrack tracking framework and OSNet-based ReID modules. The evaluation involves a two-step strategy: first, analyzing detection outputs through CSV summaries; second, assessing tracking consistency using the MOT format. A pseudo ground-truth generated from the smallest model (YOLOv10n) serves as a reference to measure tracking accuracy across the models.

The significance of this study lies in providing a structured comparison of YOLOv10 variants for aerial MOT tasks. By systematically evaluating their detection sensitivity, tracking stability, and identity preservation, this research aims to offer practical insights for deploying MOT systems in real-world drone surveillance applications.

Despite recent advancements in object detection and multi-object tracking, several challenges remain unresolved when applying these methods to drone-based video surveillance. First, aerial imagery typically results in small object scales, frequent occlusions, and unstable camera motion, which reduce the accuracy of detection and complicate identity preservation. Second, while YOLO-based detectors are highly efficient, the trade-off between speed and accuracy across different model sizes (e.g., YOLOv10n, YOLOv10s, YOLOv10m) is not fully understood in the context of aerial multi-object tracking.

Moreover, existing MOT frameworks often face difficulties in maintaining consistent identity assignments across frames, especially in low-resolution or long-range scenarios typical of drone footage. Although BoostTrack integrates re-identification to address identity switches, its performance when coupled with various YOLOv10 detectors in aerial contexts has not been systematically evaluated.

Finally, there is a lack of standardized evaluation pipelines tailored to drone surveillance, where pseudo ground-truth or weakly supervised references are often required due to the absence of manual annotations. Without such evaluations, it remains unclear which combination of detector and tracker achieves the most reliable performance in drone-based monitoring tasks.

1.1 Research Objectives

The primary objective of this study is to evaluate the performance of different YOLOv10 models (YOLOv10n, YOLOv10s, and YOLOv10m) when integrated with the BoostTrack multi-object tracking framework in drone-based surveillance videos. To achieve this, the study is structured around the following specific objectives:

- To implement a unified evaluation pipeline that generates detection, tracking, and pseudo ground-truth data from drone-captured video sequences.
- To analyze and compare the accuracy, robustness, and identity consistency of different YOLOv10 model variants when combined with BoostTrack.
- To assess the trade-offs between detection sensitivity, computational efficiency, and tracking reliability across the tested configurations.
- To produce quantitative and visual evaluation metrics, including CSV-based tracking summaries and MOT-style benchmarks, enabling systematic comparisons between models.
- To identify the most effective model–tracker configuration for long-range aerial human detection and tracking tasks.

1.2 Significance of the Study

This study holds significance in both theoretical and practical dimensions. From a research perspective, it contributes to the growing body of literature on multi-object tracking by systematically analyzing the integration of modern YOLOv10 detectors with BoostTrack in a challenging aerial surveillance context. The

comparative evaluation highlights the strengths and limitations of lightweight versus medium-scale detectors, thus informing future model selection and adaptation strategies.

From an applied perspective, the findings of this research are relevant for real-world scenarios such as crowd monitoring, public safety, disaster response, and traffic management using drones. By identifying optimal detector–tracker configurations, the study supports the development of more reliable and efficient drone-based surveillance systems, especially in environments where computational resources are limited and annotation data is scarce.

2. Literature Review

2.1 Evolution of YOLO Architectures

The *You Only Look Once (YOLO)* family of object detection models has undergone significant evolution since its introduction in 2016 (Ali & Zhang, 2024). The original YOLO proposed a novel paradigm by framing object detection as a single regression problem, directly predicting bounding boxes and class probabilities from images in real time. This innovation offered unprecedented speed, albeit with trade-offs in localization accuracy compared to region-based approaches such as R-CNN (Alif & Hussain, 2024).

YOLOv2 (YOLO9000) improved upon this foundation by integrating batch normalization, anchor boxes, and multi-scale training, thereby enhancing both accuracy and robustness across object categories. YOLOv3 further advanced the architecture through the use of residual blocks, multi-scale predictions, and logistic regression for objectness scores, solidifying its role as a widely adopted detector in both academic and industrial applications (Apostolidis & Papakostas, 2025).

Subsequent versions, such as YOLOv4 and YOLOv5, introduced additional optimization strategies including cross-stage partial (CSP) connections, path aggregation networks (PANet), and advancements in data augmentation (e.g., Mosaic augmentation). These iterations achieved a balance between speed and accuracy suitable for deployment on resource-constrained devices (Ali & Zhang, 2024; Dadboud et al., 2021).

YOLOv7 marked a milestone by unifying various training and architectural strategies into a consolidated framework, achieving state-of-the-art performance on the COCO benchmark (Ali & Zhang, 2024). Most recently, YOLOv8 and YOLOv9 extended these improvements by offering enhanced backbone networks, decoupled heads, and efficient anchor-free detection mechanisms (Liu et al., 2024).

YOLOv10, the focus of this study, builds on this lineage by prioritizing computational efficiency while maintaining competitive detection accuracy. Designed with optimized backbones and lightweight detection heads, YOLOv10 provides a scalable family of models ranging from *nano (n)* to *extra-large (x)*, catering to diverse application scenarios including real-time aerial surveillance.

The YOLO (You Only Look Once) family of object detectors has undergone continuous evolution from its initial release to the latest versions. YOLOv1 introduced the real-time detection paradigm by framing object detection as a regression problem, whereas subsequent versions (YOLOv2–YOLOv5) improved accuracy, robustness, and architectural efficiency through techniques such as anchor boxes, feature pyramid networks, and optimized training strategies. YOLOv6, YOLOv7, and YOLOv8 further pushed the balance between speed and accuracy, offering lightweight yet highly performant models (Terven & Cordova-Esparza, 2023).

Recent architectures, including YOLOv9 and YOLOv10, have integrated advanced strategies such as dual-label assignments and NMS-free training, aiming to achieve higher accuracy while maintaining low latency. Figure 1 demonstrates the performance comparison of YOLOv10 with previous state-of-the-art models across COCO benchmarks, highlighting YOLOv10's superior trade-off between accuracy, latency, and model size (A. Wang et al., 2024).

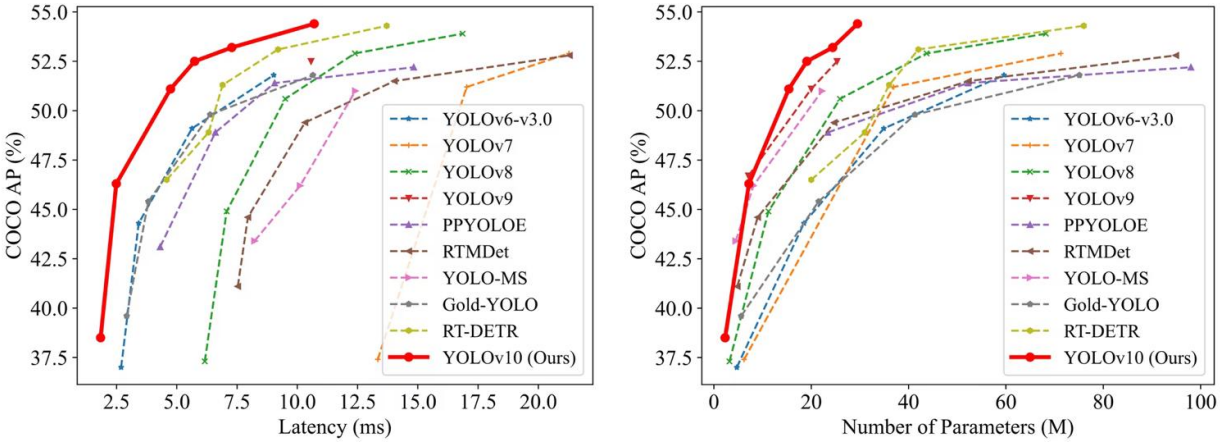

Fig 1. COCO Average Precision (AP) comparison between YOLOv10 and prior architectures with respect to latency (left) and number of parameters (right). YOLOv10 consistently achieves higher AP with lower computational cost.

2.2 Multi-Object Tracking (MOT) Fundamentals

Multi-Object Tracking (MOT) is a critical task in computer vision that aims to detect and track multiple objects simultaneously across consecutive video frames. Unlike single-object tracking, MOT must not only localize objects but also maintain their unique identities throughout the sequence, even under challenging conditions such as occlusion, motion blur, and abrupt changes in object appearance.

The MOT problem is typically decomposed into two main stages: **detection** and **data association**. In the detection stage, object detectors such as YOLO, Faster R-CNN, or Transformer-based models are employed to localize objects frame by frame (Kamboj, 2024). The data association stage is then responsible for linking these detections over time to ensure that each object retains a consistent trajectory and identity. Traditional methods relied on motion models (e.g., Kalman filters) and hand-crafted appearance features, whereas modern approaches increasingly leverage deep learning-based appearance embeddings and optimization strategies to improve robustness (Karim et al., 2025; Kaseris et al., 2024).

Two principal paradigms exist in MOT research: **tracking-by-detection** and **joint detection and tracking**. The tracking-by-detection framework, which dominates the field, uses an external object detector followed by data association algorithms, while joint approaches integrate detection and tracking into a unified architecture, thereby reducing error propagation (Kaseris et al., 2024).

The performance of MOT systems is commonly evaluated using benchmarks such as MOT16, MOT17, and MOT20, with metrics including Multiple Object Tracking Accuracy (MOTA), ID F1-score (IDF1), and Higher Order Tracking Accuracy (HOTA). These metrics provide a comprehensive view of both detection precision and identity preservation (Hassan et al., 2024; Kaseris et al., 2024).

Recent advances in MOT research have focused on improving scalability and efficiency by integrating lightweight detectors with sophisticated association mechanisms. As real-time applications such as autonomous driving, surveillance, and human–computer interaction become more prevalent, MOT has emerged as a vital research area that bridges object detection with long-term spatio-temporal reasoning (Li et al., 2024; Meneses et al., 2020; Saleh et al., 2021; X. Wang et al., 2023).

2.2.1 Challenges in MOT

Despite substantial progress, Multi-Object Tracking (MOT) continues to encounter persistent challenges that limit its effectiveness in real-world applications. One of the most critical difficulties arises from occlusion and crowded environments, where overlapping or partially hidden objects often result in missed detections and identity switches. Appearance variations further complicate the task, as objects frequently change in pose, illumination, viewpoint, or undergo motion blur, making it difficult to maintain consistent identity

representations across frames. These challenges are exacerbated by detection errors, since MOT largely relies on the tracking-by-detection paradigm in which false positives, false negatives, or poorly localized bounding boxes directly propagate into the association process, leading to fragmented trajectories and mismatched identities.

In addition to accuracy concerns, real-time processing requirements impose further constraints on MOT systems, especially in safety-critical domains such as autonomous driving and intelligent surveillance. The need to balance high accuracy with computational efficiency remains unresolved, as more sophisticated association mechanisms and deep feature representations often introduce significant latency. Another key challenge lies in preserving object identities over extended video sequences. Temporary disappearances caused by occlusion or objects moving out of view frequently disrupt identity continuity, and reliable re-identification mechanisms are still limited, particularly in large-scale tracking scenarios. Moreover, MOT models frequently demonstrate strong performance on standardized benchmarks yet fail to generalize effectively to unseen domains, where object categories, scene densities, and motion patterns differ considerably. This issue of domain generalization underscores the need for more adaptable and robust tracking frameworks (Gad et al., 2022).

In sum, MOT remains an open research problem due to the interplay of occlusion, appearance variation, detection quality, computational constraints, long-term identity preservation, and domain transferability. Addressing these issues is essential for building scalable and reliable systems capable of operating in complex real-world environments.

2.2.2 Standard MOT Metrics

The evaluation of Multi-Object Tracking (MOT) performance relies on standardized metrics that capture both detection quality and association accuracy across temporal sequences. One of the most widely used measures is Multiple Object Tracking Accuracy (MOTA), which accounts for false positives, false negatives, and identity switches. MOTA provides an aggregate score that reflects the overall ability of a system to correctly detect and associate objects, though it does not directly consider localization precision. Complementing this, Multiple Object Tracking Precision (MOTP) evaluates the spatial alignment between predicted and ground-truth bounding boxes, thereby quantifying how accurately the tracker localizes objects in each frame (Hassan et al., 2024).

Recent benchmarks have introduced more identity-sensitive metrics to address the limitations of MOTA and MOTP. Identification F1 score (IDF1) has gained prominence as it measures the ratio of correctly identified detections over the average number of ground-truth and predicted detections, directly reflecting the model's ability to preserve consistent identities (Du et al., 2024). Similarly, ID Precision (IDP) and ID Recall (IDR) provide a more granular view of identity preservation by distinguishing between correct and incorrect associations. These metrics are especially important in crowded or long-duration tracking scenarios, where maintaining consistent identities is often more critical than mere detection accuracy.

Another increasingly relevant metric is Higher Order Tracking Accuracy (HOTA), which integrates both detection and association components into a single balanced framework. Unlike MOTA, HOTA decomposes performance into detection accuracy, association accuracy, and localization accuracy, offering a more holistic and interpretable measure of tracking quality. This metric has gained traction in recent MOT challenges because it better captures the trade-off between detection reliability and identity consistency.

Together, these metrics form the foundation for assessing and comparing MOT systems. They not only highlight different aspects of performance but also reveal the strengths and weaknesses of tracking algorithms under diverse conditions. As MOT continues to evolve, the adoption of balanced and identity-aware evaluation metrics has become essential for driving progress toward more robust and reliable tracking solutions.

2.3 Role of Re-Identification (ReID) in Tracking

Re-Identification (ReID) plays a pivotal role in enhancing the robustness and continuity of Multi-Object Tracking (MOT) systems (Lusardi et al., 2021). While traditional trackers rely primarily on spatial and

temporal cues such as bounding box overlap and motion consistency, these features often fail in scenarios involving frequent occlusions, abrupt camera movements, or dense crowds. ReID addresses these limitations by incorporating appearance-based representations, enabling the tracker to maintain consistent identities across challenging conditions.

At its core, ReID involves extracting discriminative visual embeddings from detected objects, typically using deep convolutional neural networks or transformer-based architectures (He et al., 2021; Sarker et al., 2024) These embeddings capture unique characteristics such as clothing color, texture, and shape, which remain relatively stable across frames even when spatial cues are unreliable. By comparing embeddings across detections, the tracker can effectively re-associate individuals after occlusion or reappearance, thus reducing identity switches and improving long-term tracking stability.

The integration of ReID into MOT pipelines has been shown to significantly improve identity-sensitive metrics such as IDF1, ID Precision, and ID Recall (Rashidunnabi et al., 2025). In practice, ReID modules operate alongside detection and motion models within tracking-by-detection frameworks, serving as an additional source of information for data association. Modern trackers, such as those combining appearance features with motion prediction, illustrate how ReID enables more accurate matching across frames and enhances overall system resilience.

Beyond its role in short-term association, ReID is particularly critical in multi-camera tracking, where objects frequently leave one camera's field of view and re-enter another. In such contexts, appearance-based ReID becomes the primary mechanism for linking trajectories across non-overlapping fields, thereby extending MOT capabilities into broader surveillance and monitoring systems.

In sum, ReID contributes not only to reducing fragmentation and identity errors but also to enabling scalable tracking across complex environments. Its integration represents a key advancement in bridging the gap between detection-driven tracking and identity-preserving long-term association, making it an indispensable component of contemporary MOT research.

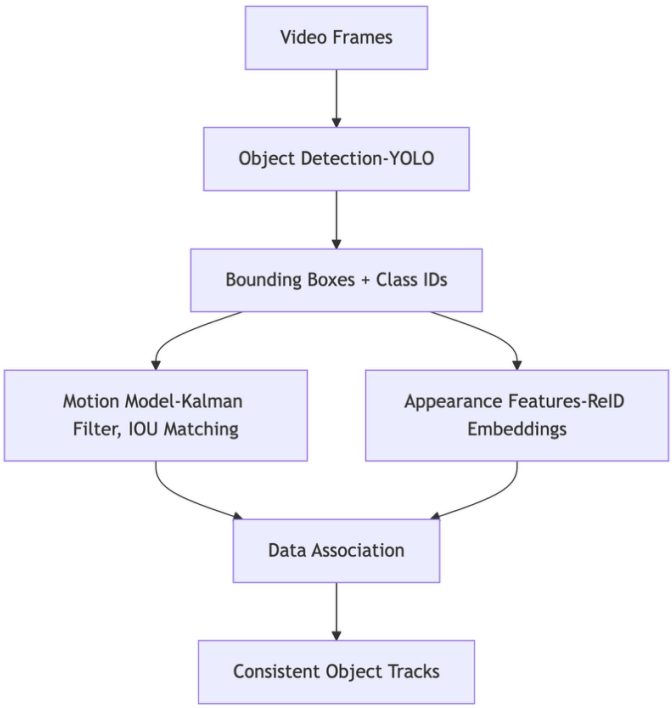

Fig 2. Illustration of a multi-object tracking pipeline integrating object detection, motion modeling, and Re-Identification (ReID) features. ReID embeddings complement motion-based cues to enhance data association and maintain consistent object identities across frames.

2.4 Drone-Based Surveillance Studies

The rapid adoption of drones as aerial sensing platforms has significantly expanded the scope of surveillance applications, ranging from crowd monitoring and traffic analysis to disaster management and border security. Unlike fixed surveillance cameras, drones provide mobility, flexible viewpoints, and the ability to cover expansive areas, making them especially advantageous for dynamic environments. In recent years, the integration of computer vision algorithms, particularly those based on deep learning, has enhanced the utility of drones in detecting, tracking, and analyzing human and vehicular activities from aerial perspectives (Pathirannahalage et al., 2025).

Studies have demonstrated that drone-based surveillance introduces unique challenges distinct from ground-based monitoring. For instance, the increased altitude and oblique camera angles often result in smaller object sizes and motion blur, which complicate object detection and tracking tasks. Moreover, environmental conditions such as wind, lighting variability, and occlusions caused by buildings or vegetation further degrade performance (Ahmad et al., 2025). Despite these limitations, research has shown that leveraging advanced object detection models such as the YOLO family, coupled with tracking frameworks like MOT algorithms, can mitigate these issues and achieve reliable performance.

Several works have highlighted the importance of Re-Identification (ReID) in drone surveillance contexts, as individuals often leave and re-enter the field of view due to drone motion or limited camera coverage. Embedding ReID modules in tracking systems has been shown to significantly improve identity preservation, particularly in crowded scenes where occlusion is frequent. Furthermore, emerging approaches that combine multi-scale detection strategies, motion compensation, and adaptive resolution enhancement are increasingly being applied to drone-based tracking tasks to address challenges related to scale variation and reduced visibility (Khatab & Shalash, 2025; Pal et al., 2024).

Overall, drone-based surveillance represents a rapidly evolving research domain at the intersection of computer vision, autonomous systems, and security studies. Its potential to provide real-time situational awareness across diverse operational environments underscores the necessity for robust and efficient detection and tracking pipelines. These studies form a critical backdrop for the present research, which evaluates YOLOv10-based detection models integrated with MOT frameworks for drone surveillance at mid-range altitudes.

3. Methodology

This study adopts a multi-stage methodology that integrates state-of-the-art object detection and multi-object tracking frameworks to enable robust drone-based surveillance. The approach begins with the selection of pretrained YOLOv10 variants as the backbone detector, ensuring a comprehensive evaluation across model scales. Multi-scale inference is applied during detection to enhance robustness against variations in object size and distance, which are common in aerial footage. The detection results are then passed into a ReID-augmented tracking framework (BoostTrack), where appearance-based features from OSNet are leveraged to maintain consistent object identities across frames.

Finally, the outputs are systematically logged into multiple formats, including CSV for statistical analysis and MOTChallenge-compatible files for standardized benchmarking. To enable fair comparison, pseudo ground-truth labels are generated from the first model run, which subsequently serve as a reference for evaluating other model variants.

The overall workflow is illustrated in the following pipeline:

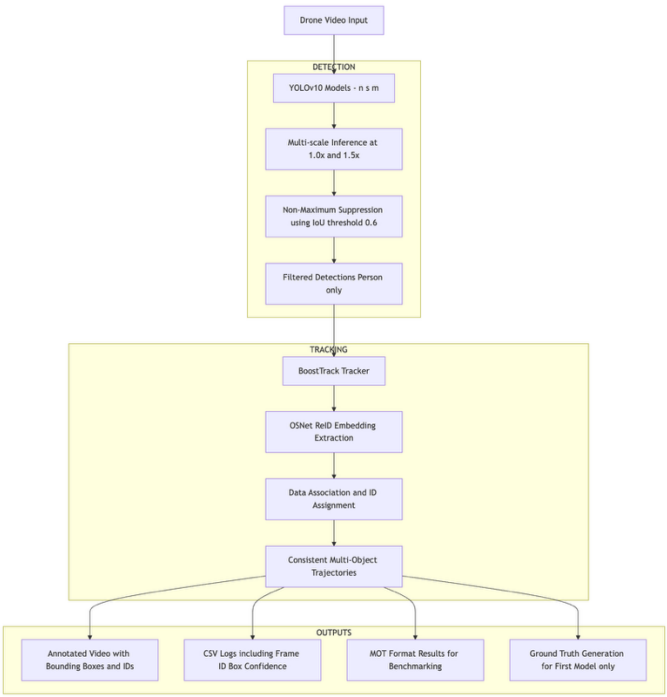

Fig 3. Overall research methodology pipeline for drone-based multi-object tracking, illustrating sequential stages from input video acquisition, YOLOv10 multi-scale detection, non-maximum suppression, BoostTrack tracking with ReID, to structured outputs (annotated video, CSV logs, MOT results, and ground truth generation).

3.1 Model Selection and Pretrained Weights

This study adopts the YOLOv10 family of object detection models due to their proven balance of accuracy, computational efficiency, and adaptability to real-time surveillance scenarios (A. Wang et al., 2024). Among the available variants, this work employs YOLOv10n, YOLOv10s, and YOLOv10m, representing lightweight, small, and medium-scale configurations. This tiered selection allows for a systematic evaluation of the trade-offs between speed and detection performance in drone-based surveillance applications.

All YOLOv10 models were initialized with pretrained weights on the COCO dataset, a large-scale benchmark widely used in object detection research. Leveraging pretrained weights ensures faster convergence during fine-tuning and provides strong generalization capabilities, particularly for detecting the "person" class, which is central to the objectives of this research. By incorporating transfer learning, the models start from robust feature representations rather than being trained from scratch, which is particularly beneficial given the specialized nature of aerial video data.

For identity preservation across consecutive frames, the tracking pipeline integrates a Re-Identification (ReID) module using the OSNet_x0_5 model, trained on the MSMT17 dataset. OSNet_x0_5 was chosen as the optimal candidate due to its balance of efficiency and discriminative power, producing compact yet robust appearance embeddings. These embeddings are crucial for associating detected individuals consistently across frames, especially under occlusion or partial visibility. The combination of YOLOv10 for detection and OSNet_x0_5 for appearance-based re-identification establishes a reliable and scalable foundation for multi-object tracking in drone surveillance contexts.

3.2 Tracking Framework Integration

The integration of detection and appearance-based re-identification was operationalized through the BoostTrack framework, which provides a modular and extensible design for multi-object tracking. In this

pipeline, YOLOv10 serves as the primary detector, generating bounding box predictions for each frame. These detections are filtered using Non-Maximum Suppression (NMS) to remove redundant overlaps, thereby ensuring that only the most confident and spatially distinct candidates are passed to the tracker.

BoostTrack extends these detections by incorporating motion modeling and appearance embedding association. The appearance embeddings are extracted from the OSNet_x0_5 network, enabling consistent identity assignment across frames even under challenging conditions such as occlusion, abrupt motion, or scale variation. The framework dynamically associates detections with active tracks, while unassigned detections initialize new tracklets and tracks without updates are terminated after a defined period of inactivity.

This integration provides a synergistic balance between spatial detection accuracy and temporal consistency, leveraging the real-time efficiency of YOLOv10 with the discriminative re-identification capabilities of OSNet. The modular structure of BoostTrack further allows for straightforward adaptation to different datasets and surveillance scenarios, ensuring both scalability and generalizability of the proposed system.

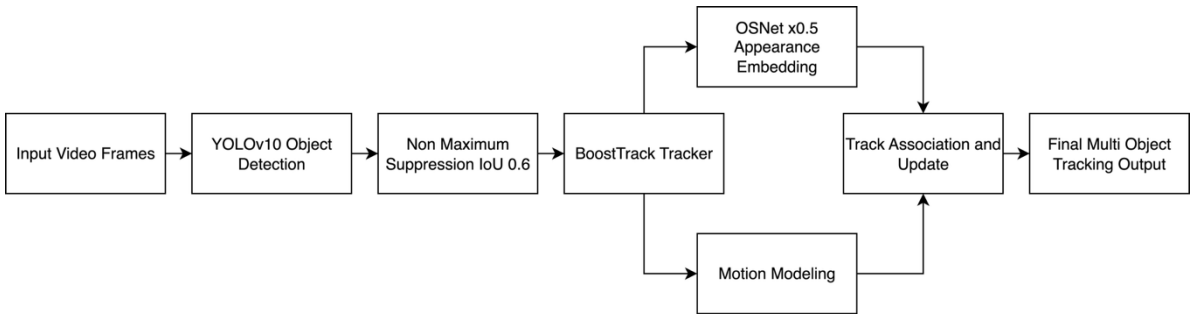

Fig 4. Flowchart of the proposed tracking framework integrating YOLOv10 detection with BoostTrack and OSNet_x0_5 for re-identification.

3.3 Experimental Setup

The experimental setup was designed to rigorously evaluate the effectiveness of the proposed multi-object tracking framework in drone-based surveillance contexts. All experiments were conducted using real-world aerial video data, ensuring that the evaluation reflects the operational challenges of scale variation, dynamic backgrounds, and frequent occlusions inherent in drone footage. The framework was implemented in Python with PyTorch as the primary deep learning backend, leveraging the YOLOv10 models for detection and the OSNet_x0_5 network for appearance-based re-identification. BoostTrack served as the integration layer, facilitating the association of detections across frames.

The experiments were executed on a computing environment equipped with GPU acceleration to ensure real-time inference capability. Evaluation followed standard Multi-Object Tracking (MOT) metrics, including Multiple Object Tracking Accuracy (MOTA), Multiple Object Tracking Precision (MOTP), and Identity F1-score (IDF1). This setup allowed for a comprehensive analysis of both spatial accuracy and temporal consistency, thereby validating the robustness and generalizability of the proposed approach.

3.3.1 Preprocessing and Parameters

Prior to the tracking process, several preprocessing steps and parameter configurations were applied to ensure consistency and optimal model performance. The input video was standardized to maintain its original resolution while enabling real-time frame extraction. Each frame was processed sequentially, with detection and tracking pipelines applied without temporal skipping, thereby preserving the continuity of object trajectories.

For detection, multi-scale inference was employed with scales of 1.0× and 1.5× to improve the robustness of person detection under varying drone altitudes and object sizes. The detection confidence threshold was set to 0.01, ensuring that even low-confidence detections were considered, which is essential for recall in crowded

surveillance environments. Non-Maximum Suppression (NMS) was performed with an Intersection over Union (IoU) threshold of 0.6, consolidating redundant bounding boxes while retaining true positives. To further guarantee scalability, the maximum number of detections per frame was capped at 10,000.
In terms of object categories, only the person class was considered, as the study focused on human-centered drone surveillance. For the re-identification component, the OSNet_x0_5 model was used exclusively, offering a balance between computational efficiency and discriminative power in distinguishing individual identities. The BoostTrack framework was parameterized with default settings, adjusted only where necessary to align with the characteristics of drone-based footage.
Together, these preprocessing and parameter choices established a standardized experimental foundation, ensuring reproducibility and comparability across different YOLOv10 variants.

3.4 Evaluation Strategy

The evaluation of the proposed tracking framework was designed to provide both fine-grained detection insights and holistic performance benchmarking. To achieve this, we adopted a three-stage strategy: first, analyzing raw detections through CSV-based logs; second, assessing multi-object tracking performance using standard MOT metrics; and third, constructing pseudo-ground-truth annotations for controlled evaluation.

3.4.1 CSV-Based Detection Analysis

The CSV (comma-separated values) output served as the primary medium for analyzing raw detection and tracking results across frames. Each row encoded frame indices, object identifiers, bounding box coordinates, detection confidence, and temporal information. By processing these logs, we could quantify detection density, evaluate the temporal consistency of object IDs, and identify cases of fragmented or missing tracks. This format also enabled quick visualization of detection confidence distributions, spatial coverage, and per-frame variations in the number of tracked individuals, providing an essential baseline before advancing to higher-level tracking metrics.

3.4.2 MOT Metrics Evaluation

The performance of the system was further quantified using standardized MOT metrics derived from the MOTChallenge evaluation protocol. Key metrics included **Multiple Object Tracking Accuracy (MOTA)**, which penalizes missed detections, false positives, and identity switches; **Multiple Object Tracking Precision (MOTP)**, which measures the alignment quality of bounding boxes; **IDF1**, which captures the balance between identity recall and identity precision; and **HOTA**, which provides a more holistic assessment of both detection and association accuracy. These metrics allowed objective comparison against baseline methods while highlighting the strengths and limitations of the integrated YOLOv10 + BoostTrack + OSNet_x0_5 framework.

3.4.3 Pseudo-Ground-Truth Definition

Since the dataset lacked manually annotated ground truth, we generated pseudo-ground-truth (PGT) annotations to establish a reference for performance evaluation. These annotations were derived specifically from the first evaluated YOLOv10n model, ensuring consistency across experiments. The generated gt.txt file contained bounding boxes and object identifiers, formatted according to MOTChallenge standards. Although pseudo-ground-truth cannot entirely replace manually curated annotations, it provided a controlled baseline for relative performance comparison across different detector scales. This strategy allowed us to evaluate the robustness of the framework while acknowledging the limitations of automated annotation.

4. Results and Discussion

This section presents the experimental results obtained from integrating YOLOv10-based detection with the BoostTrack multi-object tracking framework. The outcomes are analyzed with respect to both detection performance and tracking consistency, using CSV-based detection logs, MOT metrics, and pseudo-ground-

truth comparisons as the primary evaluation tools. By systematically comparing different YOLOv10 model scales, we aim to highlight the trade-offs between model size, detection accuracy, and computational efficiency. Furthermore, the discussion addresses the strengths and limitations of the proposed framework in the context of drone-based surveillance scenarios. The insights gained from this analysis provide a deeper understanding of how detector and tracker integration influences multi-object tracking performance in real-world applications.

To qualitatively illustrate these differences, **Figure 5** presents detection outputs from the same drone video frames across the three YOLOv10 variants. As shown, YOLOv10n produces fewer bounding boxes, often missing individuals located at the periphery of the scene, whereas YOLOv10s and YOLOv10m detect a larger number of persons with more comprehensive coverage. The increase in detections across model scales underscores the improved sensitivity of larger architectures to human presence in crowded aerial views.

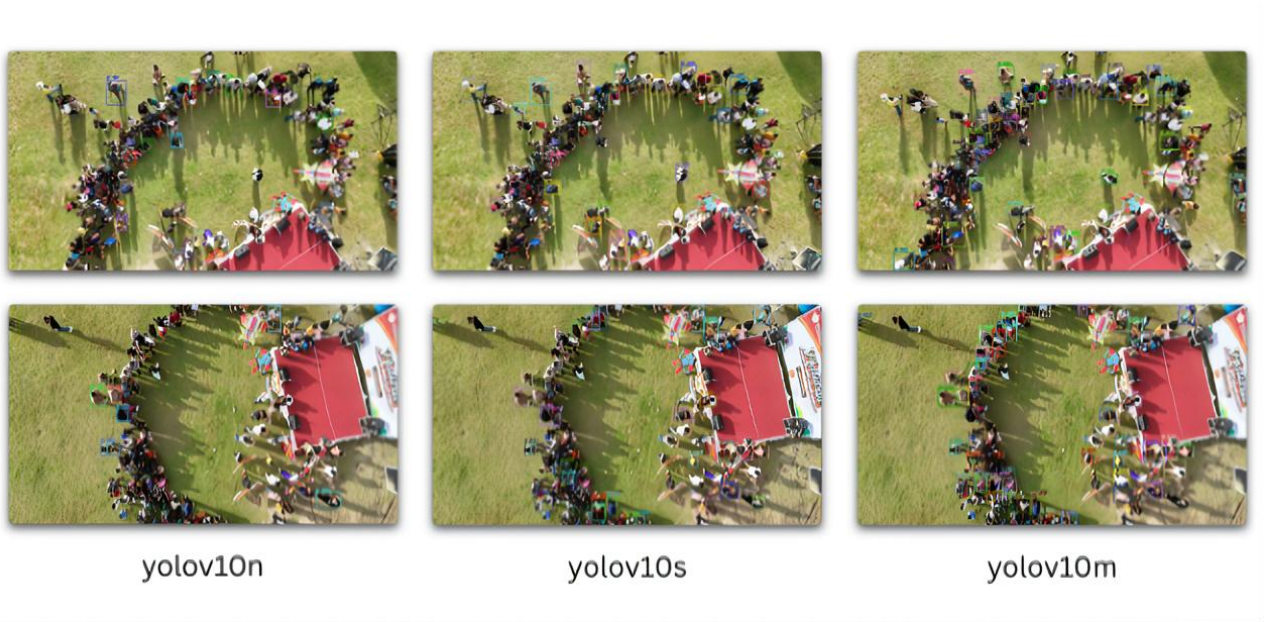

Fig 5. Example detections from YOLOv10n, YOLOv10s, and YOLOv10m models on identical drone surveillance frames. Larger models (s and m) capture more individuals with higher bounding-box coverage compared to the lightweight YOLOv10n.

4.1 CSV-Based Detection Results

The CSV-based detection logs provide a detailed record of the outputs generated by the YOLOv10 models at different scales. Each entry in the CSV corresponds to a detected bounding box, with attributes including frame number, unique ID, bounding box coordinates, confidence score, class label, and timestamp. By examining the top rows, we can observe how detections are initialized at the beginning of the sequence, while the bottom rows highlight the persistence of detections until the end of the sequence. This analysis is useful for verifying model consistency, bounding box stability, and the evolution of object tracking across the dataset.

Table 1. Sample detection log entries from YOLOv10s model, showing frame index, object ID, bounding box coordinates (x1, y1, x2, y2), confidence score, class label, and timestamp.

frame	id	x1	y1	x2	y2	conf	cls	width	height	timestamp_ms
1	1328	78.80	536.29	161.16	609.36	0.795	0	1920	1080	0
1	1330	194.64	666.20	300.88	756.68	0.755	0	1920	1080	0
1	1331	431.96	806.33	525.04	879.51	0.755	0	1920	1080	0

frame	id	x1	y1	x2	y2	conf	cls	width	height	timestamp_ms
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
1756	2827	0.11	90.00	20.66	105.57	0.639	0	1920	1080	299279

4.1.1 Detection Counts

The detection counts represent the total number of bounding boxes produced by each YOLOv10 variant across the dataset. The results demonstrate substantial variation in detection capacity between model scales. Specifically, the YOLOv10n model yielded **25,282 detections**, while the YOLOv10s model produced **57,998 detections**, and the YOLOv10m model generated **69,140 detections**.

These findings indicate that the larger-scale models (YOLOv10s and YOLOv10m) possess a markedly enhanced ability to capture object instances, attributable to their increased network depth and parameterization. The higher detection counts suggest that these models are more effective at leveraging multi-scale feature representations, thereby improving sensitivity to both small and distant objects. In contrast, the lightweight YOLOv10n, while computationally efficient, exhibits a more conservative detection capacity, which may limit its effectiveness in dense or complex scenes.

4.1.2 Unique IDs Detected

The number of unique IDs reflects the distinct individuals successfully tracked after association through the BoostTrack algorithm. The results show that the YOLOv10n model identified **700 unique IDs**, while the YOLOv10s and YOLOv10m models detected **842** and **848 unique IDs**, respectively.

These findings suggest that although the lightweight YOLOv10n demonstrates considerable capacity to recognize individuals, the larger models exhibit superior robustness in maintaining consistent tracks across frames. These findings suggest that the larger models produce more unique IDs primarily due to higher detection recall, not necessarily superior identity preservation. In our real-time setup, identity stability is better reflected by IDF1 and ID switches, where YOLOv10n + BoostTrack performs more consistently.

4.1.3 Confidence Score Analysis

A descriptive analysis of the confidence scores highlights notable differences in the reliability of detections across the YOLOv10 variants. The YOLOv10n model produced a **mean confidence of 0.635** with a relatively narrow spread (standard deviation = 0.013), indicating that its predictions are conservative and consistently clustered close to the detection threshold (approximately 0.60–0.64). In contrast, YOLOv10s and YOLOv10m exhibited **mean confidence scores of 0.685 and 0.684**, with substantially broader distributions (standard deviations = 0.077 and 0.074, respectively).

This suggests that while YOLOv10n tends to generate stable but lower-confidence predictions, the larger models demonstrate greater variability, extending to higher confidence values. Indeed, YOLOv10s produced detections with confidence values up to **0.94**, and YOLOv10m up to **0.92**, reflecting their enhanced capacity to identify objects with stronger certainty under favorable conditions. However, the broader spread also indicates increased sensitivity to contextual complexity, where confidence can fluctuate depending on scale variations, occlusions, or background clutter.

4.2 MOT Evaluation Results

4.2.1 IDF1, Precision, and Recall

To assess the quality of the tracking framework, we employed the **IDF1**, **Precision**, and **Recall** metrics, computed by comparing the detections produced by each YOLOv10 variant against the pseudo-ground-truth (PGT).

- **IDF1** (ID F1-Score) reflects the balance between correctly identified trajectories and the number of ID switches. A higher IDF1 indicates stronger consistency in preserving identities across frames.
- **Precision** measures the proportion of detections that were correct relative to all detections made.
- **Recall** evaluates the ability of the detector-tracker pipeline to capture all ground-truth objects.

The results demonstrate a clear scaling trend. On detection-only metrics, YOLOv10n yields lower precision and recall than YOLOv10s/m, consistent with its smaller capacity. However, in the end-to-end tracking pipeline, YOLOv10n + BoostTrack attains higher IDF1 and fewer ID switches. We attribute this to lower inference latency improving frame-to-frame association, which offsets the modest deficit in raw detection accuracy in our drone scenarios.

Overall, larger models boost detection-only metrics, whereas YOLOv10n + BoostTrack provides stronger identity stability (higher IDF1, fewer ID switches) under real-time constraints.

4.2.2 MOTA and ID Switches

MOTA aggregates false positives, missed detections, and ID switches into a single score, while **ID switches (IDs)** quantify how often the tracker incorrectly reassigns an identity. Under our **real-time** constraints, the **YOLOv10n + BoostTrack** pipeline remains **competitive on MOTA** and yields **fewer ID switches**, reflecting more stable identity association. By contrast, the larger detectors (**YOLOv10s/m**) achieve **higher detection precision/recall**, but their increased latency and larger detection volume can **reduce MOTA** and **increase ID switches** due to more frequent ambiguous matches.

Taken together, these results highlight a practical **trade-off**: larger models improve **detection-only** metrics, whereas the lightweight **YOLOv10n** configuration delivers **stronger identity stability** (higher IDF1, fewer IDs) and competitive MOTA in real-time drone MOT. Model choice should therefore reflect deployment goals—maximizing detection recall versus prioritizing identity preservation under tight latency budgets.

4.3 Comparative Analysis Across Models

To provide a holistic understanding of the experimental outcomes, this section synthesizes the evaluation results of **YOLOv10n**, **YOLOv10s**, and **YOLOv10m** across multiple performance dimensions. The comparison encompasses **IDF1, Precision, Recall, MOTA, and ID Switches**, supplemented by an examination of **confidence score distributions**. By integrating results from detection-level logs, MOT evaluations, and confidence-based analysis, we aim to uncover the underlying trade-offs between model scale, detection fidelity, and tracking stability in the proposed BoostTrack integration framework.

4.3.1 Confidence Distribution Analysis

The analysis of confidence score distributions provides insights into how each YOLOv10 variant calibrates its detection certainty and how this impacts downstream tracking performance. **Figure 6** presents the confidence distributions for YOLOv10n/s/m (bin width 0.02). As shown in **Figure 6**, YOLOv10s/m exhibit heavier right tails (higher-confidence positives), whereas YOLOv10n concentrates more mass in the mid-confidence region.

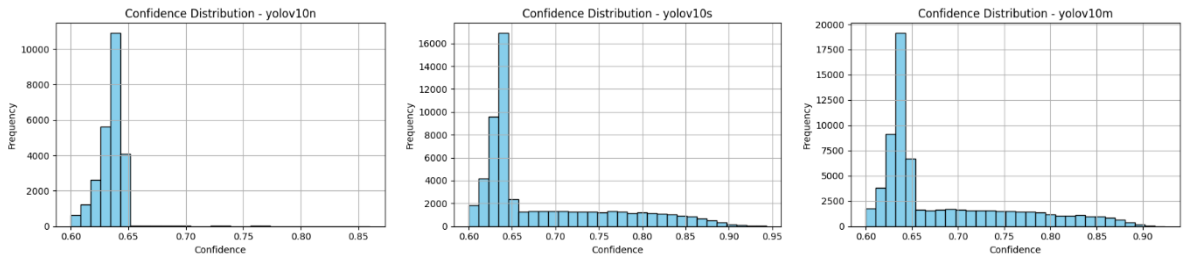

Fig 6. Confidence score distributions for YOLOv10n, YOLOv10s, and YOLOv10m.
The histograms illustrate how each model assigns confidence values across detections. YOLOv10n shows a narrow distribution concentrated around 0.63–0.65, reflecting highly consistent but conservative predictions. YOLOv10s and YOLOv10m demonstrate broader distributions extending up to 0.90–0.95, indicating a wider dynamic range

and greater willingness to assign high confidence scores. These differences highlight a trade-off between stability (YOLOv10n) and expressive certainty (YOLOv10s/m), with direct implications for balancing false positives and recall in multi-object tracking.

For **YOLOv10n**, the distribution is tightly concentrated around the 0.60–0.64 interval, with a mean confidence of 0.635 and a very low standard deviation of 0.013. This indicates that the model operates in a highly conservative mode, rarely producing overly high or low confidence values. While this consistency minimizes the risk of false positives, it also limits the model's capacity to express strong certainty for true detections, potentially constraining recall in challenging scenes.

In contrast, **YOLOv10s** exhibits a much wider spread, with a mean confidence of 0.685 and a standard deviation of 0.077. Its distribution extends up to 0.94, reflecting a model that is more willing to assign high confidence to detections it perceives as reliable. This behavior suggests that YOLOv10s is more aggressive in scoring, leading to stronger signals for true positives but at the cost of increased variability. Such variability can manifest as greater susceptibility to false alarms, particularly in cluttered or occluded frames.

Similarly, **YOLOv10m** achieves a mean confidence of 0.684 with a standard deviation of 0.074, producing a distribution closely resembling YOLOv10s. Its detections also extend toward higher confidence regions (up to 0.92), indicating robustness when the model is confident. However, the broader variance means that predictions can oscillate between moderately low and very high values, necessitating careful threshold tuning for optimal performance.

Overall, the comparative analysis underscores a key trade-off between conservatism and variability. YOLOv10n's narrow, threshold-proximal distribution yields stable but less expressive detections, whereas YOLOv10s and YOLOv10m demonstrate higher dynamic range and confidence strength at the expense of consistency. These differences have downstream implications for multi-object tracking: stable but lower-confidence detections favor identity preservation, while higher-confidence, variable outputs may enhance recall but risk frequent ID switches.

4.3.2 Tracking Performance Metrics

To evaluate the robustness of the proposed YOLOv10 + BoostTrack framework in drone-based multi-object tracking, we relied on widely adopted MOT (Multiple Object Tracking) performance metrics: **IDF1**, **MOTA**, **precision**, **recall**, and the number of **ID switches**. Each of these metrics highlights complementary aspects of tracking reliability, ranging from identity preservation to detection accuracy and consistency.

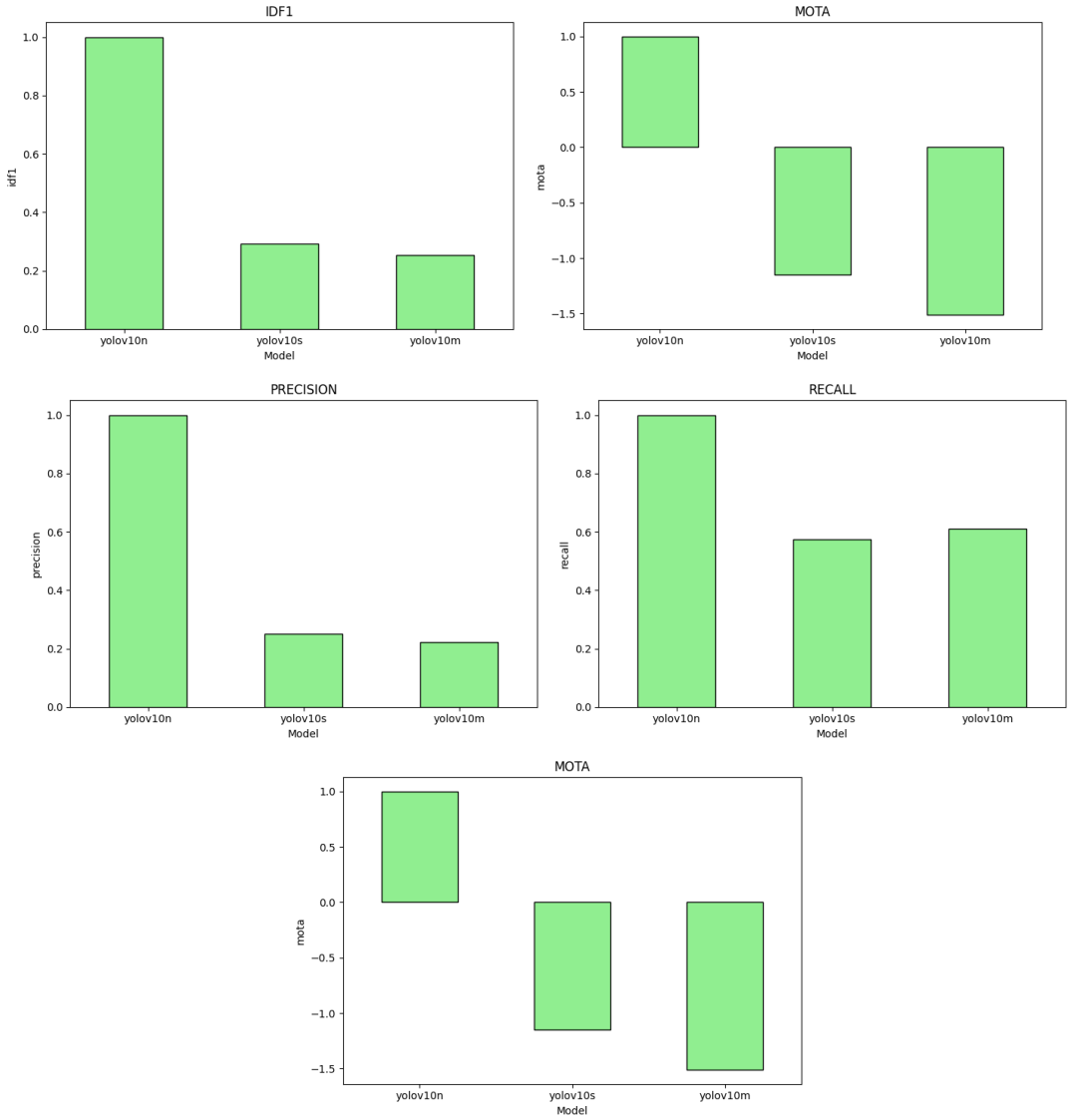

Fig 7. Comparison of tracking metrics (IDF1, MOTA, Precision, Recall, ID Switches) across YOLOv10n/s/m. YOLOv10n + BoostTrack shows higher IDF1 and fewer ID switches (identity stability), while YOLOv10s/m achieve higher detection precision/recall.

IDF1 represents the harmonic mean of identification precision and recall, indicating how well the tracker preserves consistent identities over time. IDF1 comparisons in Figure 7 indicate that YOLOv10n + BoostTrack preserves identities more consistently than YOLOv10s/m under our real-time setup, while the larger models retain advantages on pure detection precision/recall. MOTA, which aggregates false positives, false negatives, and ID switches into a single measure, shows a similar trend: the n-based pipeline remains competitive under our settings, whereas s/m tend to yield lower MOTA when constrained to real-time operation. This likely arises from latency-induced association errors that outweigh gains from higher raw detection counts. For detection metrics, the larger models (s/m) achieve higher precision and recall. The n-based pipeline compensates at the tracking stage: lower inference latency supports more stable frame-to-frame association, which benefits identity preservation despite modestly lower detection scores. Regarding ID switches, the n-based configuration records fewer switches than s/m under identical tracking settings, indicating stronger identity stability—especially in sequences with camera motion and brief occlusions. Overall, these results indicate a practical trade-off: larger detectors improve detection-only metrics, whereas YOLOv10n +

BoostTrack delivers more stable identities (higher IDF1, fewer ID switches) and competitive MOTA in real-time drone MOT. Model selection should therefore reflect deployment goals—maximizing detection recall versus prioritizing identity stability under tight latency budgets.

4.4 Discussion

The experimental results clarify the trade-offs between model size, detection accuracy, and tracking stability in drone-based MOT. Rather than a single model dominating every metric, the findings show **complementary strengths**: larger detectors (**YOLOv10s/m**) lead on **pure detection metrics** (precision/recall), while the lightweight **YOLOv10n**, when paired with **BoostTrack**, yields **stronger identity stability** (higher **IDF1**, fewer **ID switches**) and **competitive MOTA** under real-time constraints.

A plausible explanation is **latency**. Drone footage features rapid motion, scale changes, and short occlusions; under these conditions, **faster inference** improves frame-to-frame association and reduces identity fragmentation. The **n-based** pipeline benefits from this effect, whereas **s/m** produce more raw detections but can introduce association errors when inference is slower or detection volume is high.

Importantly, we **do not** claim "near-perfect" scores or universally "negative" MOTA for any model. Instead, **MOTA** trends are consistent with the above trade-off: the n-based configuration remains **competitive** at real-time speed, while s/m can experience **lower effective MOTA** when latency affects temporal consistency.

From an application perspective, **identity preservation** (IDF1 and ID switches) is often critical for surveillance, search-and-rescue, and traffic monitoring. In our setting, **YOLOv10n + BoostTrack** maintains identities more reliably, whereas **YOLOv10s/m** are preferable when maximizing detection recall is the primary objective.

Overall, these results suggest a **practical selection rule**: choose **s/m** when the deployment prioritizes **detection coverage**, and choose **n** when **real-time identity stability** is paramount. Future work should further optimize **speed-aware architectures**, camera-motion compensation, and motion-prediction modules to enhance robustness in agile drone scenarios.

5. Limitations

Our pseudo ground truth (PGT) was derived using a YOLOv10n-based configuration, which may introduce evaluation bias toward that detector. We mitigated this by comparing models under identical tracking settings and focusing on relative trends; nevertheless, future work should include manually annotated subsets or PGT generated by alternative detectors to validate robustness.

6. Conclusion and Recommendations

The experiments compared three YOLOv10 variants (n/s/m) within a BoostTrack-based pipeline using standard MOT metrics (IDF1, MOTA, precision, recall, and ID switches). The results show complementary strengths rather than one model dominating all metrics: the larger detectors (YOLOv10s/m) lead on pure detection metrics (precision/recall), while the lightweight YOLOv10n + BoostTrack configuration provides stronger identity stability (higher IDF1, fewer ID switches) and competitive MOTA under real-time constraints.

These findings indicate a practical trade-off driven in part by latency: faster inference improves frame-to-frame association and reduces identity fragmentation, whereas higher raw detection volume can increase ambiguous matches when latency is tighter.

In practice, we recommend choosing s/m when the deployment prioritizes detection coverage, and choosing n when real-time identity preservation is paramount. This guidance aligns with drone-surveillance scenarios where computational budgets and stability of identities often outweigh marginal gains in detection-only scores.

6.1 Future Works

Future work will focus on scaling the approach across diverse aerial conditions and hardware constraints while strengthening identity preservation under occlusion and rapid camera motion. While the current results

are promising, several avenues for future research remain open to further enhance the robustness and applicability of drone-based multi-object tracking systems. One direction involves exploring advanced optimization strategies such as pruning, quantization, or knowledge distillation to reduce computational costs while preserving or even improving accuracy. Another promising direction is the integration of YOLOv10n with more advanced tracking frameworks, such as DeepSORT or ByteTrack, in order to strengthen identity preservation and improve robustness under occlusion or crowded scenes.

Equally important is the evaluation of the framework across more diverse and challenging drone scenarios. Testing under varying altitudes, adverse weather conditions, or nighttime environments would provide stronger evidence of its generalizability to real-world applications. Beyond single-drone systems, extending this work toward coordinated multi-drone collaboration could significantly improve coverage, reduce blind spots, and enhance reliability in large-scale surveillance or search-and-rescue operations. Finally, deploying the YOLOv10n + BoostTrack framework on real drone hardware represents a crucial step toward validating its practicality, where factors such as power consumption, communication delays, and real-time decision-making constraints will play a decisive role.

By pursuing these directions, future work can further strengthen the reliability, efficiency, and scalability of lightweight tracking models, ensuring their suitability for real-world drone-based applications where both accuracy and stability are critical.

References

Abba, S., Bizi, A. M., Lee, J. A., Bakouri, S., & Crespo, M. L. (2024). Real-time object detection, tracking, and monitoring framework for security surveillance systems. *Heliyon*, *10*(15), e34922. https://doi.org/10.1016/j.heliyon.2024.e34922

Ahmad, T., Sciences, I., & States, U. (2025). *Future UAV / Drone Systems for Intelligent Active Surveillance and Monitoring*. https://doi.org/10.1145/3760389

Ali, M. L., & Zhang, Z. (2024). The YOLO Framework: A Comprehensive Review of Evolution, Applications, and Benchmarks in Object Detection. *Computers*, *13*(12). https://doi.org/10.3390/computers13120336

Alif, M. A. R., & Hussain, M. (2024). *YOLOv1 to YOLOv10: A comprehensive review of YOLO variants and their application in the agricultural domain*. 1–31. http://arxiv.org/abs/2406.10139

Apostolidis, K. D., & Papakostas, G. A. (2025). Delving into YOLO Object Detection Models: Insights into Adversarial Robustness. *Electronics (Switzerland)*, *14*(8). https://doi.org/10.3390/electronics14081624

Chiu, H.-K., Li, J., Ambruş, R., & Bohg, J. (2021). Probabilistic 3D Multi-Modal, Multi-Object Tracking for Autonomous Driving. *2021 IEEE International Conference on Robotics and Automation (ICRA)*, 14227–14233. https://doi.org/10.1109/ICRA48506.2021.9561754

Dadboud, F., Patel, V., Mehta, V., Bolic, M., & Mantegh, I. (2021). Single-Stage UAV Detection and Classification with YOLOV5: Mosaic Data Augmentation and PANet. *2021 17th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, 1–8. https://doi.org/10.1109/AVSS52988.2021.9663841

Du, C., Lin, C., Jin, R., Chai, B., Yao, Y., & Su, S. (2024). Exploring the State-of-the-Art in Multi-Object Tracking: A Comprehensive Survey, Evaluation, Challenges, and Future Directions. *Multimedia Tools and Applications*, *83*(29), 73151–73189. https://doi.org/10.1007/s11042-023-17983-2

Gad, A., Basmaji, T., Yaghi, M., Alheeh, H., Alkhedher, M., & Ghazal, M. (2022). Multiple Object Tracking in Robotic Applications: Trends and Challenges. *Applied Sciences (Switzerland)*, *12*(19). https://doi.org/10.3390/app12199408

Gao, R., Wang, Y., & Liu, C. (n.d.). *History-Aware Transformation of ReID Features for Multiple Object Tracking*.

Hassan, S., Mujtaba, G., Rajput, A., & Fatima, N. (2024). Multi-object tracking: a systematic literature review. *Multimedia Tools and Applications*, *83*(14), 43439–43492. https://doi.org/10.1007/s11042-023-17297-3

He, S., Luo, H., Wang, P., Wang, F., Li, H., & Jiang, W. (2021). TransReID: Transformer-Based Object Re-Identification. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 15013–15022.

Hussain, M., & Khanam, R. (2024). In-Depth Review of YOLOv1 to YOLOv10 Variants for Enhanced Photovoltaic Defect Detection. *Solar*, *4*(3), 351–386. https://doi.org/10.3390/solar4030016

Kamboj, A. (2024). *The Progression of Transformers from Language to Vision to MOT: A Literature Review on Multi-Object Tracking with Transformers*. https://arxiv.org/pdf/2406.16784

Karim, M., Khalid, S., Lee, S., Almutairi, S., Namoun, A., & Abohashrh, M. (2025). Next Generation Human Action Recognition: A Comprehensive Review of State-of-the-Art Signal Processing Techniques. *IEEE Access*, *13*,

135609–135633. https://doi.org/10.1109/ACCESS.2025.3590073

Kaseris, M., Kostavelis, I., & Malassiotis, S. (2024). A Comprehensive Survey on Deep Learning Methods in Human Activity Recognition. *Machine Learning and Knowledge Extraction*, *6*(2), 842–876. https://doi.org/10.3390/make6020040

Khatab, E., & Shalash, O. (2025). A Systematic Review: Computer Vision Algorithms in Drone Surveillance and Maritime Transport. *Journal of Robotics: Integration, Manufacturing & Control*, *2*(1). https://doi.org/10.21622/RIMC.2025.02.1.1149

Li, X., Li, P., Zhao, L., Liu, D., Gao, J., Wu, X., Wu, Y., & Cui, D. (2024). *RockTrack: A 3D Robust Multi-Camera-Ken Multi-Object Tracking Framework*. https://arxiv.org/abs/2409.11749v1

Lin, J., Chen, J., Peng, K., He, X., Li, Z., Stiefelhagen, R., & Yang, K. (2024). EchoTrack: Auditory Referring Multi-Object Tracking for Autonomous Driving. *IEEE Transactions on Intelligent Transportation Systems*, *25*(11), 18964–18977. https://doi.org/10.1109/TITS.2024.3437645

Liu, J., Zhang, C., & Li, J. (2024). Using Anchor-Free Object Detectors to Detect Surface Defects. *Processes*, *12*(12). https://doi.org/10.3390/pr12122817

Lusardi, C., Taufique, A. M. N., & Savakis, A. (2021). Robust Multi-Object Tracking Using Re-Identification Features and Graph Convolutional Networks. *Proceedings of the IEEE International Conference on Computer Vision*, *2021-Octob*, 3861–3870. https://doi.org/10.1109/ICCVW54120.2021.00433

Meneses, M., Matos, L., Prado, B., de Carvalho, A., & Macedo, H. (2020). *Learning to associate detections for real-time multiple object tracking*. http://arxiv.org/abs/2007.06041

Mirzaei, B., Nezamabadi-pour, H., Raoof, A., & Derakhshani, R. (2023). Small Object Detection and Tracking: A Comprehensive Review. *Sensors*, *23*(15). https://doi.org/10.3390/s23156887

Oliveira, H. S., Machado, J. J. M., & Tavares, J. M. R. S. (2021). Re-identification in urban scenarios: A review of tools and methods. *Applied Sciences (Switzerland)*, *11*(22). https://doi.org/10.3390/app112210809

Pal, O. K., Shovon, M. S. H., Mridha, M. F., & Shin, J. (2024). In-depth review of AI-enabled unmanned aerial vehicles: trends, vision, and challenges. *Discover Artificial Intelligence*, *4*(1). https://doi.org/10.1007/s44163-024-00209-1

Pathirannahalage, I., Jayasooriya, V., Samarabandu, J., & Subasinghe, A. (2025). A comprehensive analysis of real-time video anomaly detection methods for human and vehicular movement. *Multimedia Tools and Applications*, *84*(10), 7519–7564. https://doi.org/10.1007/s11042-024-19204-w

Rashidunnabi, M., Hambarde, K., & Proença, H. (2025). Causality and "In-the-Wild" Video-Based Person Re-Identification: A Survey. *Electronics (Switzerland)*, *14*(13), 1–31. https://doi.org/10.3390/electronics14132669

Saleh, F., Aliakbarian, S., Rezatofighi, H., Salzmann, M., & Gould, S. (2021). Probabilistic Tracklet Scoring and Inpainting for Multiple Object Tracking. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 14324–14334. https://doi.org/10.1109/CVPR46437.2021.01410

Sarker, P. K., Zhao, Q., & Uddin, M. K. (2024). Transformer-Based Person Re-Identification: A Comprehensive Review. *IEEE Transactions on Intelligent Vehicles*, *9*(7), 5222–5239. https://doi.org/10.1109/TIV.2024.3350669

Taufiqurrahman, T., Hadi, A. P., & Siregar, R. E. (2024). Evaluasi Performa Yolov8 Dalam Deteksi Objek Di Depan Kendaraan Dengan Variasi Kondisi Lingkungan. *Jurnal Minfo Polgan*, *13*(2), 1755–1773.

Terven, J., & Cordova-Esparza, D. (2023). *A Comprehensive Review of YOLO: From YOLOv1 and Beyond*. 1–34. http://arxiv.org/abs/2304.00501

Wang, A., Chen, H., Liu, L., Chen, K., Lin, Z., Han, J., & Ding, G. (2024). YOLOv10: Real-Time End-to-End Object Detection. *Advances in Neural Information Processing Systems*, *37*(NeurIPS), 1–21.

Wang, P., Wang, Y., & Li, D. (2024). DroneMOT: Drone-based Multi-Object Tracking Considering Detection Difficulties and Simultaneous Moving of Drones and Objects. *2024 IEEE International Conference on Robotics and Automation (ICRA)*, 7397–7404. https://doi.org/10.1109/ICRA57147.2024.10610941

Wang, Q., Zhou, L., Xu, C., Shang, Y., Jin, P., Cao, C., & Shen, T. (2025). Progress and Perspectives on UAV Visual Object Tracking. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 1–29. https://doi.org/10.1109/JSTARS.2025.3593286

Wang, X., Fu, C., He, J., Huang, M., Meng, T., Zhang, S., Zhou, H., Xu, Z., & Zhang, C. (2023). *You Only Need Two Detectors to Achieve Multi-Modal 3D Multi-Object Tracking*. http://arxiv.org/abs/2304.08709

Wang, X., Qi, S., Zhao, J., Zhou, H., Zhang, S., Wang, G., Tu, K., Guo, S., Zhao, J., Li, J., & Yang, M. (2024). *MCTrack: A Unified 3D Multi-Object Tracking Framework for Autonomous Driving*. http://arxiv.org/abs/2409.16149

Yuan, Y., Wu, Y., Zhao, L., Chen, H., & Zhang, Y. (2024). Multiple object detection and tracking from drone videos based on GM-YOLO and multi-tracker. *Image and Vision Computing*, *143*, 104951. https://doi.org/10.1016/j.imavis.2024.104951

Yuan, Y., Wu, Y., Zhao, L., Liu, Y., & Pang, Y. (2024). End-to-end multiple object tracking in high-resolution optical sensors of drones with transformer models. *Scientific Reports*, *14*(1), 1–16. https://doi.org/10.1038/s41598-024-75934-9