Academic Education Publishing House

Industry Science and Engineering Vol. 3 No. 3, 2026

A Multi-Object Tracking Algorithm for UAVs in Scenarios with Drastic Scale Variations and Dense Occlusions

Ziyang Qin[1], Zongshang Yang[1], Wanwan Wang[2], Jiangang Zhang[2]

[1]*School of Artificial Intelligence and Big Data, Henan University of Technology, Zhengzhou, Henan, China*

[2]*iFLYTEK Co., Ltd., Hefei, Anhui, China*

Abstract: UAV-based aerial imaging plays a crucial role in intelligent inspection applications. However, in real-world high-altitude scenarios, drastic scale variations and dense occlusions often result in missed detections, false positives, and identity switches in multi-object tracking systems. To address this issue, this paper proposes a lightweight UAV multi-object tracking method. A dynamic feature reconstruction module (DySample) is introduced at the detection stage to enhance small-object representation. To mitigate background interference caused by feature amplification, an efficient multi-scale attention mechanism (EMA) is incorporated. At the association stage, a confidence-driven adaptive Kalman filter combined with a dual-threshold matching strategy is employed to improve trajectory stability under occlusion. Ablation experiments on the VisDrone2019-MOT dataset show that, compared to the baseline model, the proposed method improves MOTA by 3.5 percentage points (from 25.3% to 28.8%), reduces ID switches from 38 to 35, and decreases false positives from 4728 to 3527, demonstrating advantages in suppressing false detections and maintaining identity consistency. However, the false negative (FN) rate increases (from 8440 to 9017), indicating room for improvement in recall under strict noise suppression. This study provides a feasible solution for UAV multi-object tracking under constrained computational resources.

Keywords: UAV Vision; Multi-Object Tracking; Dynamic Upsampling; Efficient Multi-Scale Attention; Adaptive Kalman Filtering.

1. Introduction

UAV-based aerial imaging, with its high mobility and wide field of view, plays a pivotal role in intelligent traffic management and large-scale security inspection, effectively compensating for the blind spots in fixed ground surveillance systems. However, in high-altitude perspectives, complex physical environments pose significant challenges to visual perception systems. On one hand, the fluctuations in UAV posture and altitude result in drastic scale variations, causing targets to occupy very small pixel areas. On the other hand, dense occlusions caused by crowded urban intersections or thick tree canopies place tremendous pressure on traditional visual tracking systems, making it easy for small targets to be missed and for trajectories to fragment frequently. These challenges have become the key bottleneck for the practical application of aerial intelligent systems.

In recent years, deep learning-based Tracking-by-Detection paradigms have gradually become mainstream. However, when conventional detectors (such as the standard YOLO architecture), which perform well in ground-level scenarios, are directly applied to high-altitude scenes, weak targets are easily lost during feature extraction due to the inherent downsampling mechanism in deep networks. Although conventional upsampling layers can partially recover some spatial information, they not only exacerbate the computational burden on UAV edge devices but also amplify background noise, such as road markings and shadows. As a result, the model is highly susceptible to irrelevant noise, leading to a sharp increase in false positives and difficulty in accurately focusing on fine details of real targets.

The loss of information at the front-end feature extraction further exacerbates the difficulty in data association at the back-end for multi-object tracking. In complex urban environments, targets are densely distributed and subject to unpredictable nonlinear movements and

long-term occlusions. Faced with continuous trajectory gaps, mainstream trackers often struggle to balance noise resistance and trajectory continuity. Whether relying on basic SORT algorithms based on pure linear motion prediction or complex deep feature matching networks, these trackers often suffer from severe frame drops due to computational bottlenecks, or experience frequent identity switches (IDSW) due to the accumulation of prediction errors, resulting in the repeated assignment of new IDs to the same target.

How to precisely lock onto weak targets at high altitudes within the constraints of limited computational power, while maintaining robust long-term trajectory continuity, remains a technical gap in current perception technologies. To address this, this paper proposes a lightweight tracking framework that integrates dynamic feature reconstruction and adaptive association strategies. First, a dynamic feature reconstruction module (DySample) with content-awareness is introduced into the feature pyramid to effectively recover the feature loss caused by drastic scaling. Second, an efficient multi-scale attention mechanism (EMA) is introduced to effectively suppress the background noise introduced during feature amplification through cross-spatial interactions. Finally, during the data association phase, adaptive Kalman filtering and a dual-threshold strategy are employed to dynamically penalize the error accumulation caused by nonlinear motion. The proposed method significantly enhances the ability of the base tracker to re-anchor targets in occluded environments, providing a practical and efficient solution for high-precision continuous tracking in complex high-altitude scenarios.

2. Related Work

The breakthrough progress of object detection technology in the field of unmanned aerial vehicle (UAV) vision is primarily attributed to the continuous evolution of deep learning network architectures. Early explorations mostly relied on traditional image processing or computationally redundant two-stage algorithms. Subsequently, the YOLO series algorithms proposed by Redmon et al. [1], relying on a single-stage end-to-end prediction mechanism, achieved an excellent trade-off between detection accuracy and inference speed, establishing their mainstream position in industrial-level deployments. In recent years, addressing the computing power bottleneck of UAV edge computing platforms, Hu et al. [2] proposed a lightweight aerial object detection algorithm based on YOLO, further promoting the deployment of models on mobile terminals. Pu et al. [3] systematically reviewed the application of deep learning in UAV vision, pointing out that limited computing power and the easy loss of small object features are the key factors restricting the practical implementation of these algorithms. To solve the problem of extremely small target pixel proportions from the perspective of UAV aerial photography, Ju et al. [4] realized effective recognition of small objects by optimizing the feature pyramid. Li et al. [5] propsed a dense small object detection algorithm for aerial photography based on a synergistic attention mechanism, effectively mitigating the degradation of target features caused by complex background clutter. However, the inherent mechanical mapping rules of conventional convolutional upsampling operations limit their ability to capture fine-grained context. To break this deadlock, Wang et al. [6] proposed the CARAFE dynamic upsampling operator, attempting to use content awareness to reconstruct features. Nevertheless, when processing extremely small and dense traffic targets from a UAV perspective, its massive dynamic kernel generation process often leads to severe computational redundancy and inference delays.

To overcome the limitations of continuous tracking in dynamic and complex scenes, the Tracking-by-Detection (TBD) paradigm has gradually become mainstream. Building upon the traditional SORT algorithm, Wojke et al. [7] introduced deep appearance features to construct the DeepSORT algorithm, substantially improving the re-identification capability of targets after brief occlusions. Luo et al. [8] proposed a multi-object tracking algorithm based on the joint optimization of YOLO and DeepSORT, strengthening the perception of minute appearance features through a spatial-temporal attention network. However, traditional algorithms often directly discard low-confidence detection boxes, resulting in severe trajectory interruptions. The ByteTrack algorithm proposed by Zhang et al. [9] innovatively incorporated low-score boxes into a secondary matching system, significantly alleviating missed detections caused by

occlusions. Subsequently, the OC-SORT algorithm proposed by Cao et al. [10] effectively dealt with trajectory deviations caused by non-linear motion through an observation-centric momentum update strategy. Meanwhile, the BoTSORT framework, widely adopted as an industry baseline, further elegantly integrated Camera Motion Compensation (CMC) and the Kalman filter, demonstrating outstanding performance in handling severe UAV camera movements and continuously setting new records for tracking accuracy.

Although the aforementioned studies have made remarkable progress in their respective dimensions, existing solutions still exhibit obvious missing pieces regarding the specific proposition of "high-precision multi-object tracking in complex aerial scenes." Most studies either focus on the lightweight design of the front-end detector, sacrificing the recall rate of weak signals, or they focus on stacking heavy Re-Identification (ReID) networks to maintain trajectories, ignoring the reality of restricted computing power on UAV edge devices. Furthermore, when existing models face real-world complex scenes like VisDrone, once a target enters the shade of trees causing a sharp drop in confidence, the traditional Kalman filter often fits low-quality observation boxes, triggering catastrophic trajectory crossovers and identity switches. Therefore, constructing a dedicated lightweight architecture integrating dynamic feature reconstruction (DySample), multi-scale attention purification (EMA), and confidence-adaptive Kalman association—balancing high recall and anti-occlusion capabilities without blindly surging computational costs—has become the key to filling this academic gap.

3.Experimental Principles and Methods:

3.1 BoTSORT Algorithm Architecture

BoTSORT (Bag of Tricks for SORT) is a state-of-the-art multi-object tracking model, which is built upon the classic Tracking-by-Detection paradigm. Unlike basic trackers that solely rely on linear kinematic prediction, this algorithm overcomes the limitations of relying solely on appearance feature matching by deeply integrating camera motion compensation, hierarchical data association, and Kalman filtering. In this paper, we adopt BoTSORT as the underlying association framework, directly utilizing its dual-threshold matching logic and motion correction mechanism. This feature is particularly crucial for handling complex image sequence tracking tasks that involve frequent occlusions and nonlinear background shifts.

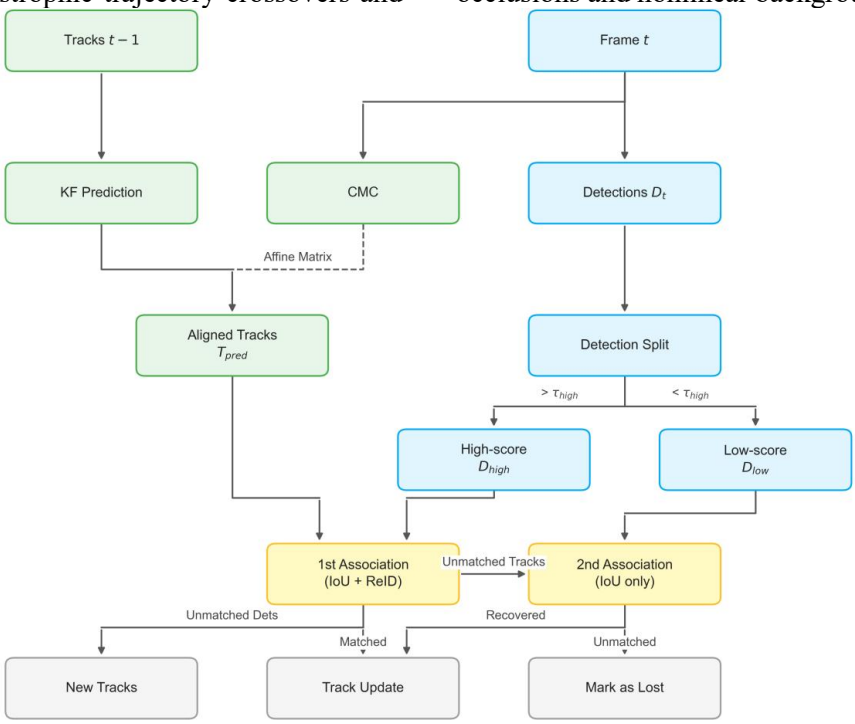

Figure 1. BoTSORT Network Architecture Diagram

As shown in Figure 1, the overall data flow of BoTSORT includes four key stages: feature detection, motion state prediction, cascade matching, and lifecycle management. In the

current video frame, the pre-processing target detector first completes feature extraction, and the resulting detection boxes are rigorously divided into high-confidence and low-confidence subsets using a confidence-based dual-threshold splitter. Meanwhile, the prediction of historical trajectories is also underway: the trajectory from the previous frame not only undergoes basic position prediction via Kalman filtering (KF) but also receives a global affine matrix extracted by the camera motion compensation (CMC) module. Using this matrix, the system performs coordinate alignment and generates the final predicted trajectory (Tpred), which strips away lens displacement errors.

Once the data association phase is reached, BoTSORT employs a meticulous downgrade matching strategy. In the first matching round, the system combines the Intersection over Union (IoU) and appearance-based re-identification (ReID) features to precisely match the predicted trajectories with high-confidence boxes. The successfully matched trajectories are immediately updated, while unmatched high-confidence boxes are initialized as new trajectories. For the predicted trajectories that were missed in the first round, the system does not discard them directly but instead feeds them together with the downgraded low-confidence boxes into the second round of matching. In this stage, the interference from appearance features is removed, and the system solely relies on IoU distances for spatial localization. This effectively rescues real targets whose scores dropped due to brief occlusions. Trajectories that remain unmatched after two rounds of filtering are eventually marked as lost, thus completing the tracking cycle.

3.2 DySample Dynamic Feature Reconstruction Module

In modern object detection networks with Feature Pyramid Networks (FPN), upsampling operations are essential for achieving multi-scale semantic fusion. Traditional feature amplification methods, such as nearest neighbor interpolation or bilinear interpolation, typically rely on fixed spatial geometric rules to estimate pixel values. This lack of contextual awareness in such mechanical operations often leads to irreversible smoothing of fine-grained textures, especially in complex images with rich details. While transposed convolutions can mitigate this issue, they inevitably introduce a high computational burden. Early dynamic upsampling operators, such as CARAFE, struggled with the heavy dynamic kernel generation process, making them difficult to deploy in lightweight networks. To strike the optimal balance between feature reconstruction quality and computational complexity, this paper introduces the DySample (Dynamic Sampling) dynamic upsampling module [11] during the feature fusion stage of the network.

As shown in Figure 2, DySample's design innovatively reduces the upsampling problem to a purely spatial point sampling process. Facing low-resolution input feature maps, the module discards complex feature aggregation calculations and directly uses a lightweight linear mapping layer to capture local contextual semantics, generating spatial offsets. These offsets, which carry structure-guiding information of the image, are then added to the standard regular sampling grid, causing the originally rigid sampling points to elastically deform and accurately focus on the edges and high-frequency texture regions of the targets. Subsequently, the system utilizes the efficient grid sample function to resample the input features using these dynamically deformed coordinates, thereby smoothly outputting high-quality feature maps with enhanced resolution and significantly improved semantic details.

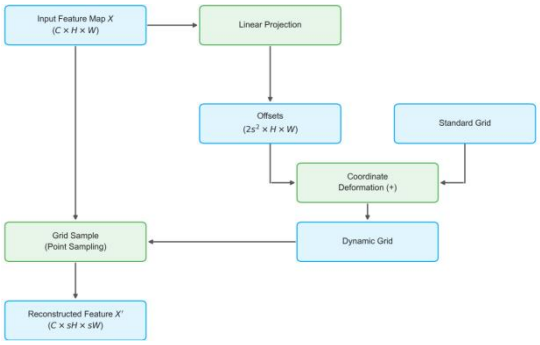

Figure 2. DySample Dynamic Feature Reconstruction Module Architecture Diagram

Thanks to this ingenious pure coordinate deformation mechanism, DySample avoids the additional resource overhead caused by the generation of large-scale dynamic convolution kernels. It not only outperforms existing learnable upsampling operators in terms of parameter scale and inference latency, but also provides the network with a 'self-adaptive focus' perspective. The introduction of this module opens up an upward channel for fine-grained information from the source of physical feature extraction, offering rich underlying detail support for subsequent dense prediction tasks.

3.3 EMA Multi-Scale Attention Mechanism

While the preceding feature reconstruction significantly enhances the recall performance of extremely small-scale targets, it is undeniably a double-edged sword. In practical experiments, we observed that as weak signals are dynamically amplified, highly deceptive background clutter—such as zebra crossings or tree canopies—is also inadvertently incorporated into the deep network. If left unchecked, the resulting surge in false positives (FPs) can severely degrade the downstream tracking pipeline. To improve the system's robustness against background noise, this work introduces an attention mechanism. However, the conventional SE module, with its global pooling, ruthlessly destroys highly valuable spatial positional information, while the CA module tends to be somewhat sluggish during cross-dimensional interactions. To address these engineering challenges, we embed an efficient multi-scale attention mechanism (EMA) [12] deep within the backbone network.

As illustrated in Figure 3, the architecture cleverly splits the input feature XXX along the channel dimension into several parallel subgroups, physically decoupling overly redundant features. Within each subgroup, the data flow is divided into three branches: two parallel 1×11 \times 11×1 convolutional branches perform one-dimensional global encoding along the X and Y axes, precisely capturing long-range structural dependencies; an independent 3×33 \times 33×3 convolutional branch functions as a "microscope," effectively preserving local multi-scale texture details. Subsequently, the architecture normalizes the parallel features via Softmax, performs cross-dimensional matrix multiplication on the resulting 2D attention weights, and applies a Sigmoid activation to produce the final aggregated weights, which are then precisely applied to the original feature map. This mechanism suppresses the propagation of background noise in dense scenes without compromising the lightweight nature of the network, enabling the visual system to maintain stable target perception even when encountering partially occluded and minute objects.

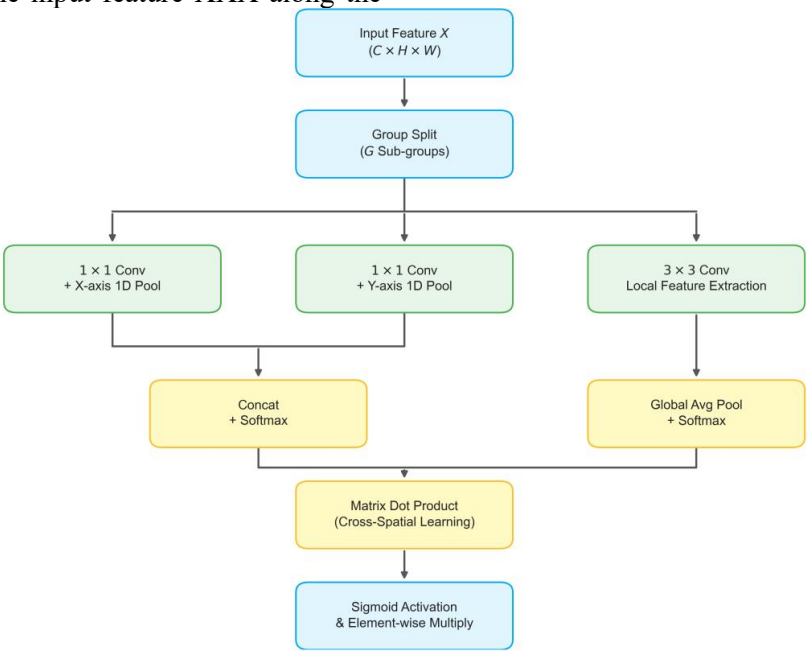

Figure 3. EMA Multi-Scale Attention Mechanism Diagram

3.4 Adaptive Association and Trajectory Management Strategy

In the context of UAV aerial perspectives, characterized by severe camera shaking and frequent occlusion by tree canopies, significantly enlarging the matching search radius can easily lead to catastrophic trajectory switches (ID Switches). Introducing excessively long historical feature sequences would further increase memory consumption and computational overhead on edge devices. To address this, we incorporate a confidence-driven Adaptive Kalman Filter [13].

The upgrade from the standard Kalman filter to its adaptive version operates as follows: when the bounding box confidence provided by the detector is very high, it indicates that the target is

clearly visible and features are rich, so the system should "trust" the current observation more. Conversely, when the target moves under a tree canopy and the confidence score drops sharply, the system should promptly reduce trust in the corrupted observation and rely instead on the Kalman filter's inherent kinematic prediction. Specifically, for a detection at frame ttt, we dynamically reconstruct its observation noise covariance matrix RtR_tRt as:

$$R_t = R_{base} \cdot \exp(1-C_t) \quad (1)$$

where RbaseR_{\text{base}}Rbase denotes the standard observation noise matrix of the Kalman filter under ideal conditions, and Ct∈(0,1]C_t \in (0,1]Ct∈(0,1] represents the real-time confidence score output by the frontend detector for that target. Through this nonlinear adaptive penalty, when the target undergoes severe occlusion or deformation (i.e., CtC_tCt approaches zero), the corresponding observation noise RtR_tRt is exponentially amplified.

This adaptively enlarged noise matrix is then directly fed into the core Kalman gain update equation:

$$K_t = P_{t|t-1}H^T(HP_{t|t-1}H^T+R_t)^{-1} \quad (2)$$

where Pt∣t−1P_{t|t-1}Pt∣t−1 is the error covariance of the prior estimate, and HHH is the state-space mapping matrix. In practical matrix inversion and state update operations, the rapidly increasing observation noise RtR_tRt in the denominator forces the Kalman gain KtK_tKt to decay quickly toward zero. This mechanism compels the tracker to actively disregard unreliable observations that could otherwise bias the trajectory when faced with low-quality or difficult samples, and instead rely on pure kinematic inertia to maintain smooth trajectory propagation.

This approach effectively mitigates nonlinear prediction drift caused by poor detection boxes. Without introducing additional network parameters, it enables the system to maintain robust frame continuity and ID persistence even under the long-term occlusions and deformations characteristic of aerial scenarios.

3.5 Overall Network Architecture

When processing raw aerial videos characterized by severe camera shake, drastic scale variations from high altitudes, and dense occlusion by tree canopies, conventional "detection–tracking" pipelines often fail at the very first stage of feature extraction. To address this, we have thoroughly redesigned the low-level perception and association logic of the baseline model, proposing a lightweight technical pathway that spans "multi-scale purification – dynamic reconstruction – adaptive tracking." As illustrated in Figure 4, the entire system is tightly integrated, consisting of a frontend visual perception network coupled with a backend association and decision-making framework.

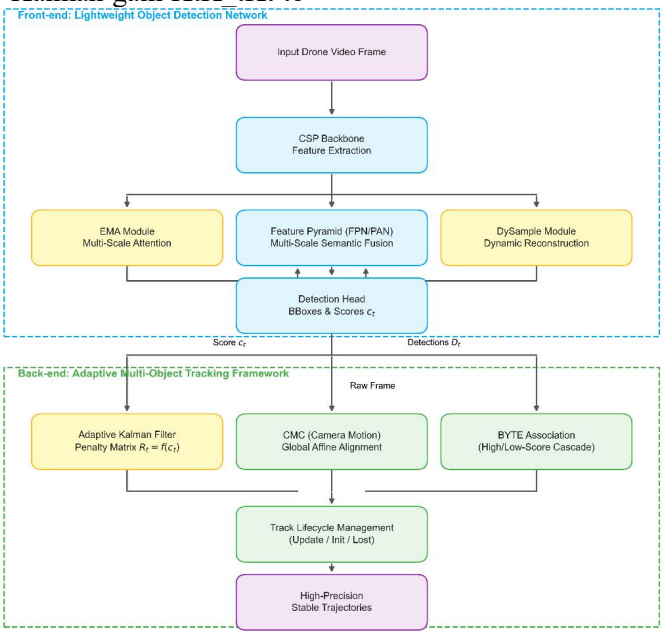

Figure 4. Overall Network Architecture Diagram

4. Experimental Process and Result Analysis

4.1 Experimental Dataset and Evaluation Metrics

This experiment uses the VisDrone2019-MOT

aerial multi-object tracking dataset as the benchmark for model training and evaluation. The dataset, collected and publicly released by the Machine Learning and Data Mining Laboratory (AISKYEYE) at Tianjin University, is one of the most representative drone vision datasets available today. The data used in this study was obtained from its official open-source channels (official website and GitHub repository), and we strictly adhered to the dataset's open-source usage agreement, which is intended for non-commercial academic research.

The VisDrone2019-MOT dataset contains videos captured by drones under various conditions, including different cities, weather conditions (sunny, cloudy, etc.), and lighting conditions (daytime, nighttime). The dataset includes top-down and oblique views. It provides over 2.5 million manually annotated bounding boxes across 10 predefined categories (e.g., pedestrians, cars, bicycles). Due to the variation in drone flight altitude, the dataset contains numerous extremely small-scale targets, dense crowds, and severe occlusions, accompanied by strong camera motion, posing high robustness challenges for tracking algorithms.

To comprehensively and objectively evaluate the tracking performance of the model, this study adopts the industry-standard CLEAR MOT evaluation metric system and identity metrics, which include the following core parameters:

1. Multi-Object Tracking Accuracy (MOTA): MOTA is the core metric for evaluating the overall performance of multi-object tracking algorithms. It takes into account false positives (FP), false negatives (FN), and identity switches (IDSW). A higher MOTA value indicates better overall tracking accuracy of the model.
2. ID F1 Score (IDF1): This metric measures the ability of the model to maintain consistent identity across the entire video sequence.
3. Identity Switches (IDSW): IDSW refers to the total number of times the same real target is incorrectly assigned different IDs during tracking. In aerial scenes with heavy occlusions, such as when a target reappears after being blocked by trees, IDSW is more likely to occur. A lower IDSW value indicates better trajectory stability.

4.2 Training Strategy

For the training strategy, all input images were uniformly resized to 640 × 640 pixels. Mosaic and MixUp data augmentation techniques were applied to improve the model's generalization capability on small-scale targets. The model was trained using the SGD optimizer with a momentum of 0.937 and a weight decay of 5×10^{-4}. The total training schedule was set to 100 epochs, with an initial learning rate of 0.01, which was decayed following a cosine annealing schedule. To fully utilize the 24 GB of physical GPU memory, the batch size was set to 32, and a warm-up mechanism was applied during the first three epochs to stabilize the initial gradients.

4.3 Ablation Study

To investigate the individual contributions of the proposed Adaptive Kalman Filter (AKF), DySample dynamic feature reconstruction, and EMA multi-scale attention mechanisms within the overall tracking pipeline, a rigorous ablation study was conducted on the VisDrone2019-MOT dataset. All experimental groups maintained identical hyperparameters and training environments. The design of the experiments and the corresponding results are presented in Table 1 and Table 2.

Table 1. Ablation Study Module Configuration

Experimental Group	Object Detector(YOLO)	Multi-Object Tracker (Tracker)	Innovation
Model 0	Official YOLOv8n	Official BoTSORTT	None
Model 1	Official YOLOv8n	BoTSORT_Adaptive	Adaptive Kalman Filter introduced
Model 2	YOLOv8n+DySample	BoTSORT_Adaptive	Dynamic upsampling introduced
Model 3(ours)	YOLOv8n+DySample+EMA	BoTSORT_Adaptive	Multi-scale attention introduced

Table 2. Ablation Study Performance Comparison

Experimental Group	MOTA (Accuracy)	IDF1 (Identity Consistency)	IDSW (Identity Switches)	FP (False Positives)	FN (False Negatives)
Model 0	25.3%	53.7%	38	4728	8440
Model 1	27.8%	55.0%	43	4281	8430
Model 2	27.7%	55.7%	37	4421	8325
Model 3(ours)	28.8%	55.6%	35	3527	9017

From the baseline model perspective, the native YOLOv8n paired with the official BoTSORT achieved a MOTA of 25.3% and an IDF1 of 53.7%. When the backend incorporated the Adaptive Kalman Filter (Model 1), all core metrics experienced a significant improvement: MOTA increased substantially to 27.8%, and IDF1 rose to 55.0%. This result partially overcomes the traditional tracker's blind reliance on current observations, demonstrating that dynamically adjusting the noise penalty matrix based on confidence is superior to rigid linear prediction when handling frequent occlusions in aerial views. It effectively prevents the model from overfitting low-quality detections and significantly reduces false positives (FP decreased from 4728 to 4281).

As summarized in Table 2, Model 3 achieved the best overall MOTA of 28.8%, while reducing IDSW to 35 and FP to 3527. This indicates that the synergistic strategy combining EMA and AKF effectively enhances association stability and robustness against interference in complex scenarios. Compared with the baseline, the system shows consistent gains in identity continuity and false positive control.

It should be noted that the FN for Model 3 is 9017, higher than that of Model 2 (8325) and the baseline (8440). This suggests that under the current configuration, EMA's noise suppression strategy, while reducing false positives, may also attenuate responses from certain weak targets, resulting in lower recall. Overall, the proposed method reflects a performance trade-off of "reducing false positives and stabilizing identity at the expense of partial recall loss." Future work will focus on optimizing attention intensity and threshold strategies to better balance this trade-off.

4.4 Visualization Analysis

To intuitively evaluate the practical performance of the proposed method in complex aerial scenarios, this section selects representative challenging clips from the VisDrone2019-MOT test set, containing extremely small targets, rapidly changing viewpoints, and occlusions. A visual comparison was conducted between the baseline model (YOLOv8n + official BoTSORT) and the improved model, as shown in Figure 5.

Figure 5. Visual Comparison Results between the Baseline Model and the Improved Model

The red-marked regions in Figure 5 primarily reflect the detection and tracking performance of distant small targets. In areas such as the road edges in Scenario 1 and the center of the intersection in Scenario 2, the targets occupy a small proportion of pixels and exhibit textures similar to the background, making detection particularly difficult. The comparison shows that the baseline model suffers from missed detections in these regions, whereas the improved model achieves more stable responses for certain small targets, successfully identifying and continuously tracking pedestrians and non-motorized vehicles that the baseline failed to detect. This observation indicates that the combination of DySample and EMA provides

tangible benefits for fine-grained target representation in complex backgrounds. It should be noted that these visualization results illustrate local improvements and are not contradictory to the overall increase in FN observed in Table 2, which reflects the global recall trade-off.

The blue zoom-in regions in Figure 5 highlight identity preservation under occlusion. In Scenario 1, a small white car is assigned different trajectory IDs by the two models (ID 12 for the baseline, ID 8 for the improved model). When the target is subsequently occluded by a billboard and trees and then reappears, the baseline model experiences a trajectory interruption and reassigns a new ID (ID 256), whereas the improved model maintains the original ID 8. This demonstrates that the confidence-driven adaptive Kalman filter used in this work reduces the impact of low-quality observations on trajectory updates, thereby improving identity consistency after occlusion.

In summary, the proposed method shows improvements in both the detection stability of small targets and post-occlusion identity preservation in complex aerial scenarios, meeting the practical requirements for UAV-based multi-object tracking. However, quantitative results indicate that a trade-off still exists between false positive suppression and recall performance, suggesting that further optimization is needed in future work.

5. Conclusion

This paper addresses the degradation of multi-object tracking performance caused by significant scale variations and dense occlusions in UAV aerial scenarios. A lightweight tracking framework is proposed, integrating DySample dynamic upsampling, EMA multi-scale attention, and Adaptive Kalman Filtering (AKF), and validated on the VisDrone2019-MOT dataset. Experimental results show that, compared to the YOLOv8n and BoTSORT baseline models, the proposed method improves Multi-Object Tracking Accuracy (MOTA) by 3.5 percentage points to 28.8%, reduces the number of Identity Switches (IDSW) from 38 to 35, and significantly suppresses False Positives (FP) from 4728 to 3527. These results demonstrate that the proposed framework offers superior robustness and identity preservation in complex scenarios, effectively overcoming the engineering dilemma in aerial tracking where "high recall inevitably leads to high false positives and occlusion causes identity loss."

However, the experiment also reveals that the number of False Negatives (FN) increased from 8440 to 9017, indicating that the stringent noise suppression mechanism of EMA, while reducing false positives, has a negative impact on the recall of extremely small targets due to over-suppression of weak true signals. This suggests that the current method still faces a performance trade-off between false positive control and target recall. Furthermore, the framework does not yet incorporate a Re-Identification (ReID) network, limiting the recognition capability of similar targets under long-term occlusion. Additionally, since the validation was conducted on a single dataset, the cross-scene generalization performance has yet to be fully evaluated.

Future work will focus on exploring collaborative optimization strategies for EMA strength and detection thresholds to alleviate the issue of weak target misdetection, integrating a lightweight ReID branch to enhance appearance feature matching, and introducing temporal attention mechanisms to improve non-linear trajectory prediction accuracy. The ultimate goal is to conduct systematic validation in more diverse datasets and real-world UAV deployment environments, advancing the research towards widespread applications in intelligent visual inspection.

References

[1] Redmon J, Divvala S, Girshick R, et al. You only look once: Unified, real-time object detection //Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR). Las Vegas, NV, USA: IEEE, 2016: 779-788.

[2] Hu, X., Pan, S. F. "A UAV Aerial Object Detection Algorithm Based on Improved Lightweight YOLOX." Computer Measurement and Control, 2024, 32(01): 57-63.

[3] Pu, L., Zhang, X. J. "UAV Visual Object Detection and Tracking Based on Deep Learning." Journal of Beijing University of Aeronautics and Astronautics, 2022, 48(05): 872-880.

[4] Ju, M. R., Luo, H. B., Wang, Z. B., et al. "Improved YOLO V3 Algorithm and Its Application in Small Object Detection." Acta Optica Sinica, 2019, 39(07): 253-260.

[5] Li, Z. H., Wang, Z. P., He, Y. T. "Aerial Small Object Detection Algorithm Based on Adaptive Collaborative Attention Mechanism." Acta Aeronautica et al Astronautica Sinica, 2023, 44(13): 244-254.

[6] Wang J, Chen K, Xu R, et al. CARAFE: Content-Aware ReAssembly of Features //Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). Seoul, Korea (South): IEEE, 2019: 3007-3016.

[7] Wojke, N., Bewley, A., Paulus, D. "Simple Online and Realtime Tracking with a Deep Association Metric." 2017 IEEE International Conference on Image Processing (ICIP). Beijing: IEEE, 2017: 3645-3649.

[8] Luo, Q., Zhao, R., Zhuang, H. S., et al. "YOLOv5 and Deep-SORT Joint Optimization for UAV Multi-Object Tracking." Signal Processing, 2022, 38(12): 2628-2638.

[9] Zhang, Y., Sun, P., Jiang, Y., et al. "ByteTrack Multi-Object Tracking by Associating Every Detection Box." European Conference on Computer Vision. Cham: Springer Nature Switzerland, 2022: 1-21.

[10] Cao J, Pang J, Weng X, et al. Observation-centric sort: Rethinking sort for robust multi-object tracking //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Vancouver, BC, Canada: IEEE, 2023: 9686-9696.

[11] Lin, Z., Yun, B., Zheng, Y. "LD-YOLO: A Lightweight Dynamic Forest Fire and Smoke Detection Model with DySample and Spatial Context Awareness Module." Forests, 2024, 15(9): 1630-1630.

[12] Sun, P., Xie, H., Lu, Q. Z., et al. "A YOLO11 Asphalt Pavement Crack Detection Algorithm Combined with Hyperbolic Enhancement and EMA." China Testing, 2025, 51(S2): 158-164.

[13] Lv, J., Ran, J. "Bagged Grape Video Counting Method Based on Improved YOLOv9s and Adaptive Kalman Filtering." Transactions of the Chinese Society of Agricultural Engineering, 2025, 41(10): 195-203.