



ARTICLE

A New Confidence-Aware Adaptive Multi-Object Tracking for Unmanned Aerial Vehicles

Long Wen^{1,2,#}, Yutao Chen^{1,#}, Yuxin Liu³, Hui Tan³, Jiangbo He² and Zhe Cheng^{4,*}

¹School of Power Engineering, Naval University of Engineering, Wuhan, China

²School of Mechanical Engineering and Electronic Information, China University of Geosciences, Wuhan, China

³National Key Laboratory of Electromagnetic Effect and Security on Marine Equipment, China Ship Development and Design Center (CSDDC), Wuhan, China

⁴Key Laboratory of Equipment State Sensing and Smart Support, National University of Defense Technology, Changsha, China

*Corresponding Author: Zhe Cheng. Email: chengzhe@nudt.edu.cn

#These authors contributed equally to this work as the first author.

Received: 18 July 2026; Accepted: 10 September 2026; Published: 21 October 2026

ABSTRACT: Multi-object tracking (MOT) for unmanned aerial vehicles (UAVs) remains challenging due to rapid camera motion, frequent occlusion, scale variation, and unreliable detections, all of which degrade the robustness of data association. Existing detection-based tracking methods generally rely on fixed fusion strategies while lacking explicit reliability modeling for detector observations and motion predictions, limiting their adaptability in dynamic UAV environments. To address this issue, this paper proposes a confidence-aware adaptive data association model for UAV multi-object tracking. The proposed model explicitly characterizes the reliability of detector observations and motion predictions within a unified association framework. A hierarchical confidence model is first developed to progressively associate detections according to their confidence levels. Subsequently, a prediction confidence model is established by exploiting the inverse covariance of the Kalman filter to quantify motion uncertainty. Finally, an adaptive association cost model is formulated by jointly incorporating detection and prediction confidence, enabling dynamic fusion of motion and appearance cues during data association. Experiments on the VisDrone2019 and UAVDT benchmarks validate the effectiveness of the proposed model. The proposed approach achieves 41.439 Multiple Object Tracking Accuracy (MOTA) and 44.001 Higher Order Tracking Accuracy (HOTA) on VisDrone2019, together with 49.000 MOTA and 52.03 HOTA on UAVDT. The results demonstrate that the proposed model provides stable and competitive tracking performance in challenging UAV scenarios.

KEYWORDS: Multi-object tracking; unmanned aerial vehicle; reliability modeling; confidence-aware association; adaptive data association; Kalman filtering

1 Introduction

Multi-Object Tracking plays an essential role in target tracking, navigation, and situational awareness for unmanned aerial vehicles [1,2]. Its importance becomes even more prominent in complex and high-precision scenarios such as dense urban environments, low-altitude high-speed inspections, and target-following missions [3,4].

At present, detection-based tracking has become the mainstream approach. In this paradigm, object detectors are used to locate targets in each video frame, and data association is adopted to link these detections across time [5,6]. To improve the reliability of association, many recent methods combine

appearance information with motion predictions produced by filtering algorithms [7,8]. Although these strategies have achieved considerable progress, they still face clear limitations in real UAV applications. One limitation lies in the separation of detection and association modules, which prevents the full use of the valuable information produced by detectors. For example, detection confidence reflects not only category reliability but also the clarity, distance, and occlusion status of the target. High-confidence detections usually correspond to clear and visible targets, while low-confidence detections often come from occluded, blurred, or distant objects [9]. These signals are extremely important in UAV scenarios that involve low altitudes, dense targets, or frequent occlusion, yet they are often ignored. Another limitation is that most existing methods use fixed weights to combine appearance and motion information [7]. Such static fusion cannot adapt to the changing reliability of different cues. In UAV scenarios, rapid camera motion, severe occlusion, and motion blur can degrade detection quality and undermine the reliability of motion-based association [10]. When fixed weights continue to be used under these conditions, association errors tend to accumulate quickly.

To overcome these issues, this paper proposes a multi-level hierarchical dynamic association strategy. First, a confidence-guided hierarchical matching mechanism is designed, in which detections are grouped according to their confidence, and high-confidence detections are matched with priority. Second, a prediction-confidence evaluation module is introduced to measure the reliability of the motion information produced by the Kalman filter, ensuring that unreliable predictions do not dominate the association process. Finally, a dynamic weighting strategy is developed to fuse appearance and motion cues. By simultaneously considering detection confidence and prediction confidence, the system adjusts the relative influence of the two information sources in real time. When one source becomes unreliable due to motion drift or visual degradation, its weight is reduced, while the weight of the more reliable source is increased. Through these designs, the proposed method significantly enhances the adaptability and robustness of data association in the highly dynamic and uncertain environments encountered by UAV tracking systems. Unlike existing confidence-prioritized association methods, the main distinction of our approach lies in jointly combining multi-level confidence stratification, Kalman-covariance-based prediction reliability estimation, and adaptive motion-appearance cost fusion.

2 Related Works

Multi-object tracking algorithms designed for UAV scenarios are generally derived from methods developed for generic environments, yet they incorporate a series of targeted improvements to adapt to the characteristics of aerial data. Among these characteristics, the frequent camera motion, complex target dynamics, drastic scale changes, and lower visual discriminability of targets introduce significant challenges for association. Consequently, many existing studies strengthen the association stage by integrating motion trajectories, appearance representations, spatial constraints, and temporal continuity. Based on their technical design and application emphasis, these methods can be reviewed from two primary perspectives: those originally proposed for general-purpose multi-object tracking, and those specifically tailored to address the unique difficulties of UAV-based tracking.

In the category of general-purpose tracking methods, SORT [1] proposed by Bewley and colleagues, established the widely adopted paradigm of detection-based tracking. SORT utilizes the Kalman filter to generate position predictions for existing trajectories and relies on the intersection-over-union (IoU) metric to perform matching. Although simple and efficient, its reliance on motion cues alone limits its robustness in occluded or crowded scenes. DeepSORT [11] introduced as an extension of SORT, alleviates this problem by incorporating an independently trained re-identification network that extracts discriminative appearance features. The addition of stable appearance cues significantly enhances the robustness of tracking in the

presence of occlusion, re-entry, and similar-looking targets. ByteTrack [12] further improves detection-based tracking by addressing the common issue of low-confidence detections being discarded too early. It proposes a two-stage association strategy that first matches high-confidence detections with existing trajectories and then allows unmatched low-confidence detections to be associated in a secondary stage. This design recovers many targets that would otherwise be missed because of temporary occlusion, blur, or insufficient detector certainty. BoT-SORT [13] continues this line of improvement by redesigning the Kalman filter state representation. Instead of relying on bounding box area and aspect ratio, it models width and height directly, which enhances stability when targets undergo scale variation. In addition, BoT-SORT introduces camera-motion compensation to address global image displacement caused by camera movement, allowing motion prediction to remain reliable even under dynamic camera conditions. UCMCTrack [14] tackles an even more challenging issue: large box displacement caused by UAV motion. It projects bounding boxes onto the physical ground plane during matching, which significantly reduces the geometric distortion caused by perspective variations and improves the robustness of data association under camera motion. Other methods seek to handle detection and re-identification jointly. FairMOT [15] presents a balanced framework that addresses the inconsistency between detection features and re-identification features in traditional pipelines. It adopts a unified architecture that allows both tasks to be optimized simultaneously with an anchor-free detection strategy that avoids spatial misalignment. QDTrack [16] pushes appearance modeling further through a quasi-dense similarity learning approach. It performs contrastive learning by densely sampling image regions, constructing large numbers of positive and negative sample pairs. This leads to more discriminative appearance representations, which strengthen tracking performance especially when targets are visually similar or appear at low resolutions.

Methods specifically designed for UAV scenarios incorporate additional mechanisms to cope with aerial-view challenges. GIAOTracker [17] uses a three-stage architecture: it compensates camera motion with Oriented FAST and Rotated BRIEF (ORB) and Random Sample Consensus (RANSAC), extracts appearance features via an identification network, and models motion with adaptive and unscented Kalman filters. It links trajectories via transformer-based temporal dependency modeling and refines results through post-processing. UAVMOT [18] enhances temporal consistency with keypoint compression and attention weights, adapts association modes by motion pattern, and improves heatmap learning with category-balancing and small-target-aware weights. FOLT [19] improves small-object tracking by aggregating features from adjacent frames through optical flow, enabling the detector to use information from multiple perspectives to improve recognition. It introduces a motion prediction module that uses a lightweight convolutional network to estimate object displacement directly, replacing the classical Kalman filter and improving performance under large motion. The overall framework combines a lightweight detector with an efficient optical flow estimator to achieve real-time performance on UAV platforms. SFTrack [20] refines association by matching high-confidence detections first and incorporating low-confidence ones only after validation. Ren et al. [21] proposed SuperMOT for UAV multi-object tracking. It employs a pyramid deformable alignment module for multi-level temporal feature fusion, a lightweight video super-resolution module to enhance small-object representation, and a motion decoupling module to separately model object and platform motion. Zheng et al. [22] proposed the DFA-MOT, a joint detection and tracking framework. It mitigates geometric distortion caused by UAV movement via the Dynamic Field-of-View Consistency Learning module and improves the prediction accuracy of target position and scale through the Scale-Aware Tracking mechanism. Ling et al. [23] proposed GraphTrack-UAV, a graph-based multi-object tracking framework for aerial videos. It builds a spatio-temporal graph with detections as nodes and inter-object relations as edges, and uses a graph pruning strategy to maintain efficiency. By message-passing graph neural network (GNN), the model captures long-range dependencies, and feature fusion and node aggregation improve robustness to occlusion

and appearance variation. Different from ByteTrack and related confidence-aware trackers, which mainly exploit detection confidence through staged matching, confidence adjustment, or adaptive thresholds, our method further introduces fine-grained multi-level stratification and explicitly evaluates Kalman prediction reliability from its covariance. The resulting prediction confidence is then incorporated into the dynamic fusion of motion and appearance costs.

3 Proposed Model

As illustrated in Fig. 1, the proposed method follows the classical SORT tracking paradigm. For clarity, the overall tracking procedure is organized into five functional stages: detector input, Kalman prediction, hierarchical confidence-based association, dynamic cost fusion, and trajectory output and update. The Kalman filter is used to estimate the motion state of each trajectory, while a re-identification (ReID) module provides appearance features that help distinguish different targets. Detection confidence and trajectory prediction confidence are treated as two distinct reliability quantities: the former determines the confidence group to which each detection belongs, whereas the latter modulates the contribution of motion information during association. The association problem is formulated as a bipartite graph matching task and is solved through the Hungarian algorithm. The cost matrix for this algorithm is constructed by combining the motion and appearance costs between trajectory predictions and current detections. The resulting association cost follows a consistent minimization direction and is used by the Hungarian algorithm for trajectory–detection assignment. The central design principle of the method is to assign higher matching priority to reliable detections. Following this principle, detections with high confidence scores are matched first, whereas detections with lower confidence are handled at later stages. To make full use of both appearance and motion information, the method introduces a reliability evaluation mechanism for motion prediction and a dynamic fusion strategy that adaptively adjusts the contribution of appearance and motion cues. Together, these components provide a unified confidence-aware association procedure in which detection reliability, motion-prediction reliability, and motion–appearance costs are explicitly coordinated during hierarchical matching.

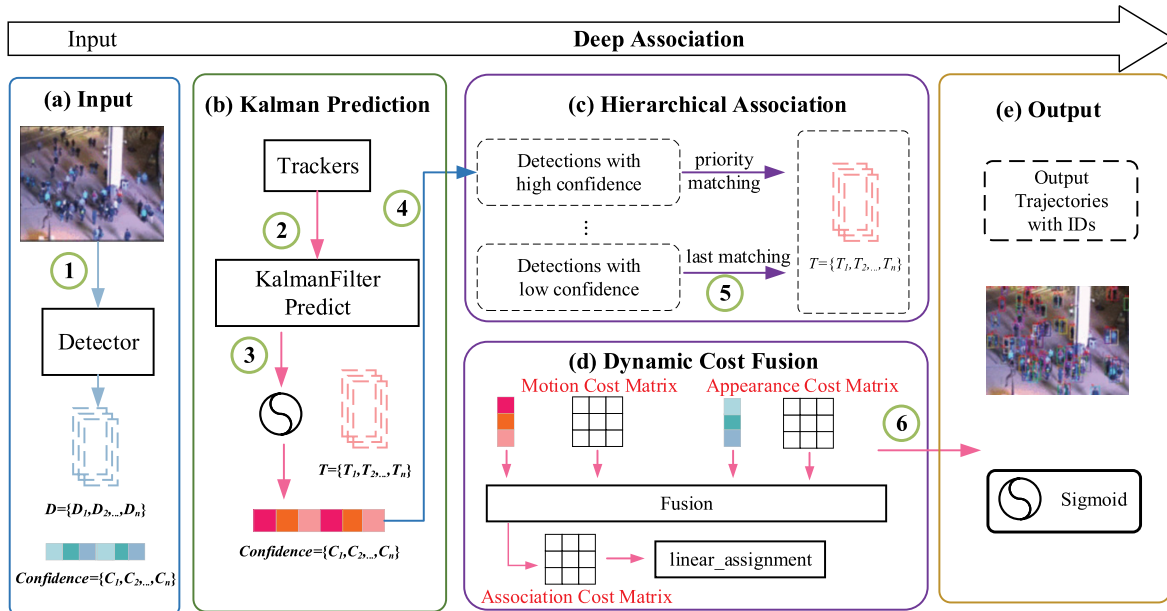


Figure 1: Multi-level hierarchical dynamic association strategy.

Fig. 1 shows the overall multi-level hierarchical dynamic association framework. The complete tracking pipeline is summarized in Algorithm 1. For each frame at time t , the detector produces a detection set $D^t = \{d_j^t\}_{j=1}^{N_D^t}$, where j denotes the detection index and N_D^t is the number of detections in the current frame. Each detection d_j^t contains a bounding box that specifies the location, width, and height of the detected target, together with a detection confidence $D_{c_i}^t$. The Kalman filter then predicts the motion of each existing trajectory, where the i -th trajectory at frame t is denoted by T_i^t . Each T_i^t is associated with a predicted bounding box and a trajectory prediction confidence $T_{c_i}^t$, which reflects the reliability of the predicted state. Here, i , j , and k denote the trajectory, detection, and confidence-group indices, respectively, while the superscript t denotes the temporal index. After obtaining the detection set D^t and the predicted trajectory set T^t , the two sets are passed into the hierarchical association module.

Within the hierarchical association process, all detection boxes are first grouped according to their confidence values using τ . The groups are then matched with the predicted trajectories one by one in descending order of confidence. During matching, the association cost matrix is constructed by combining motion cost and appearance cost. The motion cost is derived from the Intersection over Union (IoU) similarity between the predicted and detected bounding boxes, whereas the appearance cost is computed using the cosine distance between the feature vectors extracted by the re-identification module. Through dynamic weighting, the final cost matrix encourages reliable motion predictions while avoiding inaccurate associations when motion estimation becomes unreliable.

Once all confidence levels have been processed, there may still be a small number of detections and trajectories that remain unmatched. These residual unmatched elements are handled by an additional IoU-based association stage. The preceding hierarchical stages rely on the dynamically weighted combination of motion and appearance information to establish reliable correspondences progressively. However, for some residual elements, appearance features may be ambiguous or degraded because of small target size, partial occlusion, or motion blur, which are common in UAV imagery. In such cases, repeatedly applying the same motion–appearance fusion criterion may provide limited additional benefit. Therefore, the final stage relies only on spatial overlap, using the IoU-derived cost $1 - \text{IoU}$ between the remaining predicted boxes and detections, thereby avoiding additional dependence on potentially unreliable appearance features. Since this stage operates only on the residual unmatched detections and trajectories and does not modify associations already established in the preceding hierarchical stages, it serves as a supplementary fallback mechanism to recover plausible short-term correspondences and reduce unnecessary trajectory fragmentation. The resulting associations are then used to update the trajectory set T^t . Trajectories that fail to match with any detection box are not immediately removed. To maintain long-term identity consistency, unmatched trajectories are temporarily kept in the set. A trajectory is deleted only if it remains unmatched for more than thirty consecutive frames. This strategy avoids unnecessary fragmentation of tracks caused by brief occlusions or temporary detector failures.

The complete computational steps are given in Algorithm 1.

Algorithm 1: Pseudo-code of the proposed tracking method

Input: video sequence V , object detector Det , confidence interval size τ

Output: trajectory set T

- 1 Initialize T as an empty set
 - 2 For each frame f_t in video sequence V
 - 3 Obtain detection set $D^t = \{d_j^t\}_{j=1}^{N_D^t}$ from $Det(f_t)$
-

(Continued)

Algorithm 1 (continued)

```

4   For each trajectory  $T_i^t$  in  $\tau$ 
5       Use KalmanFilter to generate prediction  $D_{box_i}^t$ 
6       Compute the trajectory prediction confidence  $T_{c_i}^t$  based on the Kalman filter
          prediction uncertainty
7       Update  $T_i^t$  with the predicted state and trajectory prediction confidence  $T_{c_i}^t$ 
      End
8   Group detections in  $D^t$  according to their detector confidence scores  $d_j^t.score$ 
9   For each detection  $d_j^t$ 
10      Determine the confidence level  $k$  according to  $1 - k\tau < d_j^t.score \leq 1 - (k - 1)\tau$ 
11      Assign  $d_j^t$  to the corresponding confidence group  $D_k^t$ 
12      Continue grouping all detections in  $D^t$  in this manner
13  End
14  For each confidence group  $G_k^t, k = 1, 2, \dots, K$ 
15      Perform association between  $G_k^t$  and the current trajectory candidate set using the dynamic
          association cost
16      Remove successfully matched trajectories from the trajectory candidate set
17      Pass the unmatched detections in  $G_k^{(t)}$  to the next confidence group  $G_{k+1}^{(t)}$ 
18  End
19  Identify remaining trajectories that were not matched
20  Perform IoU-based association between the remaining unmatched detections and trajectories
    using the IoU-derived cost
21  For each matched trajectory in  $T$ 
22      Update its state based on the associated detection
23  Remove trajectories that have failed to match for more than the allowed threshold
24  For each remaining unmatched detection  $d$ 
25      Initialize a new trajectory in  $T$ 
26  End
27 End
28 Return the final set  $T$ 

```

3.1 Detector

The proposed method depends on the bounding boxes and confidence scores produced by the object detector. For each frame in the video sequence, the tracker must obtain reliable estimates of potential target locations as well as the confidence values that reflect the likelihood of each object category. These outputs form the basic input to the tracking module and are typically generated by an independent detection model.

In UAV application scenarios, real-time performance is essential. The computational capacity available on UAV platforms is often limited, yet the tracking system must still process high-frame-rate video streams to respond promptly to rapidly moving targets. As a result, the detector must achieve a balanced compromise between accuracy and inference speed. Taking these requirements into account, YOLOX-S [24] is adopted as the base detection model in this work. It provides a suitable trade-off between precision and efficiency, making it appropriate for deployment on resource-constrained UAV systems.

3.2 Multi-Level Association Strategy

Data association is the central element of multi-target tracking, whose goal is to match current-frame detections with historical trajectories in a reliable and temporally consistent manner. In UAV scenarios, the association process becomes considerably more challenging due to frequent target occlusions, rapid variations in scale, dense distributions of small objects and the inevitable missed detections and false alarms produced by the detector. These factors often cause traditional single-stage association strategies to lose track continuity or drift, especially in environments characterized by high motion dynamics or partial visibility.

To mitigate these challenges, this work proposes a confidence-driven deep association framework that improves robustness through progressive matching and adaptive weighting mechanisms. The proposed approach first introduces a confidence-based stratification procedure, in which candidate detections are grouped according to their confidence values and matched with trajectories in descending order of reliability. By prioritizing high-confidence detections, the system significantly reduces the propagation of association errors and enhances the stability of long-term trajectory maintenance. In addition, a prediction-uncertainty evaluation module is designed to quantify the reliability of motion information produced by the Kalman Filter. By converting the covariance structure within the filter into an interpretable measure of prediction confidence, the system is able to assess the trustworthiness of motion cues and avoid over-reliance on inaccurate predictions. Furthermore, the method constructs an association cost matrix through a dynamic coupling strategy that balances motion information and appearance information. The contribution of each information source is adaptively adjusted according to the reliability of the detector outputs and the temporal context of the current frame. This allows the association process to emphasize the modality that offers more stable cues under different conditions, enabling the framework to maintain robust tracking performance even in complex UAV environments.

(1) Multi-Level Matching Mechanism

The confidence scores produced by the detector contain rich information about the reliability of each observation. High-confidence detections generally correspond to clear and complete targets whose localization and category information are trustworthy. In contrast, low-confidence detections may result from partial occlusion, long viewing distance or motion-induced blur, and therefore carry a higher risk of false alarms. When all detections are treated equally during association, low-quality observations may interfere with or even corrupt existing stable trajectories. To alleviate this issue, the detection set $D^t = \{d_1^t, \dots, d_n^t\}$ is stratified according to confidence scores, while the trajectory set is denoted as $T^t = \{t_1^t, \dots, t_n^t\}$. The detection set is divided into multiple confidence levels $D_1^t, D_2^t, \dots, D_n^t$ according to

$$D_k^t = \{d_j^t \mid 1 - k\tau < d_j^t.score \leq 1 - (k-1)\tau\}, \quad (1)$$

where τ denotes the confidence interval size, $d_j^t.score$ denotes the confidence score of detection d_j^t , j denotes the detection index, and k denotes the confidence-group index. According to Eq. (1), the detection set is partitioned into descending confidence intervals using a left-open and right-closed convention. Thus, a detection whose confidence score lies exactly on an internal boundary is uniquely assigned to the adjacent lower-confidence group. The same inequality direction and boundary convention are used in Fig. 2, and matching is performed sequentially from the highest-confidence group to the lowest-confidence group. For clarity of illustration, d_1^t , d_k^t , and d_n^t in Fig. 2 denote representative detections in the first, intermediate, and last confidence groups, respectively, and should not be interpreted as a redefinition of the general detection index j used in Eq. (1). Here, the subscript of D_i^t denotes the confidence level in the hierarchical grouping, while the confidence score of each detection determines its membership in the corresponding interval. Matching is then performed sequentially from the highest-confidence group D_1^t to the lowest-confidence

group D_n^t . For each level, the dynamic cost matrix introduced in the subsequent sections is used to associate the detections with the trajectory set T^t .

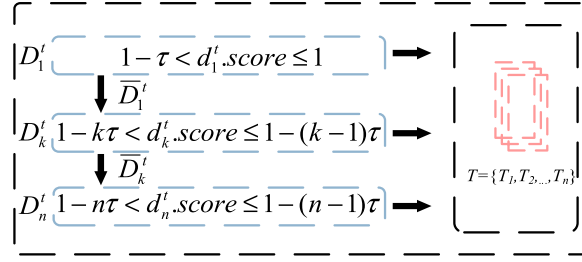


Figure 2: Confidence stratification strategy.

As shown in Fig. 2, the advantage of this multi-level strategy is that it prioritizes the establishment of the most reliable associations, thereby providing a stable foundation for the entire tracking process. When a high-confidence detection fails to match any trajectory in its current level, it is not discarded; instead, it is passed to the next confidence level and continues to participate in later matching rounds. This progressive degradation mechanism prevents the loss of newly appearing targets due to a single association failure. Meanwhile, the hierarchical filtering structure suppresses the influence of unreliable detections, effectively reducing the likelihood that low-confidence observations disrupt well-established trajectories.

(2) Accuracy Measurement Mechanism for Kalman Filter Prediction Boxes

The Kalman Filter predicts the future position of each target by modeling its underlying motion state. However, the reliability of these predictions varies over time. When a target moves smoothly with nearly constant velocity, the prediction tends to be highly accurate. In contrast, abrupt maneuvers such as sudden turns or acceleration, as well as long-term occlusion, can cause the prediction error of the constant-velocity model to increase significantly. Conventional tracking methods typically overlook this dynamic variability in prediction reliability. To explicitly quantify such uncertainty, the error covariance matrix P , an internal component of the Kalman Filter, is employed as the measurement source.

The diagonal element P_{ii} of matrix P represents the estimation variance of the i -th state variable, such as the x coordinate of the target center. A larger variance implies greater uncertainty in the estimation of that state component. Based on this principle, the raw confidence of the i -th state variable is defined as the reciprocal of the square root of its variance:

$$conf_i = \frac{1}{\sqrt{P_{ii}}} \quad (2)$$

arithmetic mean. To bound the raw confidence values while preserving their monotonic relationship with prediction uncertainty, the Sigmoid function is adopted. Specifically, the input to the Sigmoid function is the raw state-wise confidence $c_i^{raw} = 1/\sqrt{P_{ii}}$, where P_{ii} denotes the corresponding diagonal element of the Kalman prediction covariance matrix. Since c_i^{raw} is non-negative, the resulting Sigmoid output lies in $[0.5, 1]$ rather than the full interval $[0, 1]$. The Sigmoid function is expressed as

$$\sigma(x) = \frac{1}{1 + e^{-\alpha x}} \quad (3)$$

where α is a scaling parameter that controls the steepness of the function. The parameter is used to control the numerical sensitivity of the nonlinear transformation and is fixed throughout all experiments rather than being adaptively tuned for individual sequences. A larger α makes the transformed confidence more sensitive

to changes in the raw confidence, whereas a smaller α produces a smoother variation. The scaling parameter α is set to 0.5 and fixed throughout all experiments. It controls the steepness of the Sigmoid function and thus the sensitivity of the nonlinear transformation to changes in the raw confidence. In this work, α is used only to control the sensitivity of the confidence mapping and does not alter the monotonic relationship between the Kalman covariance and prediction confidence.

After applying the Sigmoid transformation to each raw confidence value, the normalized confidence becomes

$$norm_{conf_i} = \frac{1}{1 + e^{-\alpha conf_i}} \quad (4)$$

It should be emphasized that the transformed value is not interpreted as the probability that the Kalman prediction is correct. Instead, it serves as a relative reliability score for modulating the contribution of motion information in the subsequent association stage. When P_{ii} increases, indicating larger prediction uncertainty, $1/\sqrt{P_{ii}}$ decreases and the transformed confidence approaches its lower bound of 0.5. Conversely, a smaller prediction variance produces a larger confidence value approaching 1. The nonzero lower bound is retained because the proposed mechanism does not completely discard motion information when the Kalman prediction becomes uncertain. Instead, it reduces its relative contribution while preserving the temporal cue provided by the motion model for subsequent association.

Finally, the overall prediction confidence of the Kalman Filter, denoted as $conf_{KF}$, is computed as the arithmetic mean of the normalized confidence values of all state variables:

$$conf_{KF} = \frac{1}{N} \sum_{i=1}^N norm_{conf_i} \quad (5)$$

where N is the total number of state components. The arithmetic mean used here should be interpreted as an equal-weight aggregation of the transformed state-wise reliability scores rather than as a probabilistic fusion of the original covariance quantities. In other words, the proposed confidence is a heuristic reliability indicator designed for adaptive association weighting, rather than a statistically calibrated probability measure. Through variance-based quantification and normalization, this method provides an integrated confidence measurement that reflects the reliability of the Kalman Filter's prediction at each frame. The resulting metric provides a simple and interpretable relative indicator of the reliability of the Kalman prediction, thereby supporting adaptive adjustment of motion information in the subsequent association stage.

(3) Dynamic Cost Matrix

Data association ultimately depends on measuring the discrepancy between detection boxes and trajectory prediction boxes, which is represented by the association cost matrix. A common approach is to compute the motion cost matrix and the appearance cost matrix independently, and then fuse them using a fixed linear weighting scheme. However, fixed-weight fusion cannot accommodate the rapidly changing conditions encountered in UAV scenarios. To address this limitation, a dynamic coupling strategy is introduced, allowing the contributions of motion cues and appearance cues to adjust adaptively according to the context of the current frame. The overall association cost matrix C is formulated as

$$C = \frac{C_{motion} \cdot conf_{KF} + w \cdot C_{appearance}}{1 + w} \quad (6)$$

where C_{motion} is the motion cost matrix and $C_{appearance}$ is the appearance cost matrix. It should be emphasized that $conf_{KF}$ and w are not treated as two equivalent normalization weights in Eq. (6). The coefficient determines the contribution of the appearance term, whereas $conf_{KF}$ acts as a reliability modulation factor

applied to the motion term. Specifically, the nominal coefficient of the motion branch is set to 1 and is further modulated by $conf_{KF}$. Accordingly, Eq. (6) can be equivalently interpreted as

$$C = \frac{1 \cdot conf_{KF} \cdot C_{motion} + w \cdot C_{appearance}}{1 + w} \quad (7)$$

where the $1 + w$ denominator corresponds to the nominal coefficients of the motion and appearance branches. The factor $conf_{KF}$ independently adjusts the effective contribution of the motion information according to the reliability of the Kalman prediction. If $conf_{KF} + w$ were used as the denominator, the reduction of $conf_{KF}$ would simultaneously reduce the normalized motion contribution and implicitly increase the normalized relative contribution of the appearance term. In contrast, the proposed formulation allows the motion contribution to be attenuated according to its own reliability without automatically increasing the appearance contribution, whose reliability is separately characterized by w . Moreover, both component costs are represented on a common normalized scale before fusion. Given $C_{motion}, C_{appearance} \in [0, 1]$, $conf_{KF} \in [0, 1]$ and $w \geq 0$, Eq. (6) satisfies

$$0 \leq C \leq \frac{conf_{KF} + w}{1 + w} \leq 1 \quad (8)$$

which prevents uncontrolled amplification of the fused association cost and keeps it on a bounded scale for subsequent assignment. The motion cost matrix C_{motion} is derived from the Intersection over Union between the detection set D^t and the predicted trajectory set T^t generated by the Kalman Filter. IoU itself is a similarity measure, for which a larger value indicates greater spatial overlap. To maintain a consistent minimization direction with the subsequent Hungarian assignment, the IoU similarity is converted into a motion cost as

$$C_{motion} = 1 - \text{IoU}(D^t, T^t) \quad (9)$$

Accordingly, a smaller C_{motion} indicates stronger spatial consistency between a trajectory prediction and a detection. Therefore, C_{motion} follows the same minimization direction as the appearance cost and the final association cost. The motion and appearance costs are subsequently fused through Eq. (6), and the resulting association cost matrix C is minimized by the Hungarian algorithm for trajectory–detection assignment.

The appearance cost matrix $C_{appearance}$ is obtained from the cosine distance between the deep features extracted from each detection and the historical appearance features stored in the trajectories. A ReID model is used to compute these feature embeddings, enabling the differentiation of targets with similar visual characteristics. The appearance cost is therefore described as

$$C_{appearance} = \text{Cos}(D^t.\text{feature}, T^t.\text{feature}) \quad (10)$$

Accordingly, $C_{appearance}$ retains its definition as the appearance cost matrix, with each entry characterizing the appearance cost of the corresponding trajectory–detection pair.

To integrate motion and appearance cues, a comprehensive weighting coefficient w is introduced. This coefficient incorporates the detector confidence $D^t.confidence$ and the motion cost information contained in C_{motion} . Since the appearance cost depends jointly on the features extracted from the detector and those preserved in the trajectory history, the influence of both components should be fully reflected in its weighting. Hence, the coefficient is defined as

$$w = D^t.confidence \cdot C_{motion} \quad (11)$$

For further clarification, the weighting quantity w is defined pairwise for the trajectories and detections involved in the current association stage. Suppose that N_T trajectories and N_D detections participate in the association. Then w has the same pairwise dimension as the corresponding cost matrices, i.e., $w \in \mathbb{R}^{N_T \times N_D}$. Accordingly, each element w_{ij} corresponds to the weighting quantity associated with the i -th trajectory and the j -th detection. Therefore, Eq. (11) should be understood in a pairwise matrix form, consistent with the dimensions of C_{motion} , $C_{appearance}$, and the final association cost matrix C .

The central idea behind the dynamic cost matrix is to achieve context-dependent coupling of motion and appearance information. First, $conf_{KF}$ serves as a reliability modulation factor for C_{motion} , allowing the effective contribution of the motion information to vary according to the reliability of the Kalman prediction. When the Kalman prediction is accurate, motion information is emphasized. When prediction uncertainty increases, its contribution is reduced to avoid propagating unreliable estimates. Importantly, the reduction of $conf_{KF}$ does not directly increase the appearance contribution, because the appearance term is independently controlled by the adaptive coefficient w . Second, the design of the appearance weight w provides additional adaptivity. Higher detector confidence implies clearer targets and better image quality, which generally yield more reliable appearance features. Likewise, a larger C_{motion} indicates weaker spatial consistency between the predicted trajectory and the detection. In such cases, appearance information can provide complementary evidence for association, particularly when the detection confidence is high. This dynamic coupling mechanism ensures that the association process remains both robust and efficient under different operational environments, achieving an effective balance between tracking accuracy and computational efficiency.

4 Case Study and Experimental Results

To comprehensively evaluate the performance of the proposed tracker, extensive experiments were conducted on two publicly available UAV multi-object tracking datasets: VisDrone2019 [25] and UAVDT [26]. These datasets capture complex urban environments, crowded scenes and diverse motion patterns, making them well suited for assessing tracking robustness in realistic UAV applications. For quantitative evaluation, four commonly used multi-object tracking metrics were adopted, namely MOTA, MOTP, IDF1, and HOTA.

4.1 Experimental Details

All experiments were conducted following the official training and testing splits provided by the VisDrone2019 and UAVDT datasets. YOLOX-S was adopted as the baseline object detector for both datasets, with the input image resolution set to 1280×1280 . The detector was optimized using Stochastic Gradient Descent, where the learning rate was fixed at 0.01. Each dataset was trained for a total of 60 epochs. The training processes were executed on a single NVIDIA GeForce RTX 2080 Ti GPU, while inference was performed on an NVIDIA MX250 GPU. This configuration allows for a realistic evaluation of the computational efficiency of the proposed tracking framework under resource-constrained conditions. The Kalman filter adopts an eight-dimensional state vector $\mathbf{X} = [x, y, w, h, \dot{x}, \dot{y}, \dot{w}, \dot{h}]^T$ based on a constant-velocity model. The confidence-group interval is set to $\tau = 0.1$, and a trajectory is removed if it remains unmatched for more than 30 consecutive frames. For appearance representation, OSNet-x0.25 with publicly available pretrained weights is used without additional training. No appearance-feature update strategy or camera-motion compensation is applied.

For a fair comparison, all baseline trackers were evaluated under the same experimental protocol as the proposed method. In particular, the same YOLOX-S detections, detection thresholds, input preprocessing, dataset splits, track initialization and termination settings, and evaluation toolkit were used throughout the comparison. For trackers requiring appearance information, the same ReID features were provided,

whereas trackers that do not rely on appearance features retained their original association structures. Thus, the comparison mainly reflects differences in the tracking and data-association strategies rather than differences in the detector or evaluation configuration. All baseline methods were reproduced using their publicly available implementations under the unified configuration described above, and the results reported in Tables 1 and 2 were obtained from these controlled evaluations.

Table 1: Experimental results on the VisDrone2019 dataset.

Method	MOTA↑	MOTP↓	IDF1↑	HOTA↑
SORT [1]	35.431	75.496	47.268	40.25
DeepSORT [11]	34.083	75.043	49.349	40.955
BotSORT [13]	35.732	76.224	56.493	41.554
Bytetrack [12]	39.248	75.187	54.111	43.378
OCSORT [27]	39.986	76.173	50.266	42.092
StrongSORT [28]	39.309	75.704	50.301	41.81
Sparsetrack [29]	36.09	74.02	55.214	43.373
Ours	41.439	74.005	54.123	44.001

Table 2: Experimental results on the UAVDT dataset.

Method	MOTA↑	MOTP↓	IDF1↑	HOTA↑
SORT	47.55	78.578	66.107	50.018
DeepSORT	48.218	79.309	65.517	50.567
BotSORT	47.676	78.62	66.418	50.204
Bytetrack	48.359	79.247	66.915	51.079
Sparsetrack	47.55	78.578	66.107	50.018
OCSORT	48.021	79.533	65.161	50.667
StrongSORT	48.359	79.25	67.001	51.119
Ours	49	79.474	66.904	52.03

4.2 Dataset

VisDrone2019, developed by the AISKYEYE team of Tianjin University, is one of the most representative UAV multi-object tracking datasets. It contains 288 video sequences with a total of 261,908 frames captured across 14 cities in China. The dataset covers diverse environments, including urban streets, residential areas and rural roads, as well as a variety of tracking conditions such as sparse pedestrian flows and dense traffic scenarios. The Multi-Object Tracking (MOT) subset follows the official partitioning standards and includes a training set, validation set, test-dev set and test-challenge set, each containing multiple annotated sequences tailored for model training, tuning and evaluation.

UAVDT (Unmanned Aerial Vehicle Detection and Tracking) is another large-scale benchmark designed specifically for object detection and tracking from UAV perspectives. It contains 100 video sequences with approximately 80,000 frames captured in complex urban environments, such as highways, intersections and public squares. All videos are provided at a resolution of 1080×540 , emphasizing real-world challenges including fast-moving vehicles, scale variation and frequent occlusion.

4.3 Quantitative Evaluation Results

To evaluate the effectiveness of the proposed method, comparative experiments were conducted on the VisDrone2019 and UAVDT datasets. The proposed tracker was compared with several state-of-the-art (SOTA) methods across all standard multi-object tracking metrics. The quantitative results are summarized in [Tables 1](#) and [2](#).

[Table 1](#) reports the performance of our method on the VisDrone2019 dataset. In the table, $\uparrow$ indicates that a higher value corresponds to better performance, while $\downarrow$ indicates that a lower value is preferable. With respect to MOTA and HOTA, the two most indicative metrics of overall tracking performance, the proposed method achieves scores of 41.439 and 44.001, respectively, surpassing all competing approaches. These results demonstrate that our method maintains superior robustness and association accuracy in highly dynamic UAV environments. For MOTP, which evaluates the localization precision of predicted bounding boxes, our method attains the best score of 74.005, meaning that the spatial alignment between predicted trajectories and ground-truth annotations is the most accurate among all compared trackers. This improvement can be attributed to the multi-level association mechanism, which prioritizes reliable, high-confidence detections and therefore suppresses mismatches at the early stages of association. Regarding IDF1, which measures identity preservation capability, ours obtains a score of 54.123. Although this is comparable to ByteTrack and slightly lower than BotSORT (56.493), it reflects a deliberate trade-off, as our method emphasizes overall tracking robustness and localization precision rather than maximizing identity-only metrics. Overall, ours achieves the best performance in the three most critical metrics—MOTA, MOTP, and HOTA—highlighting its effectiveness across core tracking dimensions.

[Table 2](#) summarizes the results on the UAVDT dataset. The proposed method again achieves competitive performance and ranks first in both MOTA (49.000) and HOTA (52.03). These results confirm that the proposed multi-level association mechanism generalizes effectively across different UAV tracking conditions, ranging from the dense pedestrian scenarios in VisDrone to the fast-moving vehicle scenes in UAVDT. It is also worth noting that the performance ranking of individual metrics on UAVDT differs slightly from that on VisDrone. This discrepancy can be attributed to the distinct characteristics of the UAVDT dataset, where tracked objects—primarily vehicles—exhibit highly similar appearance features and relatively stable motion patterns. As a result, the differences between various trackers across individual metrics are less pronounced.

[Table 3](#) shows the tracking speed comparison. ByteTrack achieves 12.532 frames per second (FPS), while the proposed method achieves 10.001 FPS. The decrease in speed is mainly associated with the successive association operations introduced by the multi-level matching strategy. The comparison indicates that the proposed method introduces a certain computational cost due to its more elaborate association procedure. However, the overall processing efficiency remains comparable to that of ByteTrack, demonstrating that the proposed confidence-aware association strategy achieves a reasonable balance between tracking performance and computational efficiency.

Table 3: Tracking speed comparison.

Method	FPS
ByteTrack	12.532
Ours	10.001

4.4 Ablation Study and Discussion

To systematically evaluate the contribution of each component in the proposed multi-level hierarchical dynamic association strategy, three groups of ablation studies were conducted. First, we examine the fusion mechanism used to combine detection confidence, motion information and appearance information when constructing the association cost matrix. Second, we analyze the effectiveness of different strategies for generating motion prediction confidence and their influence on association decisions. Finally, several confidence-layer segmentation intervals are tested to study how hierarchical granularity affects tracking performance and to determine the optimal parameter configuration.

(1) Fusion Mechanism

To investigate the influence of information fusion on the construction of the association cost matrix and to reveal the performance differences among alternative fusion schemes, two baseline approaches were designed for comparison.

The first scheme adopts a direct additive fusion, where the appearance cost and motion cost are simply summed:

$$C = C_{motion} + C_{appearance} \quad (12)$$

The second scheme introduces a basic dynamic weighting mechanism by multiplying the motion cost with a confidence value $conf$ and combining it with the appearance cost weighted by a fixed scalar w :

$$C = C_{motion} \cdot conf + w \cdot C_{appearance} \quad (13)$$

As shown in the [Table 4](#), the fusion method proposed in this paper achieves the best performance. The reasons can be clarified by analyzing the limitations of the two comparative schemes: (1) The simple summation method demonstrates that motion information and appearance information contribute equally to the result. Although it is computationally convenient, it fails to consider the reliability differences of the two types of information in different scenarios. For example, when motion prediction has large deviations or appearance features become blurred due to occlusion, this scheme cannot dynamically adjust weights and is prone to be interfered by invalid information, leading to association errors. (2) The confidence-weighted multiplicative fusion scheme improves the adaptability of the cost matrix to the reliability of motion information to a certain extent, but it does not perform normalization on the fusion results. Since the value ranges of $conf_{KF}$ and w are different, the weight proportions of the two types of information lack a unified scale, which may cause the contribution of a certain type of information to be excessively amplified or suppressed, thereby affecting the rationality of the cost matrix.

Table 4: Different confidence fusion strategies.

Method	MOTA↑	MOTP↓	HOTA↑
Simply Summed	39.596	75.315	43.377
Fixed Scalar	40.021	75.564	42.7
Ours	41.439	74.005	44.001

(2) Confidence Generation Method

To quantitatively assess the effectiveness of the Kalman filter prediction confidence estimation module, this subsection conducts a comprehensive ablation study comparing five distinct confidence generation strategies on the VisDrone2019 validation set. These strategies aim to characterize the reliability of motion information from different theoretical perspectives and to adjust, in a data-driven manner, the relative contributions of appearance and motion cues within the association cost matrix.

The fixed-weight strategy serves as the baseline. It does not compute any dynamic confidence values; instead, it assigns a constant weight of 1.0 to the motion-based cost C_{motion} . This scheme is included to verify the necessity of incorporating a dynamic confidence mechanism.

The Mahalanobis-distance-based method evaluates prediction uncertainty by measuring the statistical deviation between the predicted state and the latest detector observation. The distance is computed as

$$d_M = \sqrt{(z - H\hat{X}_{k|k-1})^T P_{k|k-1}^{-1} (z - H\hat{X}_{k|k-1})} \quad (14)$$

and the corresponding confidence is defined as

$$conf_{mahalanobis} = e^{-\lambda d_M} \quad (15)$$

where λ is the attenuation coefficient (set to 0.5), z denotes the observed bounding box, and $H\hat{X}_{k|k-1}$ is the predicted measurement. Since the Mahalanobis distance accounts for parameter correlations, smaller values imply stronger statistical consistency between prediction and observation, thereby indicating higher confidence. Here, the residual used in the Mahalanobis distance is defined in the measurement space. Accordingly, the covariance matrix P in Eq. (14) denotes the covariance associated with this measurement-space prediction residual, rather than the unprojected covariance of the full Kalman state. To relate this notation to the conventional Kalman-filter formulation, if P_x^- denotes the predicted state covariance, H denotes the measurement matrix, and R denotes the measurement noise covariance, the corresponding measurement-space covariance can be expressed as $P = HP_x^- H^T + R$. Thus, the matrix denoted by P in Eq. (14) corresponds to the innovation covariance that is often denoted by S in standard Kalman-filter notation. Let n_z denote the dimension of the measurement vector. Then the observation z and the predicted measurement $\hat{z}$ satisfy $z, \hat{z} \in \mathbb{R}^{n_z}$, while $P \in \mathbb{R}^{n_z \times n_z}$. Therefore, the residual $(z - \hat{z}) \in \mathbb{R}^{n_z}$ is dimensionally compatible with P^{-1} in Eq. (14), and the resulting Mahalanobis distance is a scalar.

The trajectory-continuity-based strategy evaluates reliability from a temporal perspective. It assumes that trajectories exhibiting stable historical matching tend to have more reliable current predictions. Its confidence value is computed as

$$conf_{track} = \frac{N_{matched}}{N_{total}} \cdot \frac{1}{1 + \beta \cdot age} \quad (16)$$

where $N_{matched}$ denotes the number of frames with successful associations, N_{total} denotes the total number of frames of the trajectory, age is the trajectory length, and β is the decay coefficient (set to 0.05).

Table 5 summarizes the comparison of these confidence generation schemes. The results demonstrate that incorporating a dynamic confidence mechanism generally improves tracking performance, particularly in terms of MOTA and HOTA. The fixed-weight baseline achieves the lowest MOTA among all methods, while its HOTA is lower than that of the proposed method, although it remains competitive in terms of MOTP. These results indicate that a static fusion strategy is less effective for UAV scenarios characterized by rapid target motion, irregular movement patterns, and frequent occlusion. Under such conditions, the

uncertainty of motion prediction can fluctuate considerably, making dynamic adjustment of information weights beneficial for reducing association errors and improving overall tracking performance.

Table 5: Performance under different confidence generation methods.

Method	MOTA↑	MOTP↓	HOTA↑
Fixed Weight	39.596	75.315	43.377
Mahalanobis Distance	39.949	75.46	42.646
Trajectory Continuity	41.397	73.42	41.75
Ours	41.439	74.005	44.001

Among the dynamic strategies, the proposed confidence estimation mechanism achieves the highest overall performance and demonstrates clear advantages in metrics such as MOTA and HOTA, which jointly reflect detection accuracy and identity consistency. This indicates that leveraging the inverse covariance of the Kalman filter provides a more fine-grained and reliable quantification of prediction uncertainty. Although the Mahalanobis-distance strategy effectively captures instantaneous deviations between prediction and observation, its dependence on single-frame observations limits adaptability under continuously changing motion patterns. Similarly, the trajectory continuity method captures long-term stability but lacks sensitivity to rapid fluctuations in prediction reliability, thus offering only moderate improvements.

It should be noted that the proposed method does not produce the optimal MOTP score. This can be attributed to the mechanism's sensitivity to prediction uncertainty. When minor positional deviations occur in predicted bounding boxes, the system reduces reliance on motion cues and increases dependence on appearance features to prevent erroneous forced matching. While this may slightly affect local bounding box precision, it effectively mitigates more severe issues such as identity switches. Consequently, the proposed strategy achieves a favorable balance between overall tracking accuracy and identity preservation.

(3) Hierarchical Fine-Grained

The influence of the hierarchical fine-grained mechanism on the confidence-priority matching strategy is illustrated in the Fig. 3. The comparison reveals that an interval of 0.1 yields the most balanced performance across all metrics, with particularly notable improvements observed on the VisDrone2019 dataset. This interval size enables a more refined separation of detection confidence levels, which in turn facilitates smoother progressive matching and more stable association behavior. Based on these observations, the interval parameter is set to $\tau = 0.1$ as the default configuration for the proposed method.

To intuitively assess the effectiveness of the proposed method in real-world scenarios, a representative and challenging video sequence from the VisDrone2019 dataset is selected for visualization. This sequence contains dense small targets and notable UAV-induced camera motion. In such scenes, distant targets occupy only a few pixels and exhibit low visual clarity, while the motion patterns become highly complex due to the movement of the UAV platform. Under these conditions, the proposed method is visually compared with several representative trackers, including ByteTrack, OC-SORT, and BoT-SORT.

As illustrated in Fig. 4, all comparative trackers display varying degrees of performance degradation. ByteTrack employs a fixed threshold to segment detections during association; however, this static mechanism is highly sensitive to camera motion and leads to reduced association accuracy when global movement increases positional uncertainty. OC-SORT applies an observation-centric update strategy, yet its performance becomes limited in UAV scenarios because the bounding boxes generated by its lightweight detector lack the spatial accuracy required for stable association, ultimately causing deviations in trajectory

estimation. Although BoT-SORT integrates appearance features to enhance identity discrimination, its re-identification module suffers from the compounded effects of small-target blurring and motion blur, substantially diminishing the reliability of its appearance embeddings.

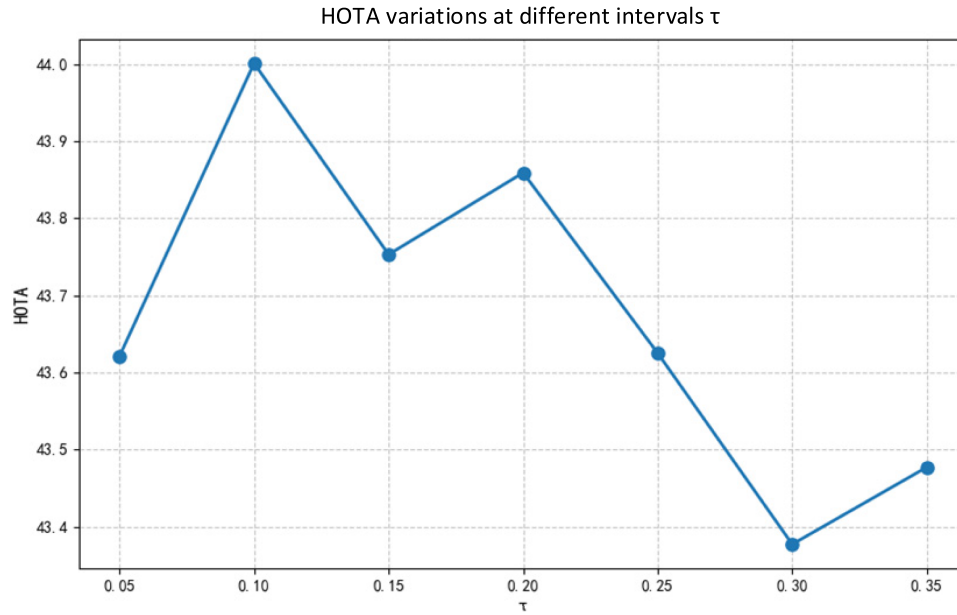


Figure 3: HOTA values under different τ on VisDrone2019.



Figure 4: Visualization of different trackers in complex scenarios.

In contrast, the proposed method maintains consistent and robust tracking behavior throughout the sequence. Its multi-level association mechanism plays a crucial role in handling small targets: by prioritizing the matching of high-confidence detection boxes with existing trajectories, the method helps maintain

more stable associations in the presence of low-quality detections. Additionally, the confidence estimation mechanism for the Kalman filter adaptively senses increases in prediction uncertainty caused by camera motion. By adjusting the weights in the dynamic cost matrix, the method reduces dependence on unreliable motion cues and increases reliance on appearance information whenever positional predictions become unstable. This adaptive trade-off allows the tracker to preserve correct associations even under abrupt or irregular target displacements.

The visualization results highlight the clear advantages of the proposed approach when dealing with scenes containing both small targets and camera motion. Through the coordinated operation of its internal components, the method exhibits stable identity assignment and maintains trajectory continuity for the majority of targets in the illustrated sequence. These observations further confirm the robustness and reliability of the proposed hierarchical dynamic association framework in complex UAV tracking environments.

5 Conclusion

This research presents a confidence-prioritized multi-object tracking framework tailored for UAV platforms. By integrating detection confidence with motion-aware association, the proposed method effectively enhances tracking stability under challenging conditions such as small object sizes, frequent occlusions, and camera motion. Extensive experiments on VisDrone2019 and UAVDT benchmarks demonstrate that the proposed approach achieves competitive performance in terms of MOTA and HOTA. The results indicate that explicitly leveraging confidence information can improve data association robustness. Despite its effectiveness, the current framework relies on fixed confidence thresholds and does not explicitly model long-term temporal dependencies. Future work will explore adaptive confidence modeling and tighter integration between detection and motion estimation modules, as well as extending the framework to more complex aerial scenarios.

Acknowledgement: Not applicable.

Funding Statement: This research was funded by the Shenzhen Science and Technology Program, grant number JCYJ20230807113708016, and the Guangdong Basic and Applied Basic Research Foundation, grant number 2024A1515011025.

Author Contributions: Long Wen and Yutao Chen contributed equally to this work. The authors confirm contribution to the paper as follows: Conceptualization, Long Wen, Yutao Chen and Zhe Cheng; methodology, Long Wen, Yuxin Liu, Hui Tan, Jiangbo He and Zhe Cheng; software, Long Wen; validation, Yutao Chen, Hui Tan and Zhe Cheng; formal analysis, Yutao Chen and Jiangbo He; investigation, Long Wen, Yuxin Liu, Hui Tan, Jiangbo He and Zhe Cheng; resources, Long Wen and Zhe Cheng; data curation, Long Wen, Yutao Chen and Jiangbo He; writing—original draft preparation, Long Wen, Yutao Chen and Jiangbo He; writing—review and editing, Long Wen, Yutao Chen and Jiangbo He; visualization, Zhe Cheng; supervision, Zhe Cheng; project administration, Long Wen and Yutao Chen; funding acquisition, Long Wen and Yutao Chen. All authors reviewed and approved the final version of the manuscript.

Availability of Data and Materials: The datasets analyzed in this study are VisDrone2019 [25] and UAVDT [26], both of which are publicly available. No new datasets were collected or generated for this study, and all data used in the experiments can be accessed from their respective public sources.

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Bewley A, Ge Z, Ott L, Ramos F, Upcroft B. Simple online and realtime tracking. In: Proceedings of the 2016 IEEE International Conference on Image Processing (ICIP); 2016 Sep 25–28; Phoenix, AZ, USA. p. 3464–8. doi:10.1109/ICIP.2016.7533003.
2. Kıratlı R, Ünlü Eroğlu H, Eroğlu A. TopoEKF: from state-space estimation to topological signatures for enhanced multi-object tracking and anomaly detection in UAVs. *Comput Model Eng Sci.* 2026;147(3):1–10. doi:10.32604/cmesci.2026.081411.
3. Wang Q, Zhou L, Xu C, Shang Y, Jin P, Cao C, et al. Progress and perspectives on UAV visual object tracking. *IEEE J Sel Top Appl Earth Obs Remote Sens.* 2025;18:20214–39. doi:10.1109/JSTARS.2025.3593286.
4. An Y, Ning J, Hou C, Ou J. Efficient low-collision UAV-based automated structural surface inspection using geometric digital twin and voxelized obstacle information. *Autom Constr.* 2025;171:105972. doi:10.1016/j.autcon.2025.105972.
5. Li S, Wang Z, Liang J, Wang Y. Small object detection in UAV scenarios based on YOLOv5. *Comput Model Eng Sci.* 2025;145(3):3993–4011. doi:10.32604/cmesci.2025.073896.
6. Qian Y, Wang Z, Gao Y, Zhang W, Wang H. A multi-object tracking method in moving UAV based on IoU matching. *IEEE Access.* 2024;12:139076–85. doi:10.1109/ACCESS.2024.3464575.
7. Delleji T, Fkih H, Kallel A, Chtourou Z. Visual tracking of mini-UAVs using modified YOLOv5 and improved DeepSORT algorithms. In: Proceedings of the 2022 6th International Conference on Advanced Technologies for Signal and Image Processing (ATSIP); 2022 May 24–27; Sfax, Tunisia. p. 1–6. doi:10.1109/ATSIP55956.2022.9805923.
8. Stanojevic VD, Todorovic BT. BoostTrack: boosting the similarity measure and detection confidence for improved multiple object tracking. *Mach Vis Appl.* 2024;35(3):53. doi:10.1007/s00138-024-01531-5.
9. Luo X, Zhang F, Firkat E, Hamdulla A, Zhu B, Dawut A. UAV vision-based object detection network with lightweight and multi-scale fusion. *Neurocomputing.* 2026;666:132211. doi:10.1016/j.neucom.2025.132211.
10. Dang Z, Sun X, Sun B, Guo R, Li C. OMCTrack: integrating occlusion perception and motion compensation for UAV multi-object tracking. *Drones.* 2024;8(9):480. doi:10.3390/drones8090480.
11. Wojke N, Bewley A, Paulus D. Simple online and realtime tracking with a deep association metric. In: Proceedings of the 2017 IEEE International Conference on Image Processing (ICIP); 2017 Sep 17–20; Beijing, China. p. 3645–9. doi:10.1109/ICIP.2017.8296962.
12. Zhang Y, Sun P, Jiang Y, Yu D, Weng F, Yuan Z, et al. ByteTrack: multi-object tracking by associating every detection box. In: *Computer Vision—ECCV 2022*. Cham, Switzerland: Springer; 2022. p. 1–21. doi:10.1007/978-3-031-20047-2_1.
13. Aharon N, Orfaig R, Bobrovsky BZ. BoT-SORT: robust associations multi-pedestrian tracking. *arXiv:2206.14651*. 2022. doi:10.48550/arXiv.2206.14651.
14. Yi K, Luo K, Luo X, Huang J, Wu H, Hu R, et al. UCMCTrack: multi-object tracking with uniform camera motion compensation. *AAAI.* 2024;38(7):6702–10. doi:10.1609/aaai.v38i7.28493.
15. Zhang Y, Wang C, Wang X, Zeng W, Liu W. FairMOT: on the fairness of detection and re-identification in multiple object tracking. *Int J Comput Vis.* 2021;129(11):3069–87. doi:10.1007/s11263-021-01513-4.
16. Fischer T, Huang TE, Pang J, Qiu L, Chen H, Darrell T, et al. QDTrack: quasi-dense similarity learning for appearance-only multiple object tracking. *IEEE Trans Pattern Anal Mach Intell.* 2023;45(12):15380–93. doi:10.1109/TPAMI.2023.3301975.
17. Du Y, Wan J, Zhao Y, Zhang B, Tong Z, Dong J. GIAOTracker: a comprehensive framework for MCMOT with global information and optimizing strategies in VisDrone 2021. In: Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW); 2021 Oct 11–17; Montreal, BC, Canada. p. 2809–19. doi:10.1109/ICCVW54120.2021.00315.
18. Liu S, Li X, Lu H, He Y. Multi-object tracking meets moving UAV. In: *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2022 Jun 18–24; New Orleans, LA, USA. p. 8866–75. doi:10.1109/CVPR52688.2022.00867.

19. Yao M, Wang J, Peng J, Chi M, Liu C. FOLT: fast multiple object tracking from UAV-captured videos based on optical flow. In: Proceedings of the 31st ACM International Conference on Multimedia; 2023 Oct 29–Nov 3; Ottawa, ON, Canada. p. 3375–83. doi:10.1145/3581783.3611868.
20. Song I, Lee J. SFTrack: a robust scale and motion adaptive algorithm for tracking small and fast moving objects. In: Proceedings of the 2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS); 2024 Oct 14–18; Abu Dhabi, United Arab Emirates. p. 10870–7. doi:10.1109/IROS58592.2024.10802537.
21. Ren L, Yin W, Diao W, Fu K, Sun X. SuperMOT: decoupling motion and fusing temporal pyramid features for UAV multiobject tracking. *IEEE J Sel Top Appl Earth Obs Remote Sens.* 2025;18:14188–202. doi:10.1109/JSTARS.2025.3563060.
22. Zheng Y, He C, Chen X, Zhang H, Qu T, Wang D. DFA-MOT: a dynamic field-aware multi-object tracking framework for uncrewed aerial vehicles. *IEEE Trans Circ Syst Video Technol.* 2025;35(11):11350–61. doi:10.1109/TCSVT.2025.3576487.
23. Ling Y, Lou Y, Ding Y, Cao Q, Rao Z. GraphTrack-UAV: robust multi-object association in aerial videos via message passing graph neural networks. In: Proceedings of the 2025 17th International Conference on Signal Processing Systems (ICSPS); 2025 Oct 24–26; Chengdu, China. p. 1046–54. doi:10.1109/ICSPS66615.2025.11347850.
24. Ge Z, Liu S, Wang F, Li Z, Sun J. YOLOX: exceeding YOLO series in 2021. arXiv:2107.08430. 2021. doi:10.48550/arXiv.2107.08430.
25. Du D, Zhu P, Wen L, Bian X, Lin H, Hu Q, et al. VisDrone-DET2019: the vision meets drone object detection in image challenge results. In: Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW); 2019 Oct 27–28; Seoul, Republic of Korea. p. 213–26. doi:10.1109/iccvw.2019.00030.
26. Yu H, Li G, Zhang W, Huang Q, Du D, Tian Q, et al. The unmanned aerial vehicle benchmark: object detection, tracking and baseline. *Int J Comput Vis.* 2020;128(5):1141–59. doi:10.1007/s11263-019-01266-1.
27. Cao J, Pang J, Weng X, Khirondkar R, Kitani K. Observation-centric SORT: rethinking SORT for robust multi-object tracking. In: Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2023 Jun 17–24; Vancouver, BC, Canada. p. 9686–96. doi:10.1109/CVPR52729.2023.00934.
28. Du Y, Zhao Z, Song Y, Zhao Y, Su F, Gong T, et al. StrongSORT: make DeepSORT great again. *IEEE Trans Multimed.* 2023;25:8725–37. doi:10.1109/TMM.2023.3240881.
29. Liu Z, Wang X, Wang C, Liu W, Bai X. SparseTrack: multi-object tracking by performing scene decomposition based on pseudo-depth. *IEEE Trans Circ Syst Video Technol.* 2025;35(5):4870–82. doi:10.1109/TCSVT.2024.3524670.