ISPRS Journal of Photogrammetry and Remote Sensing 229 (2025) 138–154

Contents lists available at ScienceDirect

ISPRS Journal of Photogrammetry and Remote Sensing

journal homepage: www.elsevier.com/locate/isprsjprs

MOSAIC-Tracker: Mutual-enhanced Occlusion-aware Spatiotemporal Adaptive Identity Consistency network for aerial multi-object tracking☆

Jian Zou [a], Wei Zhang [a,b], Qiang Li [a,*], Qi Wang [a,*]

[a] School of Artificial Intelligence, OPtics and ElectroNics (iOPEN), Northwestern Polytechnical University, Xi'an 710072, China
[b] School of Computer Science, Northwestern Polytechnical University, Xi'an 710072, China

ARTICLE INFO

Keywords:
Unmanned aerial vehicle video
Multi-object tracking
Spatiotemporal fusion
Multi-layer feature aggregation
Data association

ABSTRACT

Multi-Object Tracking (MOT) in aerial imagery remains challenging due to small object sizes, occlusions, and dynamic environments. Existing approaches predominantly rely on high precision detection and Re ID matching but neglect spatiotemporal cues and global temporal modeling of occlusion. Their static confidence weighting during association cannot adapt to real time detector confidence fluctuations, resulting in mismatches and ID switches. To alleviate these limitations, we propose MOSAIC-Tracker, a Mutual-enhanced Occlusion-aware Spatiotemporal Adaptive Identity Conservation Network with three key dimensions. First, a Spatiotemporal Occlusion Enhancement (STOE) module integrates multi-frame temporal dependencies to model global motion patterns and local dynamic features, mitigating identity switches during occlusions. Then, an Adaptive Multi-scale Feature Enhancement (AMFE) mechanism combines a Local Enhancement Mechanism with multi-scale feature aggregation to improve small object discrimination. Finally, a Dynamic Confidence Matrix Adjustment (DCMA) strategy adaptively weights detection confidence in trajectory matching to minimize association errors. Together, the three modules reduce occlusion-induced identity switches. Extensive evaluations on UAVDT and VisDrone2019 datasets demonstrate advanced performance. The code is released at: https://github.com/aJanm/MOSAIC-Tracker.

1. Introduction

Multi-Object Tracking (MOT) in aerial imagery is a fundamental Earth observation technique, widely applied in environmental monitoring, ecological conservation, disaster surveillance, and urban planning (Wu et al., 2021; Qiao et al., 2022; Zhu et al., 2024; Chen et al., 2024). With the rapid development of Unmanned Aerial Vehicles (UAVs), aerial MOT has gained increasing attention, offering new opportunities but also presenting unique challenges. As illustrated in Fig. 1, unlike conventional tracking scenarios, aerial imagery typically involves diverse occlusion conditions (Fig. 1a–c), small and densely distributed objects, motion blur, low-resolution imaging due to high-altitude capture, and frequent viewpoint variations induced by UAV motion (Xu et al., 2025; Huang et al., 2023; Ma et al., 2024b). These factors significantly complicate detection and tracking and often lead to marked performance degradation of general-purpose MOT algorithms. Therefore, specialized, more robust, and adaptive methods are required to effectively address the unique challenges of aerial multi-object tracking.

Current research in multi-object tracking primarily focuses on three directions: (1) enhancing object representation through mechanisms such as attention and multi-scale feature fusion (Chu et al., 2023; Liu et al., 2024b); (2)dynamically utilizing spatiotemporal relationships with deep learning methods and (3) improving detection-to-association accuracy via refined data-association strategies, thereby optimizing overall tracking performance.

With respect to feature enhancement, Xu et al. (2022) proposes a pixel-level, multi-scale feature query strategy built upon a Transformer architecture, which strengthens object recognition at the expense of real-time performance. DroneMOT (Wang et al., 2024c) introduces a dual-domain integrated attention module to improve small-object detection under rapid UAV motion; however, it has only been validated on single frames and fails to exploit temporal consistency. DepthMOT (Wu and Liu, 2024) leverages depth cues to augment small-object feature extraction, but its stringent depth-matching protocol leads to degraded identity preservation. Wen et al. (2023) employ an enhanced feature-selection mechanism to extract more representative visual features, achieving higher tracking accuracy. Similarly, other

☆ This work was supported by the National Natural Science Foundation of China under Grant U21B2041, 62471394, and 62301385.
* Corresponding authors.
E-mail addresses: zoujian@mail.nwpu.edu.cn (J. Zou), zhangwei707@mail.nwpu.edu.cn (W. Zhang), qiangli@nwpu.edu.cn (Q. Li), crabwq@gmail.com (Q. Wang).

https://doi.org/10.1016/j.isprsjprs.2025.08.013
Received 7 March 2025; Received in revised form 3 August 2025; Accepted 15 August 2025

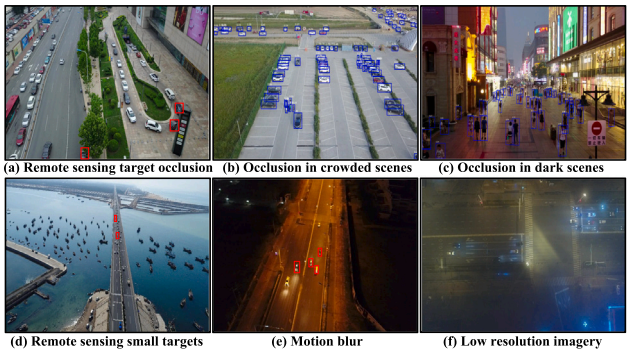

Fig. 1. Several challenges issues in UAT videos. Fig. 1 (a)–(c) illustrate the challenges of object occlusion under distinct observational scenarios. (d)–(f) demonstrate the challenge of densely distributed small objects, motion blur and low resolution degradation effects.

works (Xu et al., 2024b; Ma et al., 2024b; Liu et al., 2025) have explored various multi-scale fusion and attention-based techniques.

In the realm of spatiotemporal relationship enhancement, Yuan et al. (2025) employ short- and long-term memory modules together with a Time-Enhanced Feature Decoder to improve MOTA; however, under UAV footage where targets occupy very few pixels and suffer severe occlusions, their identity-preservation performance remains suboptimal. He et al. (2024) fuse multi-frame embeddings to enhance robustness to rapid motion, but in low–frame-rate, low-resolution, or highly jittery remote-sensing video, their encoder is vulnerable to motion blur, leading to degraded tracking accuracy. Similarly, Zhang et al. (2023b) construct a spatiotemporal neighborhood topology to reduce false matches and improve occlusion recovery; however, in densely distributed small-object scenarios with irregular motion patterns, their topology becomes unstable, indicating room for further improvement.

With respect to data association strategies, Li et al. (2025a) introduce a spatial-relation-based data-association method to improve matching under occlusion. Wang et al. (2024b) propose a motion-aware association scheme that enhances the tracking of densely distributed small objects with the cost of increased false positives. Yasir et al. (2024) design a C-BIoU association algorithm that incorporates a buffering mechanism to expand the matching window, thereby enhancing association stability. Likewise, other works (Xu et al., 2024a; Ma et al., 2024b; Liu et al., 2025) refine static association paradigms from the perspectives of similarity metrics and multi-hypothesis tracking. Although these approaches yield measurable improvements, they fundamentally rely on static association strategies, which inherently limit further gains in association accuracy and robustness. While these efforts have partially alleviated certain challenges in UAV-based MOT, two fundamental limitations persist: First, most existing approaches do not fully exploit cross-frame spatiotemporal cues or implement global temporal modeling of occlusions and background clutter, resulting in only marginal improvements for small-object tracking. Second, prevailing association strategies rely heavily on high-precision detections and ReID match scores but ignore the dynamic confidence information produced by modern detectors, instead employing static cost matrices. As a result, these methods struggle to achieve both high accuracy and robust performance in complex aerial scenarios.

To address these issues, we present a Mutual-enhanced Occlusion-aware Spatiotemporal Adaptive Identity Conservation Framework. First, a novel spatiotemporal fusion mechanism effectively resolves occlusion-induced identity switching. Then, an adaptive multi-scale perception architecture enhances small object tracking capability. Finally, a confidence-driven optimization paradigm establishes more reliable data association for aerial imagery. The main contributions are threefold:

(1)Spatiotemporal Occlusion Enhancement (STOE). It fuses global cumulative features—aggregated across past frames—with dynamically changing features from the current frame. The global features suppress background clutter, reducing ID switches, while temporal features recover information lost during occlusion.

(2)Adaptive Multi-scale Feature Enhancement (AMFE). It unites a Local Enhancement Mechanism (LEM) for fine-grained spatial refinement with a Global Multi-scale Aggregation Module (GMAM) that hierarchically fuses features at multiple resolutions. This combination markedly improves the representation and tracking of small, densely distributed objects.

(3)Dynamic Confidence Matrix Adjustment (DCMA). It injects per-detection confidence scores into the IoU and ReID association matrices, up-weighting reliable matches and down-weighting uncertain ones. By adapting association costs in real time, It delivers more accurate and robust trajectory linking.

Extensive experiments on VisDrone2019 and UAVDT demonstrate that our framework outperforms state-of-the-art MOT methods—achieving significant gains in HOTA, MOTA, and IDF1—by effectively balancing occlusion resilience, small-object perception, and dynamic association.

The remaining structure of this paper is organized as follows. Section 2 provides a review of related work on generic object tracking and small object tracking in remote sensing. The detailed structure of the proposed MOSAIC-Tracker method are presented in Section 3. The experimental results and analysis are discussed in Section 4. Finally, Section 5 summarizes the contributions of this paper.

2. Related work

This section reviews methods in generic object and remote sensing small object tracking. While generic tracking addresses challenges like appearance variation, small object tracking in remote sensing faces unique difficulties, like occlusion, low resolution and dynamic viewpoints. We highlight key advancements and limitations in tackling these issues.

2.1. Overview of generic domain object tracking

Generic-domain MOT methods offer foundational insights and guidance for remote-sensing small-object tracking. These approaches fall into two main paradigms: Joint Detection and Tracking (JDT), which integrates detection and tracking within a single network to enhance efficiency, and Tracking by Detection (TBD), which decouples detection and data association to achieve higher accuracy. Guan et al. (2025) offer a unified survey of the two principal MOT paradigms and release an open-source toolkit that accelerates entry into the field.

With the JDT paradigm, detection and tracking are integrated into a unified network, which reduces computational resources, thus improving tracking efficiency. Gao et al. (2025) reformulate tracking as an ID-prediction task by decoding identity labels directly from detections. It reduces model complexity but delivers lower accuracy and robustness than TBD approaches. Xu and Huang (2024) separate detection and ReID into distinct feature streams, then perform hierarchical matching. However it fails to generate consistent tracks on the VisDrone dataset, where small target sizes and frequent occlusions prevail. Although JDT models enjoy low parameter counts and fast inference, they sacrifice representational capacity. To alleviate this, several studies introduce auxiliary modules: attention and contextual encoders (Wang et al., 2021; Ma et al., 2022); a visualization branch for per-object occlusion estimation (Lv et al., 2023); and a channel-attention block with a DIoU-based similarity metric (He et al., 2024). These enhancements improve robustness, yet fail to resolve the conflict between feature extraction and object localization. Thus, it is important to develop hybrid architectures that better balance efficiency, representational capacity, and robustness for aerial small-object tracking.

The TBD paradigm decouples object detection and data association, combining state-of-the-art detectors—such as YOLO, SSD, and

DETR—with efficient trackers like StrongSORT (Du et al., 2023), and ByteTrack (Zhang et al., 2022). In the case of ByteTrack, it employs a detect-then-associate paradigm, combining Kalman filtering for motion prediction and the Hungarian algorithm for data association. The TBD paradigm effectively addresses challenges like occlusion and interference from similar objects, leading to improved detection and tracking outcomes. To further enhance robustness and speed, recent works introduce targeted optimizations. Li et al. (2024d) jointly learn single- and multi-shot features to strengthen association under occlusion, whereas Li et al. (2024b) fuse global and local spatiotemporal contexts for better multi-scale embedding and reduced mismatches. Addressing computational constraints, Hu et al. (2024b) propose a multitask learning framework that parallelizes detection and feature embedding, significantly accelerating inference.

Despite these gains, generic TBD methods remain insufficient for remote sensing object tracking. The unique challenges of small targets, dense clustering, and persistent occlusion in aerial imagery require specialized solutions—including improved small-object detection modules, reliability-aware association strategies, and efficient spatiotemporal fusion—to achieve high precision and real-time performance in remote sensing applications.

2.2. Remote sensing small object tracking methods

Research on small-object tracking in remote sensing scenes can be organized into three interconnected aspects. First, occlusion-robust tracking develops mechanisms to detect and recover targets that temporarily disappear under occlusion, ensuring continuity of the track. Second, feature extraction focuses on learning discriminative representations—often by integrating multi-scale appearance and motion cues—to distinguish small targets against complex backgrounds. Third, data association in MOT leverages temporal and spatial consistency to accurately link target detections into coherent trajectories. By combining robust occlusion handling, enriched feature representations, and reliable trajectory linking, these methods jointly improve tracking accuracy and resilience in challenging remote sensing environments.

2.2.1. Spatiotemporal tracking for occlusion

Occlusion-robust tracking in remote-sensing MOT relies critically on fusing information across frames to recover targets when single-frame cues fail. Traditional methods, such as optical flow and multi-frame averaging, offer partial solutions but remain limited in complex scenarios. Deep spatiotemporal fusion methods (Hu et al., 2023; Lei et al., 2024; Zhang et al., 2024b, 2020) can be grouped into memory-based schemes, clustering-augmented approaches, and topology-driven techniques, each seeking to leverage temporal continuity under heavy occlusion.

Memory-based schemes retain and update feature embeddings over time to preserve identity. Yuan et al. (2025) integrate short- and long-term memory modules with a Time-Enhanced Feature Decoder (TEFD), boosting MOTA by decoding richer temporal representations, but their performance degrades when targets occupy only a few pixels, leading to residual ID switches. Liu et al. (2024a) fuse multi-frame embeddings to improve robustness against rapid motion; however, low frame rates, motion blur (Zhuang et al., 2023), and UAV jitter undermine their encoder's stability, resulting in tracking drift. Ma et al. (2024a) enhance temporal cohesion via historical embedding modules that propagate object features across occlusions, yet their reliance on global memory banks incurs non-negligible latency and may blur fine-grained small-object details. Wang et al. (2024b) further enhance frame-to-frame continuity by integrating motion modules with hybrid scale-enhancement blocks, yielding finer multi-scale small-object tracking.

Clustering-augmented approaches refine associations by grouping spatiotemporal proposals. Zheng et al. (2022) apply DBSCAN filtering to motion-augmented detections, reducing false matches and aiding occlusion recovery; nonetheless, its high computational complexity limits real-time UAV applicability. Topology-driven techniques explicitly construct local spatiotemporal graphs to enforce smooth trajectories. Zhang et al. (2023b) build a dynamic neighborhood topology that suppresses mismatches and aids recovery from occlusion, but in densely packed scenes with irregular small-object motion, the graph structure can become unstable, leading to fragmented tracks.

Although each category advances occlusion handling—by preserving temporal embeddings, refining matches through clustering, or enforcing local topologies—they share drawbacks of computational overhead, sensitivity to sparse or blurred inputs, and instability under dense occlusion. In response, our method employs a lightweight spatiotemporal fusion mechanism that dynamically aggregates frame-level features without heavy memory or clustering operations, ensuring both robust ID preservation and real-time performance under occlusion.

2.2.2. Feature extraction for tracking

Small-object feature extraction in UAV-based MOT has attracted considerable attention, yet existing approaches remain constrained by trade-offs between representational richness, computational cost, and adaptability to scale and motion variations. And existing approaches can be broadly grouped into three categories: Transformer-based feature learning, efficient spatiotemporal fusion, and scale- and motion-guided enhancements.

Transformer-based architectures (Zhu et al., 2023) —exemplified by Xu et al. (2024a) and Hu et al. (2024a) have demonstrated superior multi-level, multi-scale feature learning, significantly boosting detection and Re-ID performance in cluttered UAV footage. However, their heavy self-attention modules and extensive parameterization introduce prohibitive inference delays, undermining real-time tracking requirements in resource-limited aerial platforms. To reduce this burden, convolutional spatiotemporal fusion has been proposed. Zhu et al. (Zhu et al. 2023) construct a multi-level similarity graph that fuses shallow motion cues with deep semantic features, offering a lighter alternative to full self-attention. Lin et al. (2024) further integrate historical trajectory priors via a motion-aware filtering framework, enhancing temporal consistency but still struggling with sudden scale variations. Rao et al. (2024) deploy cascaded conventional feature extractors for dynamic cue extraction, yet their static pipelines exhibit limited adaptability across diverse target sizes and backgrounds.

Other efforts have focused on scale- and motion-guided enhancements. Li et al. (2025b) propose a multi-branch mutual-guiding learning network with a lossless encoder to mitigate feature degradation in infrared small-object detection (Li et al., 2024c), but its thermal-domain specialization may not generalize to visible-spectrum UAV imagery. Zhang et al. (2024a) employ an optical-flow-guided feature extractor to improve motion coherence, reducing drift yet overlooking multi-scale spatial representation. Rahman et al. (2024) have proposed a effective feature-extraction module that holds potential for adaptation to the object-tracking. Chen et al. (2023) integrate hierarchical aggregation modules to refine local feature localization, but the added fusion layers inevitably raise latency and computational demand. Wu et al. (2024) develop HRTracker, employing high-resolution feature fusion to enhance target detail representation in satellite videos coupled with an adaptive data association strategy that significantly reduces false matches. MM-Tracker (Wang et al., 2024a) jointly trains detection and discriminative feature extraction in a multi task framework, significantly enhancing adaptability to appearance changes.

Although these approaches improve isolated facets of small-object representation—be it semantic richness, temporal context, or scale handling—they lack a unified, real-time solution that adapts dynamically to scale and motion changes while preserving discrimination under heavy occlusion. Our Adaptive Multi-scale Feature Enhancement (AMFE) module remedies this by fusing a lightweight Local Enhancement Mechanism with a Global Multi-scale Aggregation Module, striking an effective balance between computational efficiency and representational power for robust UAV-based tracking.

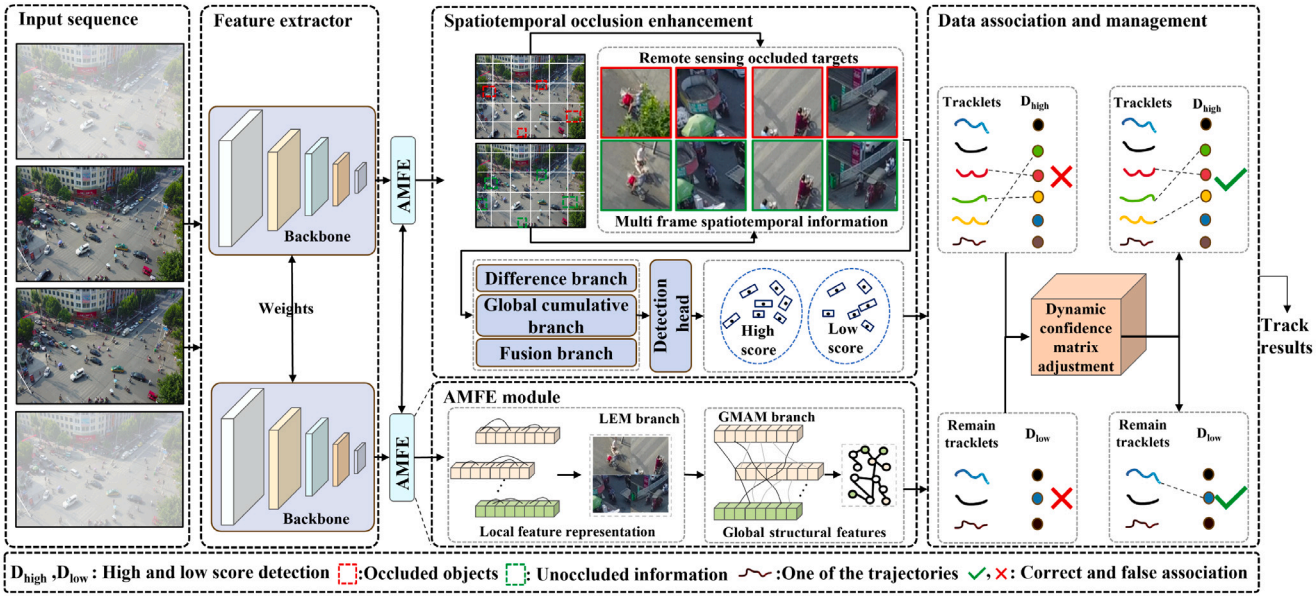

Fig. 2. Overview of our proposed tracking method. It includes Adaptive Multi-scale Feature Enhancement(AMFE) for efficient multi-scale feature extraction, Spatiotemporal Occlusion Enhancement (STOE) to restore occluded object features via multi-frame fusion, and the Dynamic Confidence Matrix Adjustment (DCMA) to improve data association by incorporating detection confidence. The rightmost panel schematically depicts the DCMA process, where some erroneous correspondences (×) are rectified into validated matches (✓).

2.2.3. Data association in MOT

Data association is a critical component of multi-object tracking, aiming to correctly link objects across frames and assign consistent identities. Classical approaches—e.g., Kalman and particle filters—rely primarily on motion cues, while IoU-based matching utilizes spatial overlap. However, in the context of remote-sensing small-object tracking, these static association strategies fail to exploit the rich temporal and confidence information inherent in UAV imagery, leading to frequent ID switches and fragmented trajectories.

Recent efforts have sought to enrich association cues, yet they remain essentially static in nature. Liu et al. (2025) introduce an IoU-guided Siamese network that fuses template confidence scores to refine matching; nonetheless, the network's weights are fixed once trained and cannot adapt to frame-by-frame variations. Liu et al. (2022) propose an ID-feature update module and an adaptive motion filter to handle UAV motion and perspective changes, but their association still treats confidence and motion cues separately and statically. Wu and Liu (2024) compensate for irregular camera movements via pose estimation in DepthMOT, reducing occlusion-induced ID switches, yet they do not adjust matching thresholds according to detection reliability. Yasir et al. (2024) expand the matching search space with C-BIoU, improving robustness but overlooking per-trajectory confidence dynamics. To partially remedy this, Wang et al. (2025) employ an adaptive exponential moving average to weight trajectory hypotheses by historical confidence, and Xu et al. (2024b) further incorporate temporal confidence memory into association. Although these methods introduce elements of adaptivity, they still operate under a static matching framework whose parameters are globally fixed, limiting their responsiveness to abrupt appearance changes or occlusions. Graph-matching formulations (e.g., Ding et al. (2025)) offer more flexible association at the cost of real-time efficiency.

In contrast, our Dynamic Confidence Matrix Adjustment (DCMA) module departs from these static paradigms by integrating detection confidence directly into the association cost matrix on a per-frame basis. By dynamically up-weighting high-confidence matches and down-weighting ambiguous ones, DCMA adapts continuously to varying imaging conditions.

3. Overview of the proposed method

This section provides an overall overview of the proposed method, followed by a detailed description of its components, including the Shared Feature Extractor, Adaptive Multi-scale Feature Enhancement Module, Spatiotemporal Occlusion Enhancement Module, and the Data Association method.

3.1. Overall framework

The proposed multi-object tracking network consists of three key components: Adaptive Multi-scale Feature Enhancement (AMFE), Spatiotemporal Occlusion Enhancement (STOE), and Dynamic Confidence Matrix Adjustment. As shown in Fig. 2. These components address small object feature extraction, occlusion recovery, and data association optimization, respectively. To achieve robust feature representation, it employs a shared-weight feature extractor built upon a backbone network. This shared-weight mechanism not only reduces computational complexity but also enhances temporal coherence across frames. Throughout the tracking pipeline, MOSAIC-Tracker dynamically adjusts feature extraction and association strategies to accommodate targets of varying scales and occlusion levels. First, the STOE module employs a cross-frame spatiotemporal attention mechanism to detect occlusion risk and motion changes in real time, adaptively reinforcing feature representations to maintain identity consistency during occlusion. Second, the AMFE module integrates a Local Enhancement Mechanism (LEM) with a Global Multi-scale Aggregation Module (GMAM), allocating adaptive weights across feature maps of different resolutions. This enables simultaneous refinement of small-object details and preservation of global semantics, thereby establishing a robust foundation for subsequent data association. Finally, the DCMA strategy incorporates detection confidence into the similarity matrix, dynamically reweighting matches to mitigate errors arising from uniform association costs. By lowering the influence of low-confidence detections and amplifying high-confidence ones, DCMA reduces ID switches and enhances overall association accuracy. Collectively, these modules enable MOSAIC-Tracker to adaptively extract critical features and preserve object identity in challenging aerial scenarios.

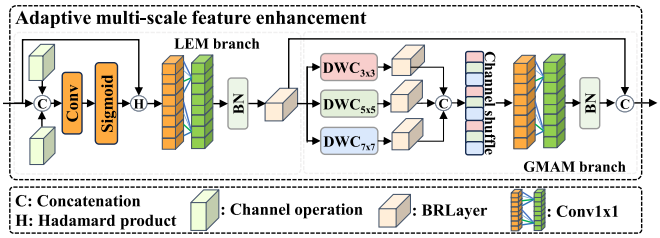

Fig. 3. Illustration of Adaptive Multi-scale Feature Enhancement (AMFE) Module.

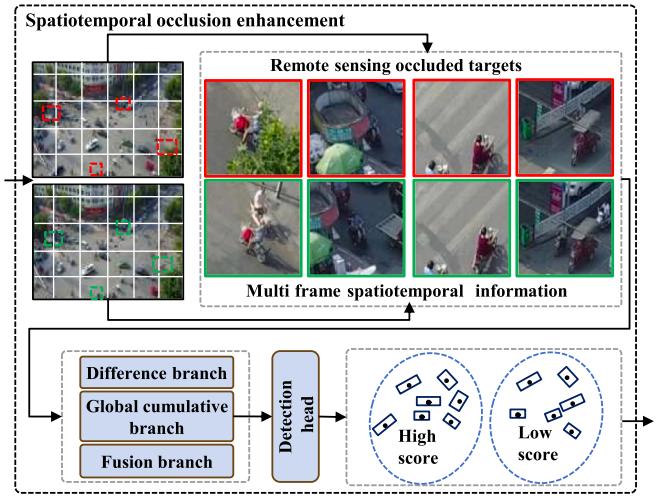

Fig. 4. Illustration of Spatiotemporal Occlusion Enhancement(STOE) module. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

3.2. Adaptive multi-scale feature enhancement

In two-stage object tracking for remote sensing, small objects face challenges such as low signal-to-noise ratio, weak feature representation, and difficulties with multi-scale information fusion. To address these issues, we propose the Adaptive Multi-scale Feature Enhancement (AMFE) module, designed to improve the feature representation of small objects in complex remote sensing imagery.

As shown in Fig. 3, it uses a hierarchical fusion strategy, combining the Local Enhancement Mechanism (LEM) and the Global Multi-scale Aggregation Module (GMAM) to enhance small object feature extraction. The LEM extracts salient regions from input feature maps by performing inter-channel aggregation with average and max pooling operations. This generates a two-dimensional attention map that captures both global and local feature responses. The map is then processed through a convolutional layer followed by a Sigmoid activation to produce spatial attention weights. This design effectively suppresses background noise, enhances object regions, and highlights small objects in cluttered environments.

Following the enhancement by the LEM, the extracted features are passed to the GMAM for multi-scale convolutional processing. It captures object features at different scales by leveraging convolutional kernels of varying sizes, enabling a more comprehensive representation of objects with diverse spatial characteristics. The feature extraction process incorporates stacked Pointwise Convolution and Depthwise Separable Convolution, facilitating both inter-channel interactions and spatial feature extraction. To ensure effective utilization of multi-scale information, it employs a parallel convolution operation strategy combined with a channel rearrangement-based fusion mechanism, which further enhances feature expressiveness while maintaining computational efficiency.

To ensure effective utilization of multi-scale information, the GMAM employs a parallel convolution operation strategy combined with a channel rearrangement mechanism for feature fusion. The resulting fused features are then passed through a channel shuffle operation to enable inter-channel interactions. The final feature representation after enhancement by both LEM and GMAM is processed through a batch normalization layer BN, and the output is obtained. This comprehensive approach enhances the representation of small objects by extracting multi-scale features efficiently while maintaining computational effectiveness.

3.3. Spatiotemporal occlusion enhancement

In remote sensing object tracking tasks, relying solely on single-frame information is often insufficient to achieve robust tracking, particularly for small objects that are susceptible to complex backgrounds and occlusions. To address this challenge, we propose a Spatiotemporal Occlusion Enhancement (STOE) module, which leverages inter-frame correlations to fully exploit temporal information and restore object features after occlusion through multi-frame fusion. This approach significantly improves the tracking accuracy and robustness of small objects in remote sensing imagery. In remote sensing datasets, the state of objects—such as position, shape, and appearance—is influenced not only by the static characteristics within individual frames but also by the dynamic relationships across consecutive frames. These relationships can be categorized into two key feature types: (1) Global cumulative features, which represent the stable and invariant characteristics of the tracked object and the surrounding scene; (2) Dynamic change features, which capture the evolving characteristics of the object under motion, deformation, or environmental changes.

A key challenge lies in comprehensively incorporating both feature types, especially in complex backgrounds or occluded scenarios, to fully utilize temporal information for robust tracking. Thus, it employs a multi-frame feature recovery mechanism, which collaboratively models global cumulative features and dynamic change features to tackle occlusion-related issues. Specifically, it first aligns consecutive frames to mitigate feature inconsistencies caused by object motion or background variations. Following alignment, it separately extracts global cumulative features and dynamic change features. In occlusion scenarios, the global cumulative features help preserve the long-term stable characteristics of objects, while the dynamic change features enable the recovery of occluded object details once they reappear. This dual-feature modeling approach effectively restores the complete representation of occluded objects, ensuring smooth tracking.

$$\begin{cases} F^t_{aligned} = \mathcal{A}(F^t, F^{t-1}), \\ Q, K, V = W_q * F^t_{aligned}, W_k * F^{t-1}, W_v * F^{t-1}, \\ A_{temporal} = \mathrm{softmax}(Q\,K^T), \\ F^t_{global} = \gamma \cdot (A_{temporal}\,V) + F^t_{aligned}, \end{cases} \tag{1}$$

The global cumulative branch captures the stable, invariant characteristics of the object and its surrounding scene across frames. As shown in the formula (1), the current frame feature F_t is aligned with the previous frame feature F_{t-1} using an alignment operator $A(\cdot)$, yielding $F_{\mathrm{aligned},t}$. This aligned feature is projected into query space via W_q, while F_{t-1} is projected into key and value spaces using W_k and W_v, respectively. The temporal attention map A_{temporal} is computed as a softmax over the similarity between Q and the transpose of K, capturing inter-frame correlations. Finally, the attention-enhanced feature $F_{\mathrm{att},t}$ is obtained by scaling the attended value $(A_{\mathrm{temporal}}V)$ with a learnable parameter γ and adding the aligned feature $F_{\mathrm{aligned},t}$. This process robustly extracts the global cumulative features.

$$\begin{cases} F^t_{diff} = \mathrm{ReLU}\Big(\mathrm{BN}_1\big(\mathrm{Conv}_1\big(|F^t - F^{t-1}|\big)\big)\Big), \\ F^t_{out} = \mathrm{ReLU}\Big(\mathrm{BN}_2\big(\mathrm{Conv}_2\big(F^t_{global} + F^t_{diff}\big)\big)\Big) + \\ \big(F^t + F^{t-1}\big), \end{cases} \tag{2}$$

The difference branch first calculates the absolute difference $|F_t - F_{t-1}|$ between consecutive frames to capture the evolving characteristics of objects, such as motion, deformation, or appearance changes. As shown in the formula The difference map is refined through a convolutional layer Conv_1, followed by batch normalization BN_1 and ReLU activation, resulting in the dynamic change feature $F_{\text{diff},t}$. The final enhanced feature $F_{\text{out},t}$ is then generated by fusing the global cumulative features $F_{\text{att},t}$ with the dynamic change features $F_{\text{diff},t}$ via another convolution Conv_2 and batch normalization BN_2 (again activated by ReLU). A residual connection $(F_t + F_{t-1})$ is added to preserve the overall object representation and ensure continuity.

Difference branch focuses on the pixel-level discrepancy between two feature maps by computing $|x_1 - x_2|$ to highlight regions of significant change. The resulting difference map is processed through convolution, batch normalization, and ReLU activation to extract local variation features. This design enables the network to be particularly sensitive to object motion or shape deformation in dynamic scenes, providing critical supplementary information for overall feature enhancement.

Global cumulative branch employs a temporal attention mechanism to accumulate the global context of the second frame into the first. Specifically, 1×1 convolutions generate the query, key, and value tensors, after which the attention weights are computed. The value features are then weighted and added back to the first frame's features, with a learnable scaling factor γ introduced to adaptively balance the original and accumulated information. This mechanism captures long-range dependencies between frames, establishing cross-temporal relationships over a broad receptive field and thereby enhancing the model's understanding of complex motion patterns.

After local difference extraction and global accumulation, Fusion branch first sums the two feature sets to achieve information complementarity. It then applies a convolution followed by activation to deeply abstract the fused features. Finally, the common features of both frames are added back to the output via a residual connection. This three-step process preserves both local variation and global context while integrating the original temporal features, thereby enhancing representational capacity without sacrificing informational completeness or training stability.

In practical remote sensing applications, the STOE module integrates low-level spatiotemporal features with high-level semantic information. As illustrated in Fig. 4, it handles occlusion scenarios involving objects such as motorcycles and pedestrians, highlighting its ability to restore missing object features. For example, when an object is partially or completely occluded by foreground elements (e.g., vehicles or buildings) in earlier frames, it may reappear in subsequent frames. Through multi-frame information alignment and fusion, it compensates for the limitations of single-frame detection, ensuring the continuity and integrity of objects in tracking tasks. Typical scenarios include (1) Motorcycle tracking: As shown in Fig. 4, when a motorcycle is partially occluded by roadside trees in earlier frames (indicated by the red box), the STOE uses both global and dynamic features to restore the object representation when it reappears in the subsequent frame (indicated by the green box). (2) Pedestrian tracking: Similarly, when a pedestrian is temporarily blocked by a passing vehicle, it identifies the occluded region and reconstructs the feature of pedestrian, ensuring tracking.

Overall, it enhances the robustness of remote sensing object tracking by addressing occlusions through multi-frame feature fusion. It provides reliable tracking in complex and dynamic environments.

3.4. Data association in MOT

Multi-Object Tracking (MOT) methods, such as SORT, DeepSORT, and Bot-SORT, rely heavily on detection results provided by object detectors. These methods use Kalman filters for state estimation and the Hungarian algorithm for data association. In MOT, it is essential to estimate each target's position and velocity at discrete time steps k in real time. The Kalman filter achieves this by combining a dynamic system state model with an observation model under the assumption of Gaussian noise, thereby providing the optimal real-time estimate of the object's motion state. The recursive process of Kalman filtering mainly includes two major stages: prediction and update. In the prediction stage, it recurses the current state based on the optimal estimation of the previous moment, and the state prediction is as follows.

$$\hat{x}_{k|k-1} = A\hat{x}_{k-1|k-1} \tag{3}$$

where A is state transition matrix, $\hat{x}_{k-1|k-1}$ is the updated state from time step $k-1$, $\hat{x}_{k|k-1}$ is the predicted state from time step k. P is the estimation covariance matrix quantifying state uncertainty:

$$P_{k|k-1} = AP_{k-1|k-1}A^T + Q \tag{4}$$

where $P_{k-1|k-1}$ is the updated covariance from time step $k-1$, $P_{k|k-1}$ is the predicted covariance from time step k, Q is process noise covariance.

During the update stage, it corrects the prediction results by using the current observed values. Firstly, the Kalman gain K_k makes a trade-off between the state prediction uncertainty and the measurement noise uncertainty. Its calculation formula is as follows.

$$K_k = P_{k|k-1}H^T \left(HP_{k|k-1}H^T + R\right)^{-1} \tag{5}$$

where R is measurement noise covariance. Secondly, the status update formula is as follows.

$$\hat{x}_{k|k} = \hat{x}_{k|k-1} + K_k \left(z_k - H\hat{x}_{k|k-1}\right) \tag{6}$$

where z_k is measurement vector. The covariance update formula is as follows.

$$P_{k|k} = (I - K_k H)P_{k|k-1} \tag{7}$$

where H is observation matrix.

However, their performance often degrades in challenging scenarios, such as low-confidence detections, object occlusion, and reappearance after occlusion. To address these limitations, this study proposes a dynamic confidence matrix adjustment method to enhance object association reliability and improve overall tracking performance in complex environments.

In the proposed approach, an object detector processes the input video sequence to generate a set of detection bounding boxes, denoted as $D = \{d_1, d_2, \ldots, d_n\}$. Each bounding box consists of coordinates and an associated confidence score, represented as $d_i = (x, y, w, h, s)$, where (x, y) denotes the position, (w, h) represents the width and height, and s is the detection confidence score. The Kalman filter then predicts the state of the object trajectories $T = \{t_1, t_2, \ldots, t_m\}$, producing a set of predicted trajectory positions $T^{\text{pred}} = \{t_1^{\text{pred}}, t_2^{\text{pred}}, \ldots, t_m^{\text{pred}}\}$. Each predicted trajectory is assigned a confidence score based on tracking history and observation consistency.

The proposed method enhances object association accuracy by incorporating two association mechanisms and Dynamic Confidence Matrix Adjustment(DCMA) method that work together to optimize trajectory consistency and robustness. The pseudocode for the algorithm is shown below.

3.4.1. Two primary association

In the primary association stage, these detection boxes are denoted as $D_{\text{high}} = \{d_j \mid s_j > \tau\}$, where s_i represents the confidence score of the detection box d_i, and τ is the predefined threshold. The similarity between a detection box and a predicted trajectory is determined by combining Intersection over Union (IoU) and Re-Identification (ReID) feature similarity, ensuring both spatial and appearance consistency. IoU measures the spatial overlap between the detection box and the predicted trajectory, and it is defined as follows:

$$\text{IoU}_{i,j} = \frac{\text{Area}B(t_i) \cap B(d_j)}{\text{Area}B(t_i) \cup B(d_j)}, \tag{8}$$

Algorithm 1 Enhanced Two-Stage Data Association

Require: Detection set D, Trajectories T, Threshold τ
Ensure: Updated trajectories T
Prediction:
for each $t_i \in T$ **do**
  $t_i^{\text{pred}} \leftarrow \text{KalmanPredict}(t_i)$
end for
Data Association:
$D_{\text{high}} \leftarrow \{d_i | s_i > \tau\}, D_{\text{low}} \leftarrow D \setminus D_{\text{high}}$
Compute similarity matrix X_1 using Eq.(10) for T^{pred} and D_{high}
$\mathcal{M}_1, \mathcal{U}_T, \mathcal{U}_D \leftarrow \text{Hungarian}(X_1)$
Update matched trajectories: $T \leftarrow \text{KalmanUpdate}(\mathcal{M}_1)$
Compute IoU matrix X_2 between $\mathcal{U}_T$ and D_{low}
$\mathcal{M}_2 \leftarrow \text{Hungarian}(X_2)$
Update remaining trajectories: $T \leftarrow T \cup \text{Update}(\mathcal{M}_2)$
Dynamic Adjustment:
Construct confidence matrix $S \leftarrow [s_1 \mathbf{1}_m; ...; s_n \mathbf{1}_m]^T$
Compute weight matrix W using Eq. (12)
Enhance cost matrix: $X' \leftarrow X - W \odot S$
Final matching: $\mathcal{M}_{\text{final}} \leftarrow \text{Hungarian}(X')$
for each unmatched detection **do**
  Initiate new trajectory if criteria met
end for
Remove trajectories with consecutive misses $> N_{\max}$

where $B(d_j)$ and $B(t_i)$ represent the detection box and the trajectory box, respectively, and $\cap$ and $\cup$ denote their intersection and union areas. A higher IoU value indicates a greater degree of spatial overlap between the detection and trajectory boxes. From this, we derive the spatial association cost distance as $1 - \text{IoU}_{i,j}$.

In addition to spatial similarity, appearance similarity is evaluated using feature vectors extracted by a deep neural network for ReID. The similarity between two feature vectors, f_i and f_j, is measured using cosine similarity, defined as:

$$\text{Sim}_{\text{ReID}}(f_i, f_j) = \frac{f_i \cdot f_j}{\|f_i\| \|f_j\|}, \tag{9}$$

where $f_i \cdot f_j$ denotes the dot product of the feature vectors, and $\|f_i\|$, $\|f_j\|$ represent their L2 norms. The similarity score ranges from $[-1, 1]$, with values closer to 1 indicating greater similarity in the characteristics of the appearance. The appearance affinity between track i and detection j is measured in *distance space* for compatibility with matching frameworks. The *cosine distance* is computed as $1 - \text{Sim}_{\text{ReID}}(f_i, f_j)$. where lower values indicate higher similarity.

The overall cost matrix between a detection box and a trajectory is computed using a combination of IoU cost distance and ReID cosine distance as follows:

$$X_1(i, j) = min(1 - \text{IoU}_{i,j}, 1 - \text{Sim}_{\text{ReID}}(f_i, f_j)) \tag{10}$$

By adjusting these parameters, the association process can be fine-tuned to accommodate different application scenarios. After computing the cost matrix X_1, the Hungarian algorithm is applied to obtain three sets: $\mathcal{M}_1$, which contains successfully matched trajectory-detection pairs, $\mathcal{U}_T$, which consists of unmatched trajectories, and $\mathcal{U}_D$ which consists of unmatched detections.

The secondary association phase addresses the relationship between low-confidence detection boxes and unmatched trajectories. The Hungarian algorithm is applied again, resulting in a new set of matches, denoted as $\mathcal{M}_2$. In this stage, matching is performed using IoU, calculated similarly matrix X_2 to the primary association phase. However, the appearance similarity based on ReID features is no longer considered. For successfully matched trajectories, their positions and velocity states are updated to ensure tracking continuity. Conversely, unmatched trajectories across consecutive frames are marked as lost if the number of successful matches falls below a predefined threshold $N_{\max}$. Additionally, low-confidence detection boxes D_{low} that meet the criteria for initiating new trajectories are initialized as new tracks for subsequent tracking.

3.4.2. Dynamic confidence matrix adjustment

Following the secondary association phase, a cost matrix X is constructed to represent the matching confidence between detection results and trajectories. It can be selectively applied to refine the outcomes of either primary or secondary matching. Specifically, X is an $m \times n$ matrix, where m represents the number of trajectories and n represents the number of detections, i.e., $X \in \mathbb{R}^{m \times n}$. Each element in the matrix denotes the confidence score between the ith trajectory and the jth detection result. Conventional methods, however, often treat all detection boxes equally, overlooking their varying confidence levels. Detection confidence is a crucial reference for trajectory matching. High-confidence detections are more stable across frames and can influence matching decisions, even when the similarity score from the Hungarian algorithm is low. Relying solely on IoU and ReID features to construct the similarity matrix X can lead to mismatches or suboptimal associations.

To address this, we propose a dynamic optimization approach that integrates detection confidence into the similarity matrix calculation. This method improves both accuracy and efficiency in matching by reducing errors caused by neglecting detection confidence. In this approach, the original cost matrix $X \in \mathbb{R}^{m \times n}$ represents the matching distance between each trajectory and all detections. The confidence vector is defined as $s \in \mathbb{R}^n$, where s_j denotes the confidence of the jth detection. To incorporate confidence, we expand the vector into a weight matrix $W \in \mathbb{R}^{m \times n}$, with each column replicated m times to represent the weighting of detections across all trajectories. The calculation is defined as follows:

$$W = \mathbf{1}_m \otimes \left(\frac{s}{\sum_{j=1}^{n} s_j} \right), \tag{11}$$

where $\mathbf{1}_m$ represents an m-dimensional column vector filled with ones. $\otimes$ denotes column-wise replication; $\frac{s}{\sum_{j=1}^{n} s_j}$ performs confidence normalization. Using the weight matrix W, the corrected IoU distance matrix X' is computed as follows:

$$x'_{i,j} = min(1 - \text{IoU}_{i,j}, 1 - \text{Sim}_{\text{ReID}}(f_i, f_j)) - w_{i,j} \cdot s_j, \tag{12}$$

where $w_{i,j}$ represents the normalized weight for the jth detection, and s_j is its confidence score.

$$X' = X - W \odot S, \tag{13}$$

where S is the expanded matrix of confidence values, with each column containing the corresponding detection confidence. $\odot$ denotes element-wise multiplication (Hadamard product).

The weight matrix W ensures that high-confidence trajectories have a stronger influence on the corrected similarity matrix, reflecting their higher reliability. By applying confidence-based weighting to each trajectory-detection pair, the corrected matrix X' more accurately represents the true matching relationships. Finally, the Hungarian algorithm is applied once more, yielding the final set of matched pairs, denoted as $\mathcal{M}_{\text{final}}$, which determines the ultimate data association results. Unlike the original similarity matrix X, the proposed method integrates detection confidence into the association optimization, allowing high-confidence detections to dominate the matching process. This significantly improves the accuracy and reliability of trajectory association.

As shown in Fig. 5, traditional data association methods use the Fused Confidence Distance Matrix for object matching after merging the ReID and IoU distance matrices. In contrast, the Dynamic Confidence Matrix Adjustment (DCMA) module refines this matrix by the individual confidence of each detection. It first generates a Fused

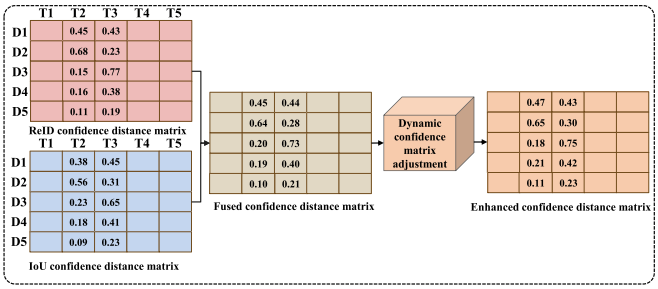

Fig. 5. Illustration of confidence-based Dynamic Confidence Matrix Adjustment (DCMA).

Confidence Distance Matrix by combining ReID and IoU matrices. Instead of directly using this fused matrix for data association, the module adjusts it by factoring in confidence score of detections. This produces the Enhanced Confidence Distance Matrix, which enhances tracking accuracy by accounting for the reliability of each detection. By incorporating detection confidence, the DCMA module improves data association, especially in challenging scenarios, leading to more reliable tracking performance in dynamic environments.

The confidence-adjusted IoU distance optimization method proposed in this study integrates detection confidence into the tracking framework, improving matching performance. In complex scenarios like object occlusion and disappearance, the method demonstrates greater robustness and scalability in multi-object tracking tasks.

3.5. Network loss

Our overall network builds upon the BotSORT tracker and follows the TBD paradigm. Within this paradigm, the loss function of the underlying YOLO detector critically determines the quality of detections—and thus the inputs to our tracker—by jointly optimizing classification and localization. We employ a composite loss:

$$\mathcal{L}_{\text{tot}} = \lambda_{cls}\mathcal{L}_{CLS} + \lambda_{box}\mathcal{L}_{BOX} + \lambda_{dfl}\mathcal{L}_{DFL} \tag{14}$$

where λ_{cls} is the classification loss weighting coefficient, this loss enhances localization robustness—particularly under occlusion or for small targets. λ_{box} is the localization loss weighting coefficient, It therefore ensures reliable category prediction even when positive samples are scarce. λ_{dfl} is the distribution focal loss weighting coefficient. This term controls the focal distribution's convergence dynamics, enhancing training stability. We adopt the default weights are 7.5,0.5 and 1.5. The specific analysis of the three loss functions is shown as follows. Box Loss focuses on the regression accuracy of bounding box coordinates and is implemented via Complete IoU (CIoU) optimization:

$$\mathcal{L}_{box} = 1 - \text{IoU} + \frac{\rho^2(b, b_{gt})}{c^2} + \alpha v \tag{15}$$

where $\rho^2(b, b_{gt})$ denote the predicted and ground-truth boxes, c is the diagonal length of the smallest enclosing box, and αv is a penalty term encoding aspect ratio dissimilarity. By jointly optimizing center point alignment, area overlap, and aspect ratio similarity, this loss enhances localization robustness—particularly under occlusion or for small targets. Then, we supervise object classification with a binary cross entropy loss:

$$\mathcal{L}_{cls} = -\sum_{i=1}^{N}\sum_{j=1}^{C}\left[y_{ij}\log(p_{ij}) + (1 - y_{ij})\log(1 - p_{ij})\right] \tag{16}$$

where $y_{ij} \in \{0, 1\}$ denotes the ground truth label, and p_{ij} is the model's predicted probability that the ith sample belongs to class j. This loss term ensures that the model correctly distinguishes foreground objects from background noise, thereby reducing identity switches in tracking scenarios. The Distribution Focal Loss (DFL Loss) is a distribution-based learning loss for discretized bounding-box modeling, defined as:

$$\mathcal{L}_{\text{df1}} = -\sum_{k=1}^{K} y_{\text{dist}}^{k} \log\left(p_{\text{dist}}^{k}\right) \tag{17}$$

where K is the number of bins, y_{dist} is the ground truth bounding box distribution after Gaussian smoothing, and p_{dist} is the predicted distribution. By modeling the box coordinates as a probability distribution instead of direct regression, this loss enhances localization sensitivity for small-scale objects.

4. Experiment and analysis

In this section, we introduce the datasets, evaluation metrics, and experimental setup, followed by a comparison with advanced methods. We also present visual tracking examples for both datasets and analyze the effectiveness of each module.

4.1. Datasets and evaluation metrics

For the remote sensing object tracking task in this study, we selected the VisDrone2019 (Du et al., 2019) and UAVDT (Du et al., 2018) datasets. The details of these datasets are as follows:

- **VisDrone2019** consists of 288 video sequences and 10,209 static images, covering 10 object categories, including pedestrians, vehicles, and bicycles, in both sparse and crowded scenes. Our training set includes 56 sequences, totaling approximately 24,000 images, while the test set consists of 17 sequences with about 7000 images.
- **UAVDT** contains approximately 80,000 annotated frames captured in diverse environments such as urban squares, highways, toll stations, and intersections. The videos are recorded at 30 FPS, with all frames having a resolution of 1080 × 540 pixels.

We adopt HOTA, MOTA, and IDF1 as the primary evaluation metric. HOTA balances detection, association, and localization performance, providing a more comprehensive assessment of tracking accuracy. MOTA is used to evaluate the overall performance of multi-object tracking methods. IDF1 measures identity consistency in multi-object tracking and is computed as the harmonic mean of Identity Precision and Identity Recall.

$$\text{HOTA} = \int_0^1 \sqrt{\text{DetA}(\alpha) \cdot \text{AssA}(\alpha)}\, d\alpha, \tag{18}$$

$$\text{MOTA} = 1 - \frac{\text{FP} + \text{FN} + \text{IDSW}}{\text{GT}}, \tag{19}$$

$$\text{IDF1} = \frac{2 \times \text{IDTP}}{2 \times \text{IDTP} + \text{IDFP} + \text{IDFN}}, \tag{20}$$

where DetA and AssA refer to detection accuracy and association accuracy, FN and FP refer to the number of missed detections and false detections, IDSW is the number of identity exchanges, and GT is the actual object quantity. IDTP represents correctly matched true objects with tracking results, IDFP denotes falsely matched objects, and IDFN refers to true objects that were not assigned to any tracking result.

4.2. Experiment setup

The experiment setup is configured with a learning rate of 0.0001 and a batch size of 4. Training is performed on a Tesla A6000 GPU and an Intel(R) Xeon(R) CPU E5-2680 v4 @ 2.40 GHz. The optimization pipeline employs the following hyperparameters: epoch of 10, a learning rate of 0.01, momentum of 0.937, and weight decay of 0.0005 to regularize model complexity. The "auto" optimizer dynamically adapts to the architecture, and the pretrained flag is enabled

Table 1
Comparison with advanced MOT methods on the VisDrone-MOT dataset. The best results are highlighted in bold and the second-best results are highlighted in underline format. The slight decline in MOTA primarily stems from our network that focuses on preserving object ID invariance, which yields significant improvements in HOTA and IDF1.

Method	HOTA ↑	MOTA ↑	IDF1 ↑	FN ↓	FP ↓	IDs ↓	MT ↑	ML ↓
SORT (Bewley et al., 2016)	35.080	33.148	42.844	112 481	20 392	3525	329	523
DeepSORT (Veeramani et al., 2018)	36.921	34.397	46.714	110 989	21 077	1784	386	517
FairMOT (Zhang et al., 2021)	31.102	12.806	37.745	114 834	59 997	3072	397	581
MOTR (Zeng et al., 2022)	–	22.800	41.400	147 937	28 407	959	272	825
ByteTrack (Zhang et al., 2022)	40.661	39.541	50.398	105 518	16 257	1581	507	538
ByteTrack+ReID (Zhang et al., 2022)	41.422	40.880	51.264	103 245	15 254	1512	510	525
BoT-SORT (Aharon et al., 2022)	42.420	**41.652**	<u>56.843</u>	103 505	14 114	1430	543	537
UAVMOT (Liu et al., 2022)	38.133	38.685	45.150	108 134	13 610	3357	463	557
TrackFormer (Meinhardt et al., 2022)	35.344	25.000	51.000	141 526	25 856	1534	515	946
OCSORT (Cao et al., 2023)	–	39.600	50.400	123 513	14 631	986	–	–
STCMOT (Ma et al., 2024a)	–	41.200	52.000	94 445	36 428	3984	667	453
MGTrack (Ren et al., 2024)	–	32.800	44.800	70 346	15 236	1800	–	–
AdaptTrack (Li et al., 2024a)	42.900	41.300	53.600	106 315	13 965	1426	–	–
UCMCTrack (Yi et al., 2024)	37.072	28.114	38.434	150 590	9244	7213	180	826
DepthMOT (Wu and Liu, 2024)	42.448	37.041	54.023	104 054	41 001	1248	626	467
TrackTrack (Shim et al., 2025)	42.530	37.958	50.956	135 375	51 351	5120	526	656
TStsm (Zhang and Ye, 2025)	<u>43.892</u>	41.256	54.637	–	–	878	–	–
Ours	**46.044**	<u>41.358</u>	**58.142**	126 112	26 895	1862	583	591

throughout to leverage transfer learning. Joint optimization is performed across all components using the aforementioned loss functions, with no resumption from previous checkpoints (resume=False). This configuration ensures deterministic training while mitigating overfitting through L2 regularization (weight decay) and momentum-based gradient updates.

In the data association phase, improvements were made based on the Bot-SORT (Aharon et al., 2022) as the baseline.

4.3. Comparison with advanced methods

In our comprehensive experiments, we conduct extensive comparisons with over ten advanced tracking methods across two benchmark datasets. Our benchmark includes both foundational methods and recent advancements: SORT (Bewley et al., 2016) as the baseline tracking framework, DeepSORT (Veeramani et al., 2018) with deep association metrics, FairMOT (Zhang et al., 2021) for joint detection-reID optimization, and the transformer-based architecture of MOTR. We further evaluate against high-performance methods including ByteTrack (Zhang et al., 2022) and its ReID-enhanced variant (Zhang et al., 2022), BoT-SORT (Aharon et al., 2022), UAVMOT (Liu et al., 2022) for aerial scenarios, and TrackFormer (Meinhardt et al., 2022) with transformer-decoder design. Special focus is given to recent innovations like: OCSORT (Cao et al., 2023) with observation-centric recovery, STCMOT (Ma et al., 2024a) employing spatiotemporal context mining, and DepthMOT (Wu and Liu, 2024) incorporating depth-aware fusion. We also include the latest 2025 advances in tracking algorithms—TStsm (Zhang and Ye, 2025), TrackTrack (Shim et al., 2025), DFA-MOT (Zheng et al., 2025), and SAM2MOT (Jiang et al., 2025)—as baselines. This diverse comparison allows for a thorough evaluation of tracking robustness across different architectural paradigms including correlation filters, ReID-based associations, transformer architectures.

Comparison on Visdrone dataset:

Experimental results demonstrate that the proposed method achieves competitive performance, effectively improving tracking accuracy. As shown in Table 1, for the HOTA metric, it attains 46.044%, surpassing the previous best-performing TStsm roughly by 2.1%. In terms of MOTA, it achieves 41.358%, which is slightly lower than BoT-SORT. BotSort serves as the baseline algorithm in our comparative study. As demonstrated in Table 1, our method achieves significant improvements over it across both evaluation metrics. It is a MOT framework that integrates motion cues with appearance features, incorporates camera motion compensation, and employs an augmented Kalman filter state vector. Its primary objective is to achieve reliable detection and tracking of all scene targets while preserving a unique identifier for each. However, Following targeted optimizations, our approach increased HOTA to 46.044% and IDF1 to 58.142%, yielding substantial improvements in the balance between detection and association (DetA and AssA) as well as in identity consistency—most notably through a marked reduction in ID switches and enhanced trajectory continuity. By contrast, MOTA decreased marginally by roughly 0.3%, a trade off stems from our more aggressive association strategy and dynamic confidence weighting: we deliberately accepted a slight increase in false positives and false negatives in order to secure more stable identity assignments and fewer erroneous matches. It still outperforms STCMOT (Ma et al., 2024a) and DepthMOT (Wu and Liu, 2024), demonstrating robust object management capabilities. The slight decline in MOTA primarily stems from our network that focuses on preserving object ID invariance, which yields significant improvements in HOTA and IDF1. Taking the recent AdaptTrack algorithm (Li et al., 2024a) as an example, although its MOTA score is comparable to ours, our approach—through the incorporation of spatiotemporal cues, multi scale feature enhancement, and dynamic confidence matrix adjustment—achieves further gains in both HOTA and IDF1 metrics. The TrackTrack (Shim et al., 2025) and TStsm (Zhang and Ye, 2025) demonstrate competitive performance on the HOTA metric, our method consistently outperforms it across all key evaluation metrics including HOTA, MOTA, and IDF1. Overall, these results confirm that the proposed method achieves enhanced identity consistency and overall tracking quality, making it a robust solution for complex tracking scenarios.

Furthermore, the proposed method is evaluated on the VisDrone dataset and compared with multiple advanced tracking methods through visual analysis, as shown in Fig. 6. Unlike UAVDT, VisDrone includes a more diverse set of objects, simultaneously tracking vehicles, pedestrians, and other objects, though it contains relatively fewer nighttime scenes. As shown in Fig. 6, the OCSORT method exhibits poor tracking performance across the four test scenarios, characterized by redundant detection boxes and position drift (as highlighted by the yellow arrow). This challenge is especially evident in complex environments, where long-range small-object tracking suffers from limited pixel information that hampers feature extraction and reduces accuracy. Additionally, all compared methods experience varying degrees of missed detections(as highlighted by the red arrow), particularly in distant or low signal-to-noise ratio conditions, where objects are susceptible to background interference, making continuous tracking difficult. In contrast, the proposed method demonstrates stable object tracking across all test scenarios, effectively reducing redundant detection boxes and position drift. It performs well under challenging conditions such as occlusion,

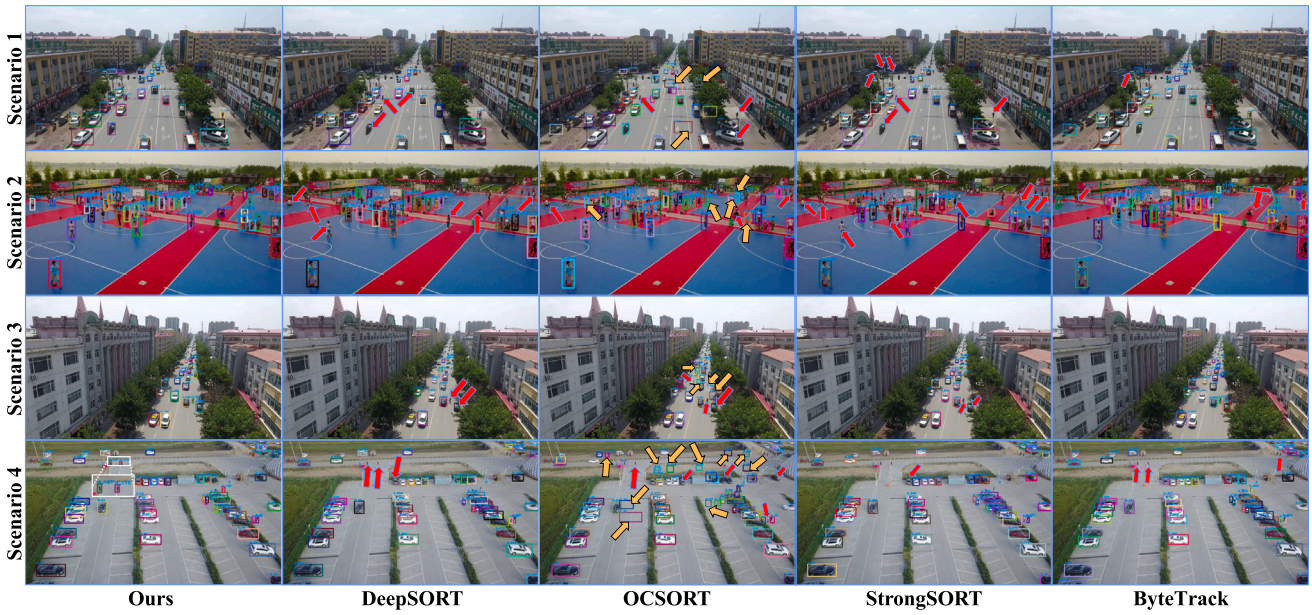

Fig. 6. The comparative experimental results of the Visdrone test dataset. Lost objects during tracking are highlighted with red arrows, while the tracking box offset is delineated using yellow arrows. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 2
Comparison with advanced MOT methods on the UAVDT-MOT dataset. The best results are highlighted in bold and the second-best results are highlighted in underline format.

Method	HOTA ↑	MOTA ↑	IDF1 ↑	FN ↓	FP ↓	IDs ↓	MT ↑	ML ↓
SORT (Bewley et al., 2016)	60.400	66.414	77.119	19 034	6505	160	201	36
DeepSORT (Veeramani et al., 2018)	61.970	68.498	<u>78.615</u>	20 035	4008	61	179	43
FairMOT (Zhang et al., 2021)	49.126	51.189	66.471	33 102	4136	110	107	87
MOTDT (Chen et al., 2018)	61.820	66.525	77.770	17 760	5825	76	175	37
ByteTrack (Zhang et al., 2022)	62.179	<u>68.754</u>	78.760	20 010	3796	102	182	42
ByteTrack+ReID (Zhang et al., 2022)	61.621	68.365	77.205	18 067	6021	118	196	44
BoT-SORT (Aharon et al., 2022)	61.369	67.741	78.508	20 296	4323	64	181	43
UAVMOT (Liu et al., 2022)	61.220	67.901	78.084	19 371	5121	69	190	42
TrackFormer (Meinhardt et al., 2022)	43.165	37.924	53.343	45 197	5585	680	67	75
OCSORT (Cao et al., 2023)	–	47.500	64.900	148 378	47 681	288	–	–
STCMOT (Ma et al., 2024a)	–	49.200	69.800	99 547	72 901	665	664	203
DroneMOT (Wang et al., 2024c)	–	50.100	69.600	112 548	57 411	129	638	178
UCMCTrack (Yi et al., 2024)	54.075	61.037	65.860	25 888	2482	984	123	51
DepthMOT (Wu and Liu, 2024)	**66.440**	62.279	78.130	28 951	3036	82	134	40
SAM2MOT (Jiang et al., 2025)	–	66.100	79.300	74 586	40 692	136	816	147
TrackTrack (Shim et al., 2025)	57.662	58.338	71.277	30 483	6617	1152	137	77
DFA-MOT (Zheng et al., 2025)	–	49.900	69.300	119 248	59 218	396	684	230
Ours	<u>64.568</u>	**70.294**	**81.667**	19 584	3137	375	190	38

and long-range tracking. For instance, in Scenario 3, a small pedestrian walking in the center of the road fails to be tracked by DeepSORT, OCSORT, and StrongSORT, whereas the proposed method successfully detect the object. Furthermore, in Scenario 4, the proposed method outperforms its counterparts by successfully tracking two pedestrians identified by ID "0_127" and "0_116", which other methods fail to track. We use white bounding boxes to highlight and enlarge these objects for clearer visualization.

Comparison on UAVDT dataset:

The tracking experiment results on the UAVDT dataset are summarized in Table 2, where the proposed method is compared against multiple advanced tracking methods, including recent approaches from the past two years, such as DepthMOT (Wu and Liu, 2024), SAM2MOT (Jiang et al., 2025), TrackTrack (Shim et al., 2025) and DFA-MOT (Zheng et al., 2025), as well as classical baselines like SORT, DeepSORT, ByteTrack, and BoT-SORT. The proposed method performs well in comparison with other advanced methods. It ranks second in the HOTA metric, while achieving first place in both MOTA and IDF1. Relative to the BotSORT baseline, our algorithm achieves substantial gains, with all three evaluation metrics showing marked improvement. Our tracker prioritizes global and local motion modeling specific to UAV footage, using multi scale fusion and adaptive weighting to deliver more consistent improvements.This superior performance can largely be attributed to the optimization of modules such as AMFE, STOE and DCMA. They focus on improving the tracking of small objects and maintaining identity consistency, which has significantly boosted the MOTA and IDF1 scores. The TrackTrack (Shim et al., 2025) and DFA-MOT (Zheng et al., 2025) achieve good results on UAVDT, making them valuable for our comparative evaluation. Although they improve performance of tracking, our proposed method still outperforms both in direct comparisons. In addition, although the SAM2MOT method employs segmentation-based cues for tracking, we integrate spatio temporal cues and multi scale feature enhancements to achieve a marked improvement in tracking quality. In comparison, while DepthMOT performs well in terms of HOTA, its MOTA score is about 8.0% lower than ours, highlighting the advantage of our method in overall tracking performance and identity preservation. For IDF1, the proposed method

Fig. 7. The comparative experimental results of the UAVDT test dataset. The lost objects during tracking are highlighted with red arrows. As indicated by the red arrows, conventional tracking methods suffer from object loss under color similarity interference between vehicles and background, whereas our algorithm maintains robust tracking performance.

reaches 81.667%, surpasses the previous best-performing DeepSORT by roughly 3.0%, further validating its robustness in identity preservation. DepthMOT integrates a per-pixel depth branch into its detection network, enabling depth-aware non-maximum suppression and overlap matching that notably enhance detection accuracy (DetA), and its camera pose compensation bolsters AssA, giving it a modest advantage in HOTA over MOSAIC Tracker. However, the added computational and error propagation overhead of depth estimation can erode its MOTA and IDF1 scores. By contrast, MOSAIC Tracker prioritizes global and local motion modeling specific to UAV footage, using multi scale fusion and adaptive weighting to maintain low FP/FN rates and minimal IDSW, thereby delivering more consistent improvements in MOTA and IDF1 while still supporting HOTA metrics. Finally, our proposed method further addresses low confidence detections to maximize target coverage and continuity in complex UAV scenarios. This indicates that our approach is particularly effective for multi-object tracking, especially in scenarios involving small objects and occlusion.

The proposed method is evaluated on the UAVDT dataset and compared against multiple advanced tracking methods through visual analysis, as shown in Fig. 7. This dataset focuses solely on vehicle tracking, excluding pedestrian and motorcycle detection and tracking. The lost objects during tracking are highlighted with red arrows. Among the four compared methods, ByteTrack demonstrates relatively superior tracking performance, maintaining stable tracking of object vehicles. However, all methods experience missed detections to some extent, particularly in low-light nighttime environments, where objects are susceptible to noise interference or blurring. For instance, in Scenario 4 processed by DeepSORT, a black car in the central region of the image is not correctly tracked, resulting in significant tracking loss. Similarly, in Scenario 3, ByteTrack struggles to maintain tracking of a object vehicle. In contrast, the proposed method consistently achieves robust tracking across all four test scenarios, effectively maintaining continuous object tracking even under challenging illumination conditions. This demonstrates that the proposed approach outperforms existing methods in object retention, occlusion recovery, and low signal-to-noise ratio environments, highlighting its strong practical potential.

4.4. Case study on two datasets

As illustrated in Fig. 8, we selected four typical scenarios for visual analysis. To enhance the readability of the results, it includes localized magnification of occluded targets and small-scale objects, with a unified layout applied across all scenarios. Experimental results were sampled approximately every 50 frames, allowing the tracking performance to be assessed through object IDs and the color-coded bounding boxes between consecutive frames. In each frame, both green and yellow bounding boxes are present, with the magnified regions of the same-colored boxes maintained across frames to facilitate reader observation.

Specifically, Scenario 1 features vehicles moving at high speeds, demonstrating the tracking performance of our algorithm in fast-motion conditions. Scenario 2 focuses on analyzing tracking performance under occlusion. It exhibits dual-motion characteristics, encompassing both mutual occlusion among moving vehicles and the motion of the UAV-based imaging platform. The continuous tracking results from Frame 1 to Frame 4 demonstrate that, despite varying degrees of occlusion—some exceeding 50%—the proposed tracking algorithm maintains consistent object IDs and stable tracking performance. Scenario 3 showcases vehicle tracking under extreme lighting conditions, validating the algorithm's robustness to illumination variations. Similarly, Scenario 4 presents tracking performance in nighttime conditions, where the algorithm is subjected to greater low-light interference compared to Scenario 1, yet overall tracking remains effective. This visual analysis not only intuitively highlights the performance advantages of the algorithm but also provides critical insights for future algorithmic improvements.

4.5. Comparison on VisDrone test-challenge dataset

To further strengthen our evaluation, we investigated additional multi-object tracking data provided by the VisDrone organizers for their annual challenge competition. Because the ground-truth annotations for this challenge competition are restricted to competitors and have not been publicly released, we performed a qualitative, visualization-based comparison on the test-challenge set (16 sequences, approximately 10 000 frames) against four representative trackers: ByteTrack, StrongSORT, OCSORT, and DeepSORT. Specifically, we download the challenge dataset and prepare the inference configuration. Then, we conducted comparative experiments with four algorithms and present the visualization results for one representative scenario from the challenge test set. In these visualizations, red arrows mark cases where

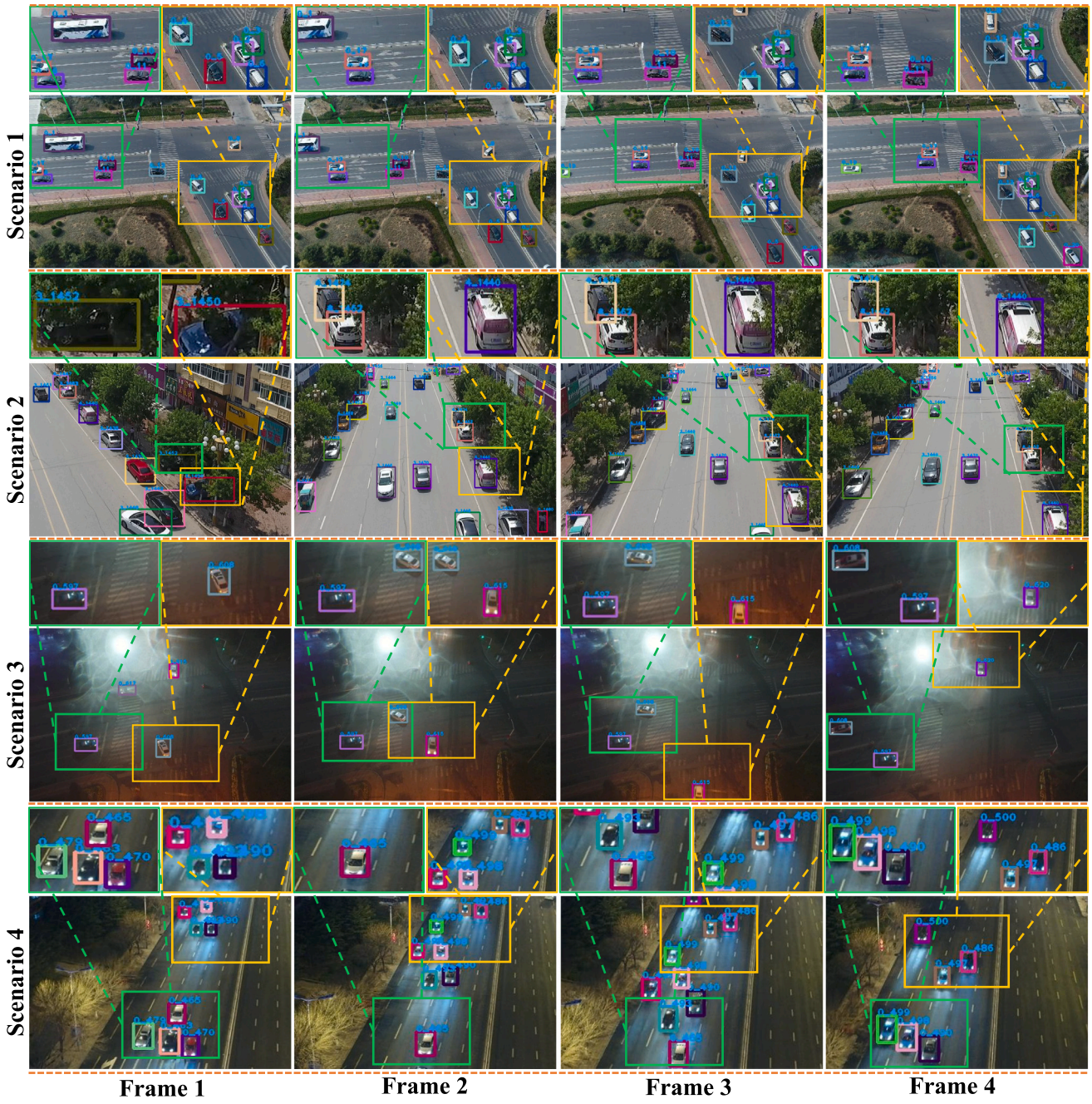

Fig. 8. Case study on two datasets. It presents representative cases from two datasets featuring occlusion, small objects, and low resolution. White bounding boxes emphasize the tracking results of selected small or occluded objects, with some objects enlarged for clearer visualization. Images were captured about every 50 frames. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

competing methods missed detections that our approach successfully captured, while yellow arrows indicate bounding-box drift or ID-switch anomalies. Frames in which all methods missed a target are not annotated. As illustrated in Fig. 9, OCSORT exhibits numerous missed detections and box misalignments in Frame 1 and Frame 2, largely because it is optimized for natural scenes and lacks specific adaptation for UAV small-object tracking. DeepSORT, though superior to OCSORT, still suffers from omissions of small targets. By contrast, our method—with its spatiotemporal occlusion enhancement and adaptive multi-scale feature modules—substantially reduces such misses. StrongSORT and ByteTrack likewise display more tracking failures relative to our approach. Overall, this qualitative comparison on the challenge dataset confirms the robustness of our tracker, in agreement with the results previously obtained on the VisDrone2019 and UAVDT benchmarks.

4.6. Testing in real-world scenarios

To demonstrate our method's applicability in real-world UAV scenarios, we conducted the following experiment. We captured a test video using a DJI Mavic 3 drone and ran our tracking pipeline on this footage. The system is implemented with a PyQt-based graphical user interface, which allows users to set all key parameters (e.g., confidence threshold, matching strategy, etc.) through an intuitive initialization panel. It performs detection and tracking on a per-frame basis within a worker thread. Each frame is first processed by the selected detector to produce bounding box proposals. Then they are passed to the tracker

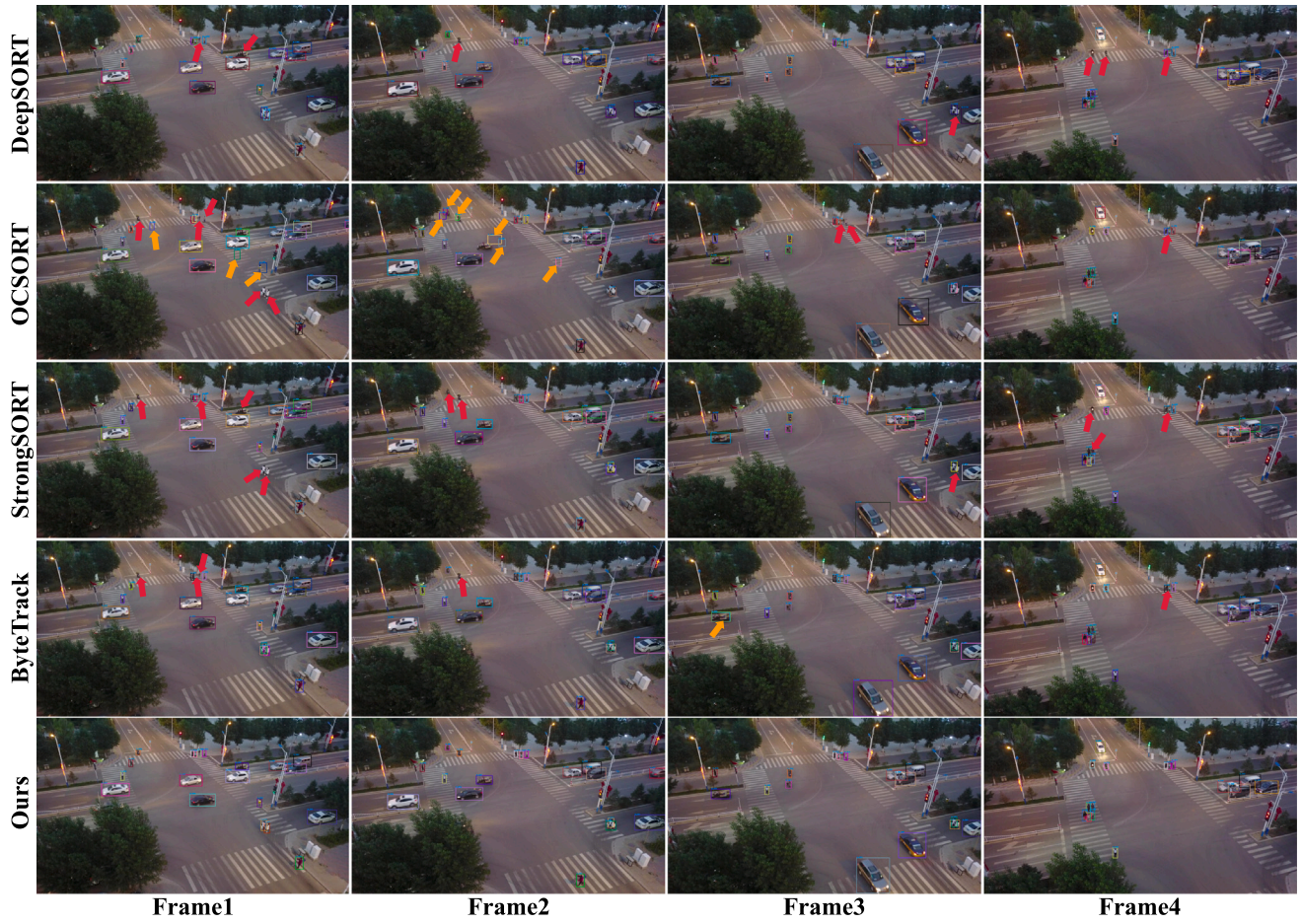

Fig. 9. Comparison on VisDrone test-challenge dataset. We conduct comparative experiments with four algorithms for one representative scenario from the challenge test set. Images were captured about every fifty frames.

Fig. 10. Testing in real-world scenarios. We captured a test video using a DJI Mavic 3 drone and ran our tracking pipeline on this footage. And we selected three representative frames to illustrate tracking performance. The system is implemented with a PyQt-based graphical user interface.

which updates the object trajectories. As shown in Fig. 10, we selected three representative frames to illustrate tracking performance. Parameter settings are displayed on the GUI, and each vehicle's trajectory is indicated by a unique ID and bounding box color. The results demonstrate tracking across varied vehicle types. This real-world evaluation confirms that our tracker can operate in real time on live UAV video streams, and that the GUI enables parameter adjustment for different conditions.

4.7. Ablation experiment

To analyze the impact of each component in our proposed method, we conducted an ablation study on the VisDrone and UAVDT datasets. The results, presented in Tables 3 and 4, compare the performance of our model with and without each module, as well as with all modules combined. We evaluated the model using HOTA, MOTA, and IDF1 metrics. The findings demonstrate the effectiveness of each component and their combined contribution to overall performance improvement.

The ablation of components with VisDrone dataset:

In the ablation study on the VisDrone dataset, we analyzed the impact of the AMFE, STOE, and DCMA modules on multi-object tracking performance. The results indicated that each module individually improves certain metrics, but may cause declines in others. For instance, using only the DCMA module yields a HOTA score of 44.762% and an IDF1 score of 56.211%, both improvements over the baseline. However, MOTA drops to 37.523%, slightly lower than with the original model. Similarly, incorporating only the STOE module increases MOTA to 38.814%, but HOTA and IDF1 exhibit minor fluctuations. This trade-off is due to the large number of small objects, such as pedestrians, in the VisDrone dataset, where a single module cannot

Table 3
Ablation study on the proposed modules for Visdrone dataset.

AMFE	STOE	DCMA	HOTA↑	MOTA↑	IDF1↑	FPS↑
–	–	–	44.722	37.553	56.203	**7.233**
✓	–	–	43.091	38.381	54.777	7.018
–	✓	–	44.154	38.814	55.789	7.135
–	–	✓	44.762	37.523	56.211	7.105
✓	✓	–	45.538	40.763	57.572	7.131
✓	✓	✓	**46.044**	**41.358**	**58.142**	7.117

Table 4
Ablation study on the proposed modules for UAVDT dataset.

AMFE	STOE	DCMA	HOTA↑	MOTA↑	IDF1↑	FPS↑
–	–	–	61.232	65.190	78.067	**14.197**
✓	–	–	63.273	68.641	80.200	14.132
–	✓	–	61.055	68.817	81.501	14.163
–	–	✓	61.378	65.045	78.246	14.118
✓	✓	–	64.329	69.947	81.503	14.090
✓	✓	✓	**64.568**	**70.294**	**81.667**	14.094

fully adapt to the varied data distribution. When both AMFE and STOE are introduced together, MOTA improves significantly to 40.763%, demonstrating that their combination effectively enhances object management. However, the best overall performance is achieved only when all three modules—AMFE, STOE, and DCMA—are integrated, resulting in optimal scores.This step-by-step accounting demonstrates that while each module delivers consistent gains in HOTA, MOTA and IDF1, the associated computational overhead remains modest.

When all three modules are combined, the overall performance improves significantly. The HOTA, MOTA, and IDF1 metrics show relative improvements of 3.0%, 10.1%, and 3.5%, respectively, over the baseline method. These results demonstrate the effectiveness of each module and their combined contribution to the overall performance enhancement of our model.

The ablation of components with UAVDT dataset:

The ablation study results on the UAVDT dataset are summarized in Table 4. We report that the baseline tracker runs at 14.197 FPS on UAVDT. Each module in isolation yields only a marginal change in throughput. Even when all three modules operate together, it still achieves 14.094 FPS, a reduction of less than 0.8% relative to baseline. In summary, the ablation study underscores the critical importance of combining these modules for optimal performance. Each module contributes its respective accuracy gains with negligible impact on real-time performance, demonstrating that our design balances tracking robustness and inference efficiency.

The speed ablation of the algorithm:

To evaluate real-time performance, we measured the inference speed (frames per second, FPS) of our method on the UAVDT test set using an NVIDIA A6000 GPU. We selected representative trackers published in 2024–2025 as baselines and ensured that all methods used identical initialization parameters (e.g., confidence threshold). The speed comparison results are summarized in Table 5. Our approach achieves an average of 14.094 FPS, closely matching UCMCTrack, while delivering superior tracking performance. Although MM-Tracker processes over 31 FPS, its multi-object tracking accuracy (HOTA, MOTA, IDF1) is substantially lower than ours. OCSORT runs slightly faster than our method but also underperforms in HOTA and IDF1. These results confirm that our algorithm not only maintains competitive inference speed but also delivers superior tracking performance compared to recent methods.

4.8. Effectiveness of AMFE

In two-stage tracking methods, the detector is an indispensable component. We further compare the performance of AMFE with these

Table 5
Algorithm speed ablation study.

Method	HOTA↑	MOTA↑	IDF1↑	FPS↑
Ours	64.568	70.294	81.667	14.094
MM-Tracker (Yao et al., 2025)	–	51.400	68.900	31.130
UCMCTrack (Yi et al., 2024)	54.075	61.037	65.860	15.430
OCSORT (Cao et al., 2023)	53.901	68.566	64.670	18.950

recent methods, including DynamicConv (Han et al., 2024), LDConv (Zhang et al., 2024c), CMUNeXtB (Tang et al., 2024), and AKconv (Zhang et al., 2023a), all of which show strong results in their respective experiments. However, in small object tracking for remote sensing, our method outperforms them in terms of HOTA, MOTA, and IDF1 scores across both datasets. This is due to the tailored optimizations specifically developed for remote sensing object tracking. AMFE enhances feature representation by combining channel-wise attention, parallel convolutions, and multi-scale feature fusion. It uses Hadamard product and concatenation to prioritize important features, while parallel convolutions capture both fine and coarse details, crucial for detecting small or occluded objects. The GMAM focuses on both global structures and multi-scale features, effectively addressing challenges like occlusion and motion blur in remote sensing.

The ablation study results presented in Table 6 highlight the superior performance of the AMFE method on both datasets. On the Visdrone dataset, AMFE outperforms other methods with a HOTA of 43.091%, MOTA of 38.381%, and IDF1 of 54.777%, demonstrating its effectiveness in challenging detection scenarios.DynamicConv follows closely with a HOTA of 42.210%, but lags behind AMFE in both MOTA and IDF1. On the UAVDT dataset, AMFE achieves the highest scores across all metrics, outperforming all competing methods. DynamicConv shows strong results, but still falls short of AMFE. Similarly, LDConv, CMUNeXtB, and AKconv also perform well but do not surpass AMFE on any of the metrics. These results highlight the effectiveness of AMFE in handling the complexities of remote sensing object tracking.

4.9. Effectiveness of STOE

The STOE module effectively addresses the challenges of remote sensing object tracking, especially for small objects that are vulnerable to occlusion and complex backgrounds. By integrating both global cumulative features and dynamic change features through multi-frame fusion, it leverages temporal information to significantly enhance tracking robustness. This approach allows the module to capture the stable, invariant characteristics of the tracked object, while also modeling its evolving features due to motion, deformation, or environmental changes. As a result, it can recover missing object details during occlusion, ensuring continuous and accurate tracking. we further compare the performance of STOE with a recent advanced method. (Ma et al., 2024a).

As shown in Table 7, the incorporation of spatiotemporal information leads to improved tracking performance. Specifically, STOE excels in maintaining the identity of tracked objects. On the Visdrone dataset, despite a slight trade-off in MOTA, STOE achieves a notable two-point improvement in IDF1, highlighting its ability to preserve object identity. The key strength of it lies in its capacity to align consecutive frames and model inter-frame correlations, which are crucial for tracking objects through occlusion. In scenarios where objects, such as a motorcycle blocked by roadside trees or a pedestrian hidden by a passing vehicle, are temporarily occluded, global features help maintain stable object characteristics, while dynamic features enable the recovery of the object once it reappears. The fusion of these features ensures that the identity of object remains intact, even during occlusion. Furthermore, on the UAVDT dataset, which presents a relatively simpler tracking scenario, the STOE module delivers even better

Table 6
Ablation study of the proposed AMFE Module on the Visdrone and UAVDT datasets. The best results are highlighted in bold.

Method	Visdrone			UAVDT		
	HOTA↑	MOTA↑	IDF1↑	HOTA↑	MOTA↑	IDF1↑
AMFE	**43.091**	**38.381**	**54.777**	**63.273**	**68.641**	**80.200**
DynamicConv (Han et al., 2024)	42.210	31.240	52.314	58.352	61.263	74.284
LDConv (Zhang et al., 2024c)	38.588	26.610	46.593	59.338	60.789	74.785
CMUNeXtB (Tang et al., 2024)	40.726	29.615	49.958	59.534	60.074	75.363
AKconv (Zhang et al., 2023a)	38.424	27.053	46.348	59.833	62.116	75.508

Table 7
Ablation study of the proposed STOE Module on the Visdrone and UAVDT datasets. The best results are highlighted in bold and the second-best results are highlighted in underline format.

Dataset	Method	HOTA↑	MOTA↑	IDF1↑
Visdrone	STOE	**43.091**	38.381	**54.777**
	TDRM (Ma et al., 2024a)	–	**38.500**	52.000
UAVDT	STOE	**63.273**	**68.641**	**80.200**
	TDRM (Ma et al., 2024a)	–	49.200	69.800

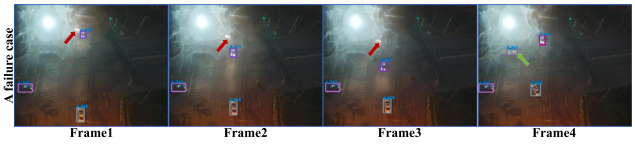

Fig. 11. An example of tracking failure. To demonstrate the limitations of our approach, we present a representative failure case where the method underperforms. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

results, further demonstrating its effectiveness. These findings emphasize the advantage of incorporating spatiotemporal features in object tracking, particularly in maintaining object identity across frames, and underscore the robustness of the STOE module in complex tracking tasks.

4.10. Effectiveness of DCMA

The DCMA method enhances multi-object tracking by incorporating detection confidence into the association process. It adjusts the similarity matrix based on the confidence of each detection, ensuring that high-confidence detections have a greater influence on matching. As it offers greater stability, ensuring that identity information remains consistent in challenging scenarios such as occlusion or abrupt appearance changes. This approach improves tracking accuracy, especially in challenging scenarios such as occlusion, by prioritizing reliable detections. By dynamically encoding detection scores into association decisions, it alleviates matching ambiguities in occlusion and deformation scenarios while maintaining computational efficiency, establishing a balanced optimization pathway for accuracy, robustness. As shown in Tables 3 and 4, the DCMA method enhances tracking performance across two datasets. This highlights the effectiveness of the method in MOT.

4.11. Display of the failure case

To demonstrate the limitations of our approach, we present a representative failure case where the method underperforms. As shown in Fig. 11, Frame 1 shows a vehicle (indicated by the red arrow) that is clearly visible to the human eye but, due to severe overexposure and low contrast, cannot be reliably detected by our model. Similarly, Frame 2 and Frame 3 exhibit tracking failures under comparable lighting interference. Only once the vehicle moves into a better illuminated region (Frame 4, green arrow) does our tracker re identify it correctly (assigned ID “0_617”). These failures stem from the scarcity of extreme glare examples during training. In future work, we plan to augment our framework with specialized modules—such as glare robust feature normalization or synthetic overexposure augmentation—to improve tracking performance under such adverse lighting conditions.

5. Conclusion

In this paper, we address the challenges of occlusion, small-object detection, viewpoint variation, and complex backgrounds in UAV-based remote sensing scenarios by introducing three core modules: Spatio-Temporal Occlusion Enhancement (STOE), Adaptive Multi-Scale Feature Enhancement (AMFE), and Dynamic Confidence Matrix Adjustment (DCMA). STOE employs global, dynamically fused feature representations to mitigate the impact of occlusions over time; AMFE reinforces object representations across multiple scales to improve the detection and tracking of small targets under varying viewpoints; and DCMA leverages a confidence-driven data-association strategy to dynamically adjust the tracking affinity matrix, thereby enhancing robustness against background clutter. Extensive experiments on the VisDrone2019 and UAVDT benchmarks demonstrate that our approach significantly improves both precision and robustness in UAV tracking tasks. In the future, we plan to integrate a semantically aware motion model to refine trajectory prediction in complex scenes and to develop a lightweight, real-time tracker by applying neural network pruning and adaptive computation allocation techniques.

CRediT authorship contribution statement

Jian Zou: Writing – original draft, Software, Methodology. **Wei Zhang:** Visualization, Methodology. **Qiang Li:** Validation. **Qi Wang:** Supervision, Project administration.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

Aharon, N., Orfaig, R., Bobrovsky, B.-Z., 2022. BoT-SORT: Robust associations multi-pedestrian tracking. arXiv preprint arXiv:2206.14651.
Bewley, A., Ge, Z., Ott, L., Ramos, F., Upcroft, B., 2016. Simple online and realtime tracking. In: 2016 IEEE International Conference on Image Processing. ICIP, pp. 3464–3468.
Cao, J., Pang, J., Weng, X., Khirodkar, R., Kitani, K., 2023. Observation-centric sort: Rethinking sort for robust multi-object tracking. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. CVPR, pp. 9686–9696.
Chen, L., Ai, H., Zhuang, Z., Shang, C., 2018. Real-time multiple people tracking with deeply learned candidate selection and person re-identification. In: 2018 IEEE International Conference on Multimedia and Expo. ICME, pp. 1–6.
Chen, Y., Tang, Y., Xiao, Y., Yuan, Q., Zhang, Y., Liu, F., He, J., Zhang, L., 2024. Satellite video single object tracking: A systematic review and an oriented object tracking benchmark. ISPRS J. Photogramm. Remote Sens. 210, 212–240.
Chen, L., Zhao, Y., Kong, S.G., 2023. Sfa-guided mosaic transformer for tracking small objects in snapshot spectral imaging. ISPRS J. Photogramm. Remote Sens. 204, 223–236.

Chu, P., Wang, J., You, Q., Ling, H., Liu, Z., 2023. Transmot: Spatial-temporal graph transformer for multiple object tracking. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. WACV, pp. 4870–4880.
Ding, J., Li, W., Yang, M., Zhao, Y., Pei, L., Tian, A., 2025. Seatrack: Rethinking observation-centric sort for robust nearshore multiple object tracking. Pattern Recognit. 159, 111091.
Du, D., Qi, Y., Yu, H., Yang, Y., Duan, K., Li, G., Zhang, W., Huang, Q., Tian, Q., 2018. The unmanned aerial vehicle benchmark: Object detection and tracking. In: Proceedings of the European Conference on Computer Vision. ECCV, pp. 370–386.
Du, Y., Zhao, Z., Song, Y., Zhao, Y., Su, F., Gong, T., Meng, H., 2023. Strongsort: Make deepsort great again. IEEE Trans. Multimed. 25, 8725–8737.
Du, D., Zhu, P., Wen, L., Bian, X., Lin, H., Hu, Q., Peng, T., Zheng, J., Wang, X., Zhang, Y., et al., 2019. VisDrone-DET2019: The vision meets drone object detection in image challenge results. In: Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops. ICCVW, pp. 213–226.
Gao, R., Qi, J., Wang, L., 2025. Multiple object tracking as id prediction. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 27883–27893.
Guan, Z., Wang, Z., Zhang, G., Li, L., Zhang, M., Shi, Z., Jiang, N., 2025. Multi-object tracking review: retrospective and emerging trend. Artif. Intell. Rev. 58 (8), 1–46.
Han, K., Wang, Y., Guo, J., Wu, E., 2024. ParameterNet: Parameters are all you need for large-scale visual pretraining of mobile networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. CVPR, pp. 15751–15761.
He, B., Yuan, L., Lv, K., 2024. FSTrack: One-shot multi-object tracking algorithm based on feature enhancement and similarity. IEEE Signal Process. Lett. 31, 775–779.
Hu, X., Sun, F., Sun, J., Wang, F., Li, H., 2024a. Cross-modal fusion and progressive decoding network for RGB-D salient object detection. Int. J. Comput. Vis., (IJCV) 1–19.
Hu, W., Wang, S., Zhou, Z., Gao, J., Li, Y., Maybank, S., 2024b. One-stage anchor-free online multiple target tracking with deformable local attention and task-aware prediction. IEEE Trans. Pattern Anal. Mach. Intell. 46, 11446–11463.
Hu, M., Zhu, X., Wang, H., Cao, S., Liu, C., Song, Q., 2023. Stdformer: Spatial-temporal motion transformer for multiple object tracking. IEEE Trans. Circuits Syst. Video Technol. 33 (11), 6571–6594.
Huang, B., Li, J., Chen, J., Wang, G., Zhao, J., Xu, T., 2023. Anti-uav410: A thermal infrared benchmark and customized scheme for tracking drones in the wild. IEEE Trans. Pattern Anal. Mach. Intell. 46, 2852–2865.
Jiang, J., Wang, Z., Zhao, M., Li, Y., Jiang, D., 2025. SAM2MOT: A novel paradigm of multi-object tracking by segmentation. arXiv preprint arXiv:2504.04519.
Lei, X., Cheng, W., Xu, C., Yang, W., 2024. Joint target and background temporal propagation for aerial tracking. ISPRS J. Photogramm. Remote Sens. 211, 121–134.
Li, K., Wang, L., Ren, H., Cao, Y., 2024a. ?AdaptTrack: Multi-object tracking by adaptive correlation. In: 2024 IEEE International Symposium on Parallel and Distributed Processing with Applications. ISPA, IEEE, pp. 1697–1703.
Li, M., Zhang, Y., Jia, Y., Yang, Y., 2025a. Advancing multi-object tracking through occlusion-awareness and trajectory optimization. Knowl.-Based Syst. 310, 112930.
Li, R., Zhang, B., Liu, W., Li, Z., Fan, J., Teng, Z., Fan, J., 2024b. HCgNet: A hierarchical context-guided network for multi-object tracking. Knowl.-Based Syst. 297, 111859.
Li, Q., Zhang, W., Lu, W., Wang, Q., 2025b. Multi-branch mutual-guiding learning for infrared small target detection. IEEE Trans. Geosci. Remote Sens. 63, 1–10.
Li, Q., Zhang, M., Yang, Z., Yuan, Y., Wang, Q., 2024c. Edge-guided perceptual network for infrared small target detection. IEEE Trans. Geosci. Remote Sens..
Li, Y., Zhou, S., Qin, Z., Wang, L., Wang, J., Zheng, N., 2024d. Single-shot and multi-shot feature learning for multi-object tracking. IEEE Trans. Multimed. 26, 9515–9526.
Lin, B., Zheng, J., Xue, C., Fu, L., Li, Y., Shen, Q., 2024. Motion-aware correlation filter-based object tracking in satellite videos. IEEE Trans. Geosci. Remote Sens. 62, 1–13.
Liu, Z., Huang, H., Dong, H., Xing, F., 2025. Iou-guided siamese network with high-confidence template fusion for visual tracking. Neurocomputing 614, 128774.
Liu, S., Li, X., Lu, H., He, Y., 2022. Multi-object tracking meets moving UAV. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. CVPR, pp. 8876–8885.
Liu, Y., Li, B., Zhou, X., Li, D., Duan, Q., 2024a. FishTrack: Multi-object tracking method for fish using spatiotemporal information fusion. Expert Syst. Appl. 238, 122194.
Liu, L., Song, X., Song, H., Sun, S., Han, X.-F., Akhtar, N., Mian, A., 2024b. Yolo-3DMM for simultaneous multiple object detection and tracking in traffic scenarios. IEEE Trans. Intell. Transp. Syst. 25 (8), 9467–9481.
Lv, W., Zhang, N., Zhang, J., Zeng, D., 2023. One-shot multiple object tracking with robust id preservation. IEEE Trans. Circuits Syst. Video Technol. 34 (6), 4473–4488.
Ma, F., Shou, M.Z., Zhu, L., Fan, H., Xu, Y., Yang, Y., Yan, Z., 2022. Unified transformer tracker for object tracking. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. CVPR, pp. 8781–8790.
Ma, J., Tang, C., Wu, F., Zhao, C., Zhang, J., Xu, Z., 2024a. STCMOT: Spatio-temporal cohesion learning for UAV-based multiple object tracking. In: 2024 IEEE International Conference on Multimedia and Expo. ICME, pp. 1–6.
Ma, S., Zhao, B., Hou, Z., Yu, W., Pu, L., Yang, X., 2024b. SOCF: A correlation filter for real-time UAV tracking based on spatial disturbance suppression and object saliency-aware. Expert Syst. Appl. 238, 122131.
Meinhardt, T., Kirillov, A., Leal-Taixe, L., Feichtenhofer, C., 2022. Trackformer: Multi-object tracking with transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. CVPR, pp. 8844–8854.
Qiao, X., Zhao, Y., Chen, L., Kong, S.G., Chan, J.C.-W., 2022. Mosaic gradient histogram for object tracking in DoFP infrared polarization imaging. ISPRS J. Photogramm. Remote Sens. 194, 108–118.
Rahman, M.M., Munir, M., Marculescu, R., 2024. EMCAD: Efficient multi-scale convolutional attention decoding for medical image segmentation. In: 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. CVPR, pp. 11769–11779.
Rao, Z., Li, X., Xiong, B., Dai, Y., Shen, Z., Li, H., Lou, Y., 2024. Cascaded recurrent networks with masked representation learning for stereo matching of high-resolution satellite images. ISPRS J. Photogramm. Remote Sens. 218, 151–165.
Ren, L., Yin, W., Diao, W., Fu, K., Sun, X., 2024. Motion-guided multi-object tracking model for high-speed aerial objects in satellite videos. IEEE Trans. Geosci. Remote Sens..
Shim, K., Ko, K., Yang, Y., Kim, C., 2025. Focusing on tracks for online multi-object tracking. In: Proceedings of the Computer Vision and Pattern Recognition Conference. CVPR, pp. 11687–11696.
Tang, F., Ding, J., Quan, Q., Wang, L., Ning, C., Zhou, S.K., 2024. Cmunext: An efficient medical image segmentation network based on large kernel and skip fusion. In: 2024 IEEE International Symposium on Biomedical Imaging. ISBI, IEEE, pp. 1–5.
Veeramani, B., Raymond, J.W., Chanda, P., 2018. DeepSort: deep convolutional networks for sorting haploid maize seeds. BMC Bioinformatics 19, 1–9.
Wang, Y., Kitani, K., Weng, X., 2021. Joint object detection and multi-object tracking with graph neural networks. In: 2021 IEEE International Conference on Robotics and Automation. ICRA, IEEE, pp. 13708–13715.
Wang, Z., Li, M., Li, Z., Zhang, X., Li, M., Li, Z., Ding, W., Wu, X., 2024a. Mm-tracker: Visual tracking with a multi-task model integrating detection and differentiating feature extraction. IEEE Trans. Emerg. Top. Comput. Intell..
Wang, Y., Li, R., Zhang, D., Li, M., Cao, J., Zheng, Z., 2025. CATrack: Condition-aware multi-object tracking with temporally enhanced appearance features. Knowl.-Based Syst. 308, 112760.
Wang, B., Sui, H., Ma, G., Zhou, Y., 2024b. MCTracker: Satellite video multi-object tracking considering inter-frame motion correlation and multi-scale cascaded feature enhancement. ISPRS J. Photogramm. Remote Sens. 214, 82–103.
Wang, P., Wang, Y., Li, D., 2024c. Dronemot: Drone-based multi-object tracking considering detection difficulties and simultaneous moving of drones and objects. In: 2024 IEEE International Conference on Robotics and Automation. ICRA, IEEE, pp. 7397–7404.
Wen, J., Chu, H., Lai, Z., Xu, T., Shen, L., 2023. Enhanced robust spatial feature selection and correlation filter learning for UAV tracking. Neural Netw. 161, 39–54.
Wu, X., Li, W., Hong, D., Tao, R., Du, Q., 2021. Deep learning for unmanned aerial vehicle-based object detection and tracking: A survey. IEEE Geosci. Remote. Sens. Mag. 10 (1), 91–124.
Wu, J., Liu, Y., 2024. DepthMOT: Depth cues lead to a strong multi-object tracker. arXiv preprint arXiv:2404.05518.
Wu, Y., Liu, Q., Sun, H., Xue, D., 2024. HRTracker: multi-object tracking in satellite video enhanced by high-resolution feature fusion and an adaptive data association. Remote. Sens. 16 (17), 3347.
Xu, Y., Ban, Y., Delorme, G., Gan, C., Rus, D., Alameda-Pineda, X., 2022. TransCenter: Transformers with dense representations for multiple-object tracking. IEEE Trans. Pattern Anal. Mach. Intell. 45 (6), 7820–7835.
Xu, L., Huang, Y., 2024. Rethinking joint detection and embedding for multiobject tracking in multiscenario. IEEE Trans. Ind. Inform..
Xu, Q., Wang, L., Sheng, W., Wang, Y., Xiao, C., Ma, C., An, W., 2024a. Heterogeneous graph transformer for multiple tiny object tracking in RGB-T videos. IEEE Trans. Multimed. 26, 9383–9397.
Xu, X., Wang, X., Wu, F., Zhang, Z., Chang, L., 2024b. ERMOT: Evidence reasoning-based robust multiple object tracking method. IEEE Trans. Ind. Inform. 1–10.
Xu, Q., Xu, Z., Wang, H., Chen, Y., Tao, L., 2025. Online learning discriminative sparse convolution networks for robust UAV object tracking. Knowl.-Based Syst. 308, 112742.
Yao, M., Peng, J., He, Q., Peng, B., Chen, H., Chi, M., Liu, C., Benediktsson, J.A., 2025. MM-tracker: Motion mamba for UAV-platform multiple object tracking. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 39, pp. 9409–9417.
Yasir, M., Liu, S., Pirasteh, S., Xu, M., Sheng, H., Wan, J., de Figueiredo, F.A., Aguilar, F.J., Li, J., 2024. YOLOShipTracker: Tracking ships in SAR images using lightweight YOLOv8. Int. J. Appl. Earth Obs. Geoinf. 134, 104137.
Yi, K., Luo, K., Luo, X., Huang, J., Wu, H., Hu, R., Hao, W., 2024. UCMCTrack: Multi-object tracking with uniform camera motion compensation. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 38, pp. 6702–6710. http://dx.doi.org/10.1609/aaai.v38i7.28493.
Yuan, Y., Wu, Y., Zhao, L., Liu, Y., Pang, Y., 2025. TLSH-MOT: Drone-view video multiple object tracking via transformer-based locally sensitive hash. IEEE Trans. Geosci. Remote Sens..
Zeng, F., Dong, B., Zhang, Y., Wang, T., Zhang, X., Wei, Y., 2022. Motr: End-to-end multiple-object tracking with transformer. In: European Conference on Computer Vision. ECCV, pp. 659–675.
Zhang, W., Li, Q., Yuan, Y., Wang, Q., 2024a. Visual consistency enhancement for multi-view stereo reconstruction in remote sensing. IEEE Trans. Geosci. Remote Sens. 62, 1–11.

Zhang, Y., Liang, Y., Leng, J., Wang, Z., 2024b. SCGTracker: Spatio-temporal correlation and graph neural networks for multiple object tracking. Pattern Recognit. 149, 110249.

Zhang, X., Song, Y., Song, T., Yang, D., Ye, Y., Zhou, J., Zhang, L., 2023a. AKConv: Convolutional kernel with arbitrary sampled shapes and arbitrary number of parameters. arXiv preprint arXiv:2311.11587.

Zhang, X., Song, Y., Song, T., Yang, D., Ye, Y., Zhou, J., Zhang, L., 2024c. LDConv: Linear deformable convolution for improving convolutional neural networks. Image Vis. Comput. 149, 105190.

Zhang, Y., Sun, P., Jiang, Y., Yu, D., Weng, F., Yuan, Z., Luo, P., Liu, W., Wang, X., 2022. ByteTrack: Multi-object tracking by associating every detection box. In: European Conference on Computer Vision. pp. 1–21.

Zhang, J., Wang, M., Jiang, H., Zhang, X., Yan, C., Zeng, D., 2023b. STAT: Multi-object tracking based on spatio-temporal topological constraints. IEEE Trans. Multimed. 26, 4445–4457.

Zhang, Y., Wang, C., Wang, X., Zeng, W., Liu, W., 2021. Fairmot: On the fairness of detection and re-identification in multiple object tracking. Int. J. Comput. Vis. 129, 3069–3087.

Zhang, P., Xu, J., Wu, Q., Huang, Y., Ben, X., 2020. Learning spatial-temporal representations over walking tracklet for long-term person re-identification in the wild. IEEE Trans. Multimed. 23, 3562–3576.

Zhang, Z., Ye, C., 2025. Real-time multi-object tracking algorithm for UAV based on deep learning. In: 2025 4th Asia Conference on Algorithms, Computing and Machine Learning. CACML, pp. 1–9.

Zheng, Y., He, C., Chen, X., Zhang, H., Qu, T., Wang, D., 2025. DFA-MOT: A dynamic field-aware multi-object tracking framework for unmanned aerial vehicles. IEEE Trans. Circuits Syst. Video Technol. 1–1.

Zheng, Y., Yu, Z., Wang, S., Huang, T., 2022. Spike-based motion estimation for object tracking through bio-inspired unsupervised learning. IEEE Trans. Image Process. 32, 335–349.

Zhu, F., Cui, J., Dou, K., 2023. Spatio-temporal hierarchical feature transformer for UAV object tracking. ISPRS J. Photogramm. Remote Sens. 204, 442–452.

Zhu, Q., Huang, X., Guan, Q., 2024. TabCtNet: Target-aware bilateral CNN-transformer network for single object tracking in satellite videos. Int. J. Appl. Earth Obs. Geoinf. 128, 103723.

Zhuang, K., Li, Q., Yuan, Y., Wang, Q., 2023. Multi-domain adaptation for motion deblurring. IEEE Trans. Multimed. 26, 3676–3688.