remote sensing

MDPI

Article

TASONet: A Spatial Enhancement and Temporal Modeling Framework for UAV Small Object Tracking

Ruiqi Ma [1], Changcai Lai [1], Qinghua Sheng [1,2], Zehao Tao [1] and Xiaorun Li [3,*]

[1] The School of Electronics and Information, Hangzhou Dianzi University, Hangzhou 310018, China
[2] Zhejiang Key Laboratory of Intelligent Vehicle Electronics Research, Hangzhou 310018, China
[3] The College of Electrical Engineering, Zhejiang University, Hangzhou 310027, China
* Correspondence: lxr@zju.edu.cn

Highlights

What are the main findings?

- TASONet introduces a lightweight spatial–temporal enhancement framework that significantly improves the representation of ground small objects in UAV-view imagery while effectively suppressing background interference.
- The integrated SOE and TEM modules collaboratively enhance spatial discriminability and temporal consistency, delivering substantial performance gains on UAVDT and VisDrone-MOT with minimal computational overhead.

What are the implications of the main findings?

- The findings demonstrate that lightweight spatial enhancement coupled with temporal modeling can overcome the common trade-off between small object amplification and false positive suppression in small object MOT.
- The efficiency and accuracy of TASONet make it well suited for real-time UAV-based applications such as aerial surveillance, traffic monitoring, and large-scale scene analysis.

Abstract

Multi object tracking (MOT) in UAV imagery is challenged by weak feature representation of small objects due to limited resolution, which leads to frequent missed detections. However, enhancing small object features often amplifies background noise and increases false positives. To address this contradiction, we propose the Temporal Aware Small Object Enhancement Network (TASONet), which integrates spatial enhancement and temporal modeling for robust tracking. The Small Object Enhancement (SOE) module combines depthwise separable convolutions with contrast-aware attention mechanisms (SimAM and LCDAttn) to improve local discriminability. It further incorporates the Small Target Enhancement Path (STEP), which uses motion-difference cues and a confidence adaptive suppression strategy to strengthen spatial features while mitigating noise. The Temporal Enhancement Module (TEM), consisting of Temporal Feature Alignment (TFA) and a Target Memory Unit (TMU), aggregates multi-frame information through adaptive inter-frame fusion and memory of high confidence historical features, improving temporal consistency and reducing false positives potentially introduced by SOE. Experiments show that TASONet achieves significant gains over state-of-the-art methods: on UAVDT, MOTA increases from 68.33 to 75.97 and IDF1 from 83.50 to 88.51; on VisDrone-MOT, MOTA rises from 61.15 to 73.52 with an IDF1 of 88.83. These results validate the effectiveness of jointly enhancing spatial features and temporal coherence for UAV small-object MOT.

Check for updates

Academic Editor: Wen Yang

Received: 16 January 2026
Revised: 5 February 2026
Accepted: 9 February 2026
Published: 11 February 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

Keywords: multi object tracking; small object detection; ByteTrack; attention mechanism; temporal modeling; Detector Tracker Integration

1. Introduction

Multi object tracking (MOT) is a fundamental task in the field of computer vision, with broad applications in intelligent transportation, UAV surveillance, urban security, and autonomous driving. In recent years, deep learning based detection algorithms such as the YOLO series [1–5] have achieved remarkable progress in medium and large scale object scenarios. However, detecting small objects in complex environments remains highly challenging. According to the COCO [6] and VisDrone [7] benchmarks, small objects are typically defined as those with a resolution lower than 32×32 pixels. Such objects occupy only a small portion of the image and contain weak feature representations. They are also sensitive to illumination variation, motion blur, and background clutter. These factors hinder the preservation of semantic integrity in high level feature maps, often causing missed detections, false positives, and temporal instability across frames.

The limited performance of existing detection frameworks in small object scenarios can be attributed to three main causes. First, in the spatial dimension, small objects exhibit weak saliency, and their valid features are often overwhelmed by surrounding background information, lacking sufficient discriminability. Second, in the temporal dimension, mainstream detectors usually infer from single frame images without modeling inter-frame contextual information, resulting in unstable detection results and frequent missing detections across frames, which subsequently lead to trajectory interruptions in MOT. Third, in the confidence dimension, many low confidence features are mistakenly amplified in complex backgrounds due to the absence of effective false positive suppression mechanisms, thereby degrading the overall stability and reliability of detection. Consequently, achieving joint optimization of spatial feature enhancement, temporal consistency modeling, and adaptive false positive suppression under real-time constraints remains a key challenge in small object detection research.

Although existing studies have made notable progress in attention mechanisms and video based detection, several critical limitations still persist. Mainstream attention modules such as SE-Net [8], CBAM [9], and ECA [10] predominantly focus on channel weighting or global feature modeling but fail to explicitly capture the local contrast between objects and their surrounding backgrounds, thereby limiting their ability to emphasize the salient characteristics of small objects. Meanwhile, multi-frame detection methods frequently depend on optical flow estimation [11,12] or recurrent network structures [13] for temporal feature alignment, which not only incur substantial computational and training overhead but also exhibit degraded performance under challenging conditions such as low frame rates or rapid object motion. Furthermore, current detectors often neglect the adverse effect of noise propagation induced by low confidence samples during feature enhancement, consequently leading to an increased false positive rate. Therefore, it remains a challenge to develop a unified framework capable of simultaneously overcoming the issues of insufficient spatial saliency, inadequate temporal exploitation, and ineffective suppression of false positive in small object detection.

To address these challenges, we propose a Temporal Aware Small Object Enhancement Network (TASONet). Built upon an improved YOLOPAFPN architecture with a CSPDarknet backbone, the network integrates two complementary core modules, the Small Object Enhancement (SOE) and the Temporal Enhancement Module (TEM), to strengthen small object representation from both spatial and temporal perspectives while maintaining

real-time inference efficiency. First, to tackle the low contrast and weak discriminability of small objects in cluttered backgrounds, a Contrastive Attention Mechanism (CAM) is proposed within the SOE module. It synergistically combines SimAM channel attention for semantic refinement and Local Contrast Driven Attention (LCDAttn) for spatial contrast enhancement, explicitly modeling luminance and texture differences between objects and their neighborhoods via local mean and variance computation to improve feature robustness under low signal to noise conditions. Second, to mitigate false positive and noise propagation from low confidence samples, the SOE module further incorporates a Small Target Enhancement Path (STEP), which integrates motion difference features, depthwise separable convolutions, and a confidence-adaptive suppression strategy to stabilize fine grained representations without compromising real-time performance. Finally, to address the insufficient utilization of inter-frame information and unstable temporal consistency, the TEM module is designed with a Temporal Feature Align (TFA) unit and a Target Memory Unit (TMU). TFA employs gated convolutions for adaptive fusion of adjacent frame features, while TMU maintains a sliding memory pool to preserve high confidence historical representations, enabling stable cross frame detection without relying on optical flow estimation.

The main contributions of this paper are summarized as follows:

- A CAM is proposed, which jointly leverages SimAM and LCDAttn to model channel semantics and local spatial contrast, significantly enhancing the discriminability between small objects and background clutter.
- A STEP is constructed, integrating multi dimensional motion features and confidence adaptive suppression to achieve comprehensive small object enhancement and noise resilient detection.
- A lightweight TEM with TFA and TMU is designed, realizing adaptive short term feature alignment and long term memory modeling to improve temporal consistency and tracking performance under dynamic UAV scenarios.

In summary, this work establishes a unified enhancement framework that jointly optimizes spatial saliency, temporal consistency, and false positive suppression across complementary dimensions. The proposed TASONet significantly improves detection and tracking continuity in complex dynamic scenes, providing an efficient and scalable solution for real-time UAV and mobile platform applications in video object detection.

2. Related Work

2.1. Detection Based MOT

The Tracking by Detection (TBD) paradigm has become the dominant framework in video surveillance, autonomous driving, and aerial monitoring, owing to its clear two-stage design that follows the principle of "detection first, association later." It first employs a detector to generate candidate bounding boxes and subsequently performs inter-frame identity association through motion modeling and feature matching. Representative algorithms such as SORT [14], DeepSORT [15], FairMOT [16], and ByteTrack [17] constitute the core lineage of MOT, continuously advancing performance toward higher accuracy, real-time efficiency, and robust stability.

The SORT [14] algorithm proposed by Bewley et al. in 2016 established the foundational TBD pipeline with its minimalist combination of a Kalman filter and the Hungarian algorithm. The Kalman filter predicts target motion states to ensure spatiotemporal continuity, while the Hungarian algorithm performs optimal matching based on the Intersection over Union (IoU) metric. Despite its outstanding computational efficiency on CPUs, SORT relies solely on motion cues and thus becomes unreliable under occlusion or in scenes with similar motion patterns, often leading to identity switches (IDS). To mitigate these issues,

DeepSORT [15] introduced discriminative appearance modeling into the motion-based framework. By extracting high-dimensional features through a pre-trained Re-Identification (ReID) network, it constructs a joint cost matrix combining motion and appearance similarities. The cascade matching mechanism and improved track lifecycle management further enhance recovery after occlusion and reduce trajectory fragmentation. However, the added appearance modeling increases computational overhead and maintains a decoupled "detect then associate" design, limiting end-to-end optimization.

FairMOT [16] subsequently proposed an end-to-end joint learning paradigm integrating detection and ReID within a unified network. Based on CenterNet with a shared backbone (e.g., DLA-34), it simultaneously predicts target centers and produces discriminative embeddings optimized under a multi-task loss. This design synchronizes detection and appearance feature generation, achieving a better trade-off between accuracy and speed in crowded scenes. Nonetheless, training instability may arise from the imbalance between detection and ReID losses, often requiring pre-training or careful loss-weight tuning.

In 2022, ByteTrack [17] achieved a new balance between accuracy and efficiency by revising the detection filtering strategy. Its key insight is the reuse of low-confidence detections, often discarded by previous methods, as valid candidates for occluded or partially visible targets. By sequentially associating high- and low-confidence detections, it effectively reduces track fragmentation while maintaining real-time performance. ByteTrack further adopts a lightweight autoregressive motion model instead of Kalman filtering, demonstrating superior MOTA and speed on benchmarks such as MOT17 and BEnSMOT. However, it still relies on single-frame detection and lacks temporal modeling, leading to localization drift in fast-motion scenarios and limited robustness for small objects.

Building upon ByteTrack, several subsequent works further enhanced motion modeling and association robustness. BoT-SORT [18] integrates appearance embeddings and camera aware motion estimation into ByteTrack's association framework, improving identity consistency under occlusion. OC-SORT [19] reformulates the Kalman-based prediction step by exploiting observation-centric motion modeling, enabling better trajectory continuity when detections are missing. DeepOC-SORT [20] extends this by combining deep appearance features with OC-SORT's adaptive motion update, achieving more stable re-identification in crowded or long-term tracking scenarios. BoTTrack [21] further refines this line by introducing a boosting-based trajectory correction mechanism that adaptively adjusts association thresholds based on detection confidence and motion uncertainty. Recently, the Improved Association Tracker advances association strategies through adaptive cost weighting and motion-aware gating, offering consistent improvements in accuracy, recall, and tracking continuity across multiple benchmarks.

Recent advances in MOT research from 2023 to 2025 largely continue along this association-centric and motion-driven trajectory. For instance, SAM2MOT [22] introduces segmentation aware tracking to strengthen inter-frame association robustness, achieving state-of-the-art performance on UAVDT and DanceTrack. View-Centric MOT [23] addresses viewpoint changes in moving UAV scenarios via homography based matching, improving association reliability under camera motion. Other approaches, such as DMTrack [24], propose deformable state space motion modeling with uncertainty aware fusion to enhance trajectory estimation under complex UAV motion dynamics. In addition, recent appearance guided motion modeling methods jointly exploit appearance and motion cues to stabilize identity continuity in aerial videos [25].

Despite different modeling emphases, these methods predominantly extend or refine the core association heuristics and motion modeling principles established by ByteTrack, BoT-SORT, OC-SORT, and DeepOC-SORT. Accordingly, these trackers remain widely

adopted and recognized as strong baselines in contemporary MOT research, and are therefore selected as the primary comparison methods in this work.

At present, the evolution of TBD frameworks is converging toward a hybrid paradigm that integrates lightweight temporal modeling, adaptive association strategies, and shared appearance learning. These advancements collectively enhance tracking accuracy and robustness while maintaining real-time inference efficiency, further extending the applicability of MOT in complex environments with dense small objects or dynamic interference.

2.2. Small Object Detection

Small object detection remains a fundamental challenge in object detection and tracking due to the extremely limited feature information of small objects. With minimal pixel coverage, small objects lack sufficient texture and contour cues, making them vulnerable to background noise, occlusion, and scale variation. Consequently, models often exhibit high miss rates and poor localization accuracy. Although the YOLO series has progressively enhanced performance through multi-scale feature fusion, e.g., YOLOv3 [2] introducing Feature Pyramid Networks (FPN) for semantic–spatial fusion and YOLOv5 incorporating Spatial Pyramid Pooling (SPP) for stronger local aggregation, significant deficiencies persist in extreme conditions. In long-range surveillance or partially occluded traffic scenes, YOLOv5 still shows high miss rates for distant pedestrians or small vehicles. The core issue lies in the simple stacking of multi-scale features, which fails to reconcile the semantic spatial trade off between high and low level representations. Moreover, fixed anchor-box designs often lead to mismatched positive sample assignments, further limiting robustness and recall for small objects.YOLOX [5] addresses these limitations through an anchor-free paradigm that adaptively generates predictions within target-centered regions, mitigating anchor scale mismatch. It introduces the Simple Online and Target-aware Assignment (SimOTA) strategy, dynamically selecting positive samples based on IoU and feature similarity, thereby improving the representational balance for small objects. At the feature fusion level, YOLOX refines the Path Aggregation Network (PANet) to enhance interaction between semantic and spatial features, effectively preserving fine grained details. Experiments on the COCO small object subset show that YOLOX-L outperforms YOLOv5-L of comparable scale, achieving higher precision and recall, especially in small vehicle monitoring tasks.

Beyond YOLOX, various targeted strategies further enhance small object detection. The Bidirectional Feature Pyramid Network (BiFPN) [26] improves fusion efficiency via weighted cross scale aggregation, achieving balanced multi-scale representation. At the detection head level, YOLOv7 [4] integrates an Extended Efficient Layer Aggregation Network (ELAN) with an additional small object branch, refined anchor sizes, and richer prediction channels, significantly boosting recall. Similarly, the anchor-free FCOS [27] introduces a centrality branch to refine localization confidence, improving small object discrimination.Lightweight convolutional architectures also play a key role in efficient deployment. Depthwise separable convolution preserves local detail while drastically reducing parameters, enabling real-time small object detection on resource limited devices. MobileNetV3-based [28] detectors achieve competitive COCO performance with high CPU efficiency, while RepVGG [29] employs structural re-parameterization to merge multi-branch training networks into single-path inference structures, achieving high accuracy and low latency in UAV inspection tasks.

Overall, research in small object detection is evolving toward the coordinated integration of feature enhancement, architectural optimization, and efficiency balance. Emerging multimodal approaches, such as fusing infrared and visible light data, further expand the potential for reliable detection in complex real-world environments.

2.3. Application of Attention Mechanism in Object Detection

Attention mechanisms enhance salient features while suppressing redundancy and are crucial for overcoming bottlenecks in object detection, particularly for small objects. They have evolved from single dimensional to multi-dimensional designs and from heavy to lightweight models. Channel attention, first realized by SE-Net [8], uses global pooling and fully connected layers to model inter-channel dependencies, improving sensitivity to key features. However, for small objects, dimensionality reduction and pooling can obscure fine spatial details, limiting performance SE-Net-ResNet50 achieves only 26.8% mAP on COCO small objects. Joint spatial channel attention has been proposed to address this. CBAM [9] sequentially refines "what" and "where" to focus on, combining max and average pooled channel features and generating a spatial attention map. Integrated into Faster R-CNN with ResNet50, CBAM improves AP by 2.3% on VOC 2007 while remaining lightweight. ECA-Net [10] avoids dimensionality reduction and models local channel dependencies via adaptive 1D convolution, balancing accuracy and efficiency, making it suitable for small object detection under resource constraints. Recently, parameter free and local contrast attention methods have gained traction. SimAM [30] computes per-pixel importance without extra parameters, while LCDAttention enhances fine grained and boundary features through local contrast modeling. These approaches improve edge representation and localization for small objects, offering promising directions for efficient and robust detection.

2.4. Temporal Modeling in Video Object Detection and Multi Object Tracking Integration

The core challenge in video object detection versus static images lies in handling appearance degradation caused by motion, occlusion, and illumination changes. Temporal information is crucial to mitigate these issues, as associating features across frames can supplement missing semantic and spatial cues, enhancing detection stability. Mainstream temporal modeling approaches include cross-frame feature alignment, motion-differential feature fusion, and memory based mechanisms. Early methods rely on cross frame alignment and global context modeling; for example, MEGA [31] introduces a spatiotemporal pyramid framework with cross frame attention to capture long term dependencies. Memory mechanisms further aid occlusion handling. For instance, the Spatio-Temporal Memory Network (STMN) series [32] construct a global spatio-temporal memory bank to store historical features, but its undifferentiated memory update and global feature matching lack targeted optimization for small object scenarios. However, these methods are often poorly integrated with MOT, as detection and tracking remain largely independent stages. Temporal features are rarely used for association, leading to errors with similar or intersecting targets. On MOT17, combining MEGA for detection with DeepSORT yields 42.3 ID Switches, whereas integrating MEGA's global features into tracking reduces ID Switches to 28.6, highlighting the potential of deeply integrating temporal modeling with MOT.

3. Methods

3.1. Overall Framework

As shown in Figure 1, TASONet is a multi target tracking framework for small objects in UAV imagery, built upon an improved YOLOPAFPN architecture. It adopts CSPDarknet as the backbone and integrates two core modules, the SOE module and TEM to collaboratively optimize spatial feature enhancement for small objects and temporal consistency for tracking.

Given an input video frame, the CSPDarknet backbone extracts hierarchical features (dark3, dark4, dark5) that integrate low-level structural cues and high-level semantic information. These features are fed into the Feature Pyramid Network (FPN) for top-

down multi-scale fusion to produce an initial fused feature map. This map still exhibits insufficient local contrast for small objects, noticeable background interference, and lack of temporal correlation, thus requiring spatial purification via the SOE module.

The SOE module first uses depthwise separable convolution to extract fine-grained spatial details with reduced computational overhead. Aided by the Contrastive Attention Mechanism (CAM, Section 3.2), it applies SimAM to filter channel-wise semantic correlations, followed by the LCDAtten mechanism to enhance brightness and texture differences between targets and their local neighborhoods, generating a spatial attention mask.

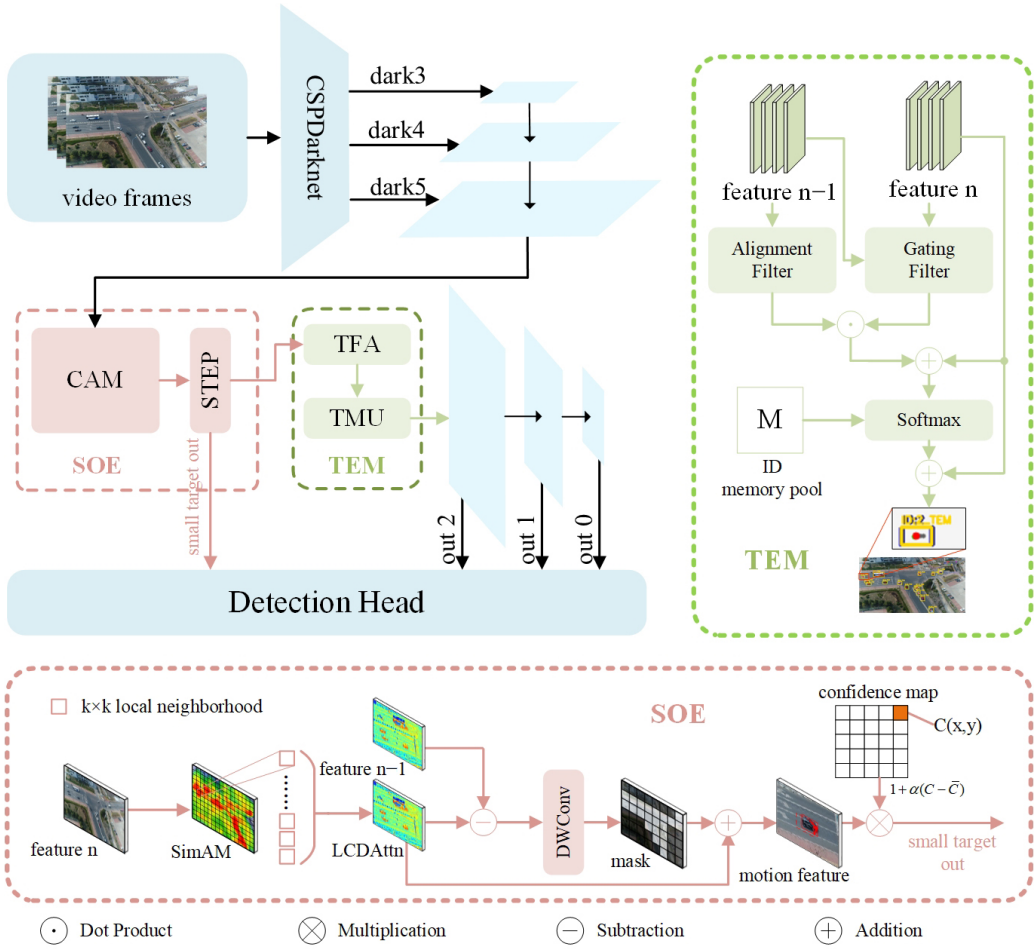


Figure 1. Overall framework of TASONet. TASONet is built on YOLOPAFPN with a CSPDarknet backbone and integrates the SOE module for spatial enhancement and the TEM module for temporal consistency. CSPDarknet extracts multi-level features fused by the FPN, after which SOE refines small object features and TEM models temporal cues through TFA and TMU. Their outputs are merged in the PAN and fed into the detection head for UAV small object tracking.

Following this, the STEP introduced in Section 3.3 incorporates inter frame motion difference cues and combines them with the detection confidence map to implement a confidence adaptive suppression strategy. This suppresses responses in low-confidence regions while strengthening features in high confidence small object areas. Ultimately, the SOE module outputs spatially enhanced features that retain key semantics, preserve structural details, and effectively suppress background noise.

Meanwhile, consecutive frame features from the backbone are transmitted to the TEM module, which combines the TFA mechanism (Section 3.4) with the TMU to establish temporal consistency. TFA employs gated convolution to dynamically regulate fusion weights between current and historical features, enabling lightweight inter-frame alignment without relying on optical flow. TMU maintains a sliding memory pool of high confidence

features and retrieves long term temporal information via QKV attention, compensating for feature degradation caused by occlusion or motion blur and generating temporally enhanced representations that mitigate single frame detection fluctuations.

Finally, the spatially enhanced SOE features and temporally enhanced TEM features are jointly fed into the Path Aggregation Network (PAN) for bottom-up secondary fusion, producing the final enhanced multi scale outputs (PAN out0, PAN out1, PAN out2). These feature maps are simultaneously delivered to the detection head, which predicts bounding boxes, categories, and confidence and to the ID branch, which leverages TMU's historical features for identity association. This forms a complete processing chain encompassing backbone feature extraction, FPN multi scale fusion, SOE spatial enhancement, TEM temporal modeling, PAN aggregation, and final detection tracking output, providing more stable and reliable detection inputs for subsequent small object tracking.

3.2. Contrastive Attention Mechanism

Small object detection is inherently challenged by low feature saliency and pronounced background interference. Conventional attention mechanisms or contrast enhancement methods often fail to address these two issues simultaneously. Channel attention can filter semantically irrelevant channels but lacks targeted enhancement of local spatial details, while local contrast methods can improve edge visibility but are prone to amplifying background noise in the absence of semantic guidance. To tackle this dilemma, we propose a CAM that focuses on scene-specific adaptation and the sequential collaboration of existing components. Through this design, CAM achieves joint optimization of channel-level semantic filtering and local spatial contrast enhancement tailored for small object scenarios.

Let $F \in \mathbb{R}^{C\times H\times W}$ denote the input feature map. CAM operates in two sequentially coupled stages, each specifically designed to accommodate the characteristics of small objects, so that their respective advantages complement each other while suppressing noise propagation.

In the first stage, SimAM is adopted for channel-level semantic refinement, and an additional false positive suppression constraint is introduced to mitigate the risk of noise amplification that commonly arises in small object settings. Following the original SimAM formulation, neuron-wise importance is evaluated based on a linear separability criterion, and the corresponding energy function is defined as:

$$E = \frac{(x-\mu)^2}{4(\sigma^2+\lambda)} + \frac{1}{2}, \tag{1}$$

where μ and σ^2 denote the mean and variance computed within each channel, and λ is a stabilizing constant. Minimizing E emphasizes channels with concentrated semantic responses, such as those corresponding to small object textures, while suppressing channels dominated by background noise. The energy is further transformed into a smooth attention response via a sigmoid function.

However, when SimAM is directly applied to small object scenarios, weak background responses may still be unintentionally amplified due to the low signal-to-noise ratio. Rather than modifying the original SimAM energy formulation, we introduce a learnable and constrained scaling parameter α to regulate the post-attention feature modulation:

$$F_s = F + \alpha \cdot (\mathrm{sigmoid}(E) - 0.5) \cdot F, \tag{2}$$

where α is initialized to 0.1 and constrained within $[0, 0.5]$ when the fp-suppress mechanism is enabled. This constraint limits excessive amplification of background responses and provides a more reliable semantic foundation for subsequent spatial refinement.

Based on the semantically filtered feature map F_s, a LCDAttn is constructed. Unlike static local normalization methods, LCDAttn incorporates learnable parameters to adaptively respond to variations in small object scale and local contrast. First, a normalized local contrast map (LCE) is computed within a 3×3 neighborhood, which is selected through grid search to better match small object characteristics:

$$LCE = \frac{F_s - \mu_{\text{local}}}{\sqrt{\sigma^2_{\text{local}} + \epsilon}}, \tag{3}$$

where μ_{local} and σ^2_{local} represent the local mean and variance, respectively, and ϵ is a stabilizing constant. Rather than directly applying the LCE as in traditional local normalization, a learnable scaling factor γ is introduced to regulate the contribution of contrast information and prevent over-amplification of trivial edge noise. An adaptive spatial attention map A is then generated as follows:

$$A = \text{sigmoid}(W(F_s + \gamma \cdot LCE)), \tag{4}$$

where W is a 1×1 convolution used to dynamically fuse the original semantic feature and the contrast information. This design ensures that contrast enhancement is concentrated on semantically meaningful regions guided by F_s, rather than being applied indiscriminately across the spatial domain. Finally, a learnable fusion ratio β, initialized to 0.1 and constrained within $[0, 0.5]$ under fp-suppress, is used to balance the original and contrast-enhanced features:

$$F_{\text{CAM}} = F_s \cdot (1 - \beta) + (F_s \cdot A) \cdot \beta, \tag{5}$$

this formulation avoids excessive distortion of small object representations while preserving the benefits of localized contrast enhancement.

The effectiveness of CAM arises from its scene-specific adaptation and the sequential synergy between its components, rather than from any individual module alone. Semantic filtering via SimAM is performed prior to contrast enhancement, ensuring that spatial refinement operates only on semantically relevant channels and reducing the risk of amplifying background noise in irrelevant regions. Design choices such as the 3×3 neighborhood size, the constrained ranges of α and β, and the learnable contrast scaling parameter γ are optimized for small objects at approximately 32×32 pixels, consistent with the definitions used in COCO and VisDrone. Compared with fixed local normalization or contrast scaling strategies, LCDAttn leverages an adaptive attention map and learnable parameters to dynamically adjust to background clutter and illumination variations, thereby improving robustness in UAV-based scenarios. Overall, CAM establishes a closed-loop process that integrates semantic guidance, contrast enhancement, and noise suppression, effectively mitigating the low discriminability of small objects without introducing significant computational overhead, as validated by ablation results demonstrating a notable reduction in false negatives (Section 4.4). The scaling parameters α and β are empirically constrained within the range [0, 0.5] to maintain a balanced attention modulation. In UAV based small object tracking scenarios, feature responses are inherently weak and easily affected by background clutter and illumination variations. Excessively large scaling factors tend to over amplify background regions and introduce unstable activations, leading to increased false positives. In contrast, moderate scaling provides sufficient enhancement for small objects while preserving the overall feature distribution. Therefore, α and β are fixed within this range across all experiments.

3.3. Small Target Enhancement Path for Adaptive False Detection Suppression

Despite improved spatial saliency, small-object features often remain fragile and prone to temporal degradation or misclassification. To further strengthen high-resolution representations, TASONet introduces a dedicated STEP integrated into the FPN. STEP combines three core strategies: depthwise separable convolutional refinement, motion aware feature modulation, and confidence guided regularization to achieve stable fine grained enhancement.

Let F_{cur} and F_{prev} denote the current and previous frame feature maps. The motion-aware enhancement process begins with computing inter-frame difference maps:

$$F_{\text{motion}} = F_{\text{cur}} + \beta \cdot \text{DWConv}\big(|F_{\text{cur}} - F_{\text{prev}}|\big) \cdot 1\big(|F_{\text{cur}} - F_{\text{prev}}| > \tau\big), \tag{6}$$

where DWConv denotes a depthwise separable convolution, τ is a motion sensitivity threshold used to suppress spurious background responses caused by camera jitter or illumination variations, and β controls fusion intensity. This operation adaptively accentuates motion-relevant regions, improving the visibility of small targets moving against dynamic backgrounds without introducing redundant flow estimation. The correlation between the feature space difference $|F_{\text{cur}} - F_{\text{prev}}|$ and actual motion arises from the distinct spatiotemporal characteristics of small objects and background variations in UAV scenarios. Specifically, small-object motion induces local, sparse feature changes concentrated in limited receptive fields, whereas global background fluctuations, such as illumination shifts or camera jitter, manifest as smooth, low-frequency patterns across the feature map. The depthwise separable convolution captures these fine-grained, motion-related variations by leveraging spatial locality, inherently suppressing global background noise while avoiding cross-channel interference. The motion sensitivity threshold τ is adaptively learned during training based on the motion statistics of small objects in the VisDrone-MOT and UAVDT datasets, ensuring that only meaningful motion differences exceeding the threshold are amplified and reliably captured.

To mitigate false positive propagation, an adaptive confidence regularization mechanism is further applied. Given a confidence map $C(x, y)$ and its global mean $\bar{C}$, the enhanced representation is expressed as:

$$F_{\text{enh}} = F_{\text{motion}} \cdot (1 + \alpha(C - \bar{C})), \tag{7}$$

where α controls the modulation strength. Regions with low detection confidence are automatically suppressed, while highly confident activations are selectively reinforced. To avoid potential self-reinforcing loops between feature enhancement and detection confidence, the confidence map $C(x, y)$ is derived from multi-scale FPN-PAN fusion outputs rather than single-scale features, reducing dependence on noisy early-stage representations. The modulation strength α is dynamically adjusted by the local contrast value LCE (computed via LCDAttn in Section 3.2), i.e., $\alpha = \alpha_0 \cdot \text{LCE}$ (α_0 is a learnable base parameter), which attenuates α for small objects overwhelmed by noise (low LCE) to prevent excessive suppression and retains appropriate enhancement intensity for high-contrast regions. We also impose L2 regularization on the term $(C - \bar{C})$ to avoid extreme values amplifying early stage noise. This self-regulating adjustment bridges semantic reliability and feature activation strength, effectively filtering out spurious responses caused by background clutter or illumination noise. The integration of motion guided amplification and confidence aware suppression enables STEP to simultaneously enhance discriminative features and maintain stability across varying environmental conditions.

3.4. Temporal Feature Alignment and Target Memory Modeling

Temporal coherence is crucial for stable small object detection in videos, especially under occlusion, deformation, or abrupt motion. TASONet thus integrates a two-level temporal modeling strategy: the TFA module for short-term adaptation and the TMU for long-term dependency modeling, both tailored to the unique challenges of UAV-based small object tracking. The TFA module achieves adaptive inter-frame feature alignment via gated convolutions, optimized to capture small objects' inter-frame motion magnitude without optical flow estimation. During training, an implicit inter-frame difference feature $F_{diff} = |F_{cur} - F_{prev}|$ quantifies small targets' motion intensity, larger differences indicate stronger motion blur or spatial displacement that easily degrades the inherently weak small-object features. Given features F_{cur} and F_{prev}, the alignment process is defined as:

$$F_{\text{aligned}} = F_{\text{cur}} + \sigma\Big(W_g * [F_{\text{cur}}, F_{\text{diff}}]\Big) \odot \Big(W_a * [F_{\text{cur}}, F_{\text{prev}}]\Big), \tag{8}$$

where W_g and W_a denote convolutional filters for gating and alignment, respectively, σ is the sigmoid activation, and $\odot$ represents element-wise multiplication. The gating branch $\sigma(W_g * [F_{\text{cur}}, F_{\text{diff}}])$ is trained to learn correlations between F_{cur} and F_{diff}, such as edge displacement and subtle texture variations of small objects, enabling it to regulate fusion adaptively according to motion magnitude. For fast-moving small targets with pronounced inter-frame changes, the gating value decreases to suppress interference from blurred features in the previous frame and prevent localization drift, while for static or slow-moving targets, the gating value increases to enhance feature fusion and reinforce temporal continuity. This lightweight formulation effectively balances accuracy and efficiency, which is essential for UAV platforms operating under constrained computational budgets.

To capture longer-range temporal dependencies and compensate for the intrinsic sparsity of small-object features, the TMU incorporates a sliding memory pool that stores high-confidence local discriminative features of small objects from preceding frames. Unlike general-purpose memory mechanisms such as the STMN family that store global frame features and apply uniform update strategies, the TMU focuses on preserving targeted small-object features to avoid information dilution by background clutter or large-object components. During inference, the current feature F_{cur} interacts with the memory through a query–key–value (QKV) attention mechanism. The query is produced by applying a 1×1 convolution to F_{cur} to compress channel dimensions while retaining essential semantics, and reshaping it into a spatial matrix. The keys and values are generated by linearly projecting the memory matrix M. This local feature-guided retrieval allows the weak current-frame small-object features to be reinforced by relevant historical cues, enhancing the separability between targets and background. The attention operation is formulated as:

$$F_{out} = F_{cur} + softmax\left(\frac{Q \cdot (KM)^T}{\sqrt{C}}\right) \cdot VM, \tag{9}$$

where $Q = F_{\text{cur}}$, and K and V denote the projected keys and values obtained from memory, with $\sqrt{C}$ serving as a normalization factor for numerical stability.

The TMU memory pool is updated selectively according to a frame-level confidence score s_t, defined as the average confidence of all detections in frame t. The update rule is expressed as:

$$M_t = \begin{cases} F_{out}, & \text{if } s_t > \delta, \\ M_{t-1}, & \text{otherwise}, \end{cases} \tag{10}$$

which prevents low-confidence or noisy features that are common in small-object detection from contaminating the memory pool, mitigating the noise accumulation issue observed in

STMN's uniform update strategy under complex backgrounds. The confidence threshold δ is optimized on the validation set and fixed to 0.5 in all experiments.

The sliding memory pool with size 20 retains only recent high-confidence features, and suspending updates during consecutive low-confidence frames helps preserve reliable historical representations and alleviate memory degradation. Although alternative schemes such as sliding-average aggregation or feature decay were investigated, they offered marginal improvements for small-object scenarios. Therefore, a confidence-gated sliding memory is adopted for robustness and stability. The design of the TMU involves two key hyperparameters: the memory size and the update threshold. The memory size determines the temporal coverage of historical features, while the update threshold controls the reliability of features being incorporated into the memory. Short memory lengths are insufficient to compensate for feature degradation under occlusion or missed detections, whereas overly long memories may accumulate outdated or noisy representations, especially in highly dynamic UAV scenes. Similarly, a permissive update threshold allows unreliable features to contaminate the memory, while an overly strict threshold restricts effective temporal aggregation. In practice, we observe that moderate values of these parameters lead to stable and robust tracking performance, and the selected configuration provides a good balance between temporal completeness and noise suppression.

By combining TFA and TMU, TASONet achieves dual-scale temporal reasoning. The short-term alignment ensures immediate consistency between consecutive frames for small objects vulnerable to motion blur, while the long-term memory supplements weak per-frame features and supports trajectory recovery after occlusion. This joint mechanism enables the network to maintain stable and robust representations of small objects even under dynamic, occluded, or visually degraded UAV conditions, as visually demonstrated in Figure 2.

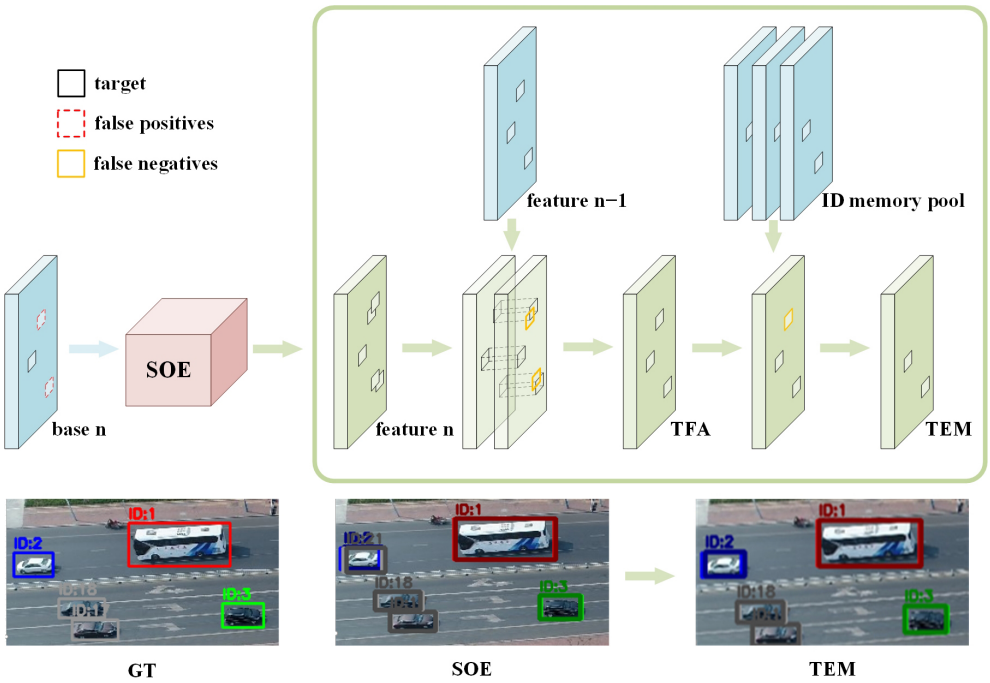


Figure 2. Optimization effects of the TEM module on object detection and tracking. By integrating inter-frame features and leveraging the identity memory pool to establish temporal associations, TEM effectively mitigates errors introduced by SOE and the detection process, thereby improving detection accuracy and temporal consistency, which consequently enhances tracking performance.

3.5. Comparative Analysis with Representative MOT Frameworks

To better contextualize the design choices of TASONet within the landscape of MOT, we provide a concise comparative analysis with representative frameworks such as BoT-SORT, OC-SORT, and DeepOC-SORT. These methods predominantly rely on global attention mechanisms or single-dimensional channel or spatial attention, which are effective for medium to large targets but tend to dilute fine-grained local contrast cues that are critical for detecting small objects ($\leq 32 \times 32$ pixels) in UAV imagery. As a result, small objects are easily overwhelmed by surrounding background textures during feature aggregation, leading to weak and unstable activations. Tracking-oriented frameworks typically emphasize association-stage optimization through motion cues or re-identification embeddings, but lack explicit spatial enhancement tailored to small object perception. In contrast, the proposed SOE module incorporates a CAM that combines channel-level semantic refinement with local contrast modeling. Specifically, SimAM suppresses semantically irrelevant responses, while LCDAttn captures localized luminance and texture differences through learnable 3×3 contrast patterns, thereby enhancing target saliency under cluttered UAV backgrounds.

In temporal modeling, existing approaches frequently depend on optical flow or Kalman-based alignment strategies, which implicitly assume relatively smooth motion patterns and stable viewpoints. However, these assumptions are often violated in UAV scenarios characterized by camera jitter, rapid motion, and frequent short-term occlusions, resulting in unstable alignment and trajectory fragmentation. To address this issue, TASONet adopts a lightweight dual-level temporal enhancement design. The TFA branch employs gated convolutions for adaptive inter-frame fusion without requiring explicit motion estimation, while the TMU maintains a confidence-gated sliding memory queue that selectively accumulates reliable historical representations. This design jointly strengthens short-term alignment and long-term temporal consistency, enabling more robust trajectory formation under challenging UAV dynamics.

From an efficiency perspective, complex spatiotemporal frameworks such as SSTM and BoostTrack substantially increase model parameters and reduce inference throughput, limiting their practical deployment on resource-constrained UAV platforms. TASONet maintains a lightweight profile via depthwise separable convolutions in SOE and compact gating operations in TEM, resulting in only a 2.88% increase in parameters and maintaining FLOPs at 7.12 G. The model sustains real-time inference speeds of 11.64 FPS on VisDrone-MOT and 12.68 FPS on UAVDT. Importantly, TASONet achieves these efficiency gains without relying on complex association strategies, multi-stage post-processing, or handcrafted filtering rules. This design choice not only simplifies the overall workflow but also reduces engineering implementation difficulty and improves inference stability, providing a clear advantage over methods such as BoT-SORT and DeepOC-SORT that depend heavily on post-processing for false detection suppression.

Furthermore, while many existing methods suppress false detections primarily during the association stage using post-filtering or fixed confidence thresholds, the proposed architecture mitigates noise earlier in the pipeline. The STEP integrates motion-difference cues, depthwise convolutions, and confidence-adaptive modulation to attenuate low-confidence activations directly at the detection stage. Combined with L2 activation regularization, this design stabilizes feature responses and effectively reduces false positives in complex UAV environments. By reducing the reliance on downstream post-processing, TASONet not only simplifies deployment but also improves the robustness and consistency of tracking results across diverse UAV scenes.

Overall, TASONet unifies spatial contrast enhancement, lightweight temporal alignment, and early-stage false detection suppression to directly address the core challenges of

small object tracking in UAV scenarios, including weak visual cues, background interference, and temporal instability. TASONet is particularly suited for medium- to high-altitude UAV scenarios with dense small targets, complex backgrounds, dynamic motion, and relatively stable viewpoints, such as aerial surveillance, traffic monitoring, and large-scale scene analysis. While its small-object-oriented design provides limited gains in low-altitude or highly volatile scenarios dominated by large targets, it offers clear engineering and practical advantages over general MOT methods within its targeted application domain.

4. Experiments

4.1. Dataset

We utilized the VisDrone-MOT [7] and UAVDT [33] datasets, both widely adopted for small object research in UAV based scenarios. VisDrone-MOT contains 10,209 images and 288 video sequences with over 260,000 frames and 2.6 million annotated bounding boxes, covering pedestrians, vehicles, and other object categories in complex urban and rural scenes. Its small, densely distributed, and frequently occluded targets under varying illumination make it a representative benchmark for evaluating small object detection and tracking. UAVDT consists of 100 drone captured video sequences over diverse urban environments, with precise vehicle annotations and attributes such as category and occlusion. The dataset's dynamic scenes and object scale variations complement VisDrone-MOT, providing a robust foundation for assessing detection and motion based tracking performance on small, fast moving objects from aerial perspectives.

4.2. Loss Functions for Detection and ID Branches

In our tracking framework, the model extends YOLOX by incorporating an additional ID branch for multi-object tracking. To optimize training, we adopt separate loss functions for the detection head and the ID branch, reflecting their distinct learning objectives. The detection head is responsible for predicting object bounding boxes and their classes. Following the design of YOLOX, its loss consists of three components: IoU loss for bounding box regression, which encourages the predicted boxes to closely overlap with the ground truth; objectness loss, implemented as binary cross-entropy, to distinguish foreground objects from background; and classification loss, which ensures correct prediction of object categories. The combined detection loss is denoted as

$$\mathcal{L}_{det} = \mathcal{L}_{iou} + \mathcal{L}_{obj} + \mathcal{L}_{cls}, \tag{11}$$

the ID branch generates appearance embeddings for each detected object to facilitate data association. These embeddings are optimized using a classification-based identity loss, where embeddings of the same object are encouraged to be close, and embeddings of different objects are pushed apart. This loss, denoted as $\mathcal{L}_{id}$, is typically implemented as cross-entropy over the object identity labels. The total loss used for training is a weighted combination of detection and identity losses,

$$\mathcal{L}_{total} = \mathcal{L}_{det} + \lambda \mathcal{L}_{id}, \tag{12}$$

the total loss is formulated as the sum of detection loss and identity-related loss, weighted by a scalar parameter λ. In all experiments, λ is fixed to 0.5. This value is selected based on empirical observations during training and validation, where it consistently provides a balanced optimization behavior. When λ is set too large, the optimization becomes dominated by identity supervision, leading to unstable gradients due to noisy appearance features, especially for small objects. Conversely, overly small λ values weaken identity constraints and result in fragmented trajectories. With $\lambda = 0.5$, the training process exhibits

stable convergence across epochs and consistent validation performance, indicating that the proposed framework does not rely on fine-grained tuning of this parameter.

4.3. Evaluation Metrics

4.3.1. False Positives

False positives (FP) quantify the number of non-existent objects erroneously detected or localized by a model. In object detection, an FP refers to a predicted bounding box whose IoU with any ground truth falls below a defined threshold. Within tracking tasks, FPs arise from both detection level errors and association level errors. Elevated FP counts reduce tracking precision and degrade MOTA performance, often reflecting overfitting or inadequate discrimination between targets and background clutter.

4.3.2. False Negatives

False negatives (FN) measure a model's inability to detect true targets. In detection tasks, they occur when an object is missed or its predicted box fails to meet the IoU threshold. In tracking, FNs include both detection level misses and association level losses caused by trajectory discontinuities or identity confusion. A high FN rate weakens recall and lowers overall MOTA. Reducing FNs typically involves refining detection thresholds, strengthening feature extraction in complex scenes, and enhancing temporal association stability.

4.3.3. Identity Switches

Identity Switches (IDs) represent the number of times a tracker incorrectly changes the identity of a target during tracking. An identity switch occurs when the same object is assigned different IDs across frames or when an ID is reassigned to another object. Frequent IDs break trajectory continuity and reduce identity consistency, leading to performance degradation in long term tracking. High IDs values usually indicate weaknesses in data association or feature discrimination. Reducing identity switches requires stronger appearance modeling, temporal consistency enhancement, and more reliable association strategies.

4.3.4. Multiple Object Tracking Accuracy

The Multiple Object Tracking Accuracy (MOTA) metric provides a comprehensive evaluation of tracking performance by jointly considering FN, FP and IDs. It is computed as

$$\text{MOTA} = 1 - \frac{\sum(FN_i + FP_i + IDs_i)}{\sum GT_i}, \tag{13}$$

MOTA theoretically ranges from negative infinity to 1, with higher values indicating superior performance. The metric primarily reflects global sensitivity to detection errors, as false negatives and false positives exert a greater influence on overall performance than identity mismatches. Due to its interpretability and integrative nature, MOTA remains one of the most widely adopted indicators in MOT evaluation.

4.3.5. IDF1

The Identity F1 score (IDF1) measures the balance between detection accuracy and identity preservation by taking the harmonic mean of identity precision and recall. It evaluates how effectively a tracker maintains consistent identities throughout trajectories. High IDF1 scores indicate robust identity continuity and low switching rates, which are essential in long-duration or crowded scenes. Improving IDF1 generally requires simultaneous enhancement of detection reliability and association robustness to minimize identity fragmentation.

4.3.6. MT

The Mostly Tracked (MT) metric quantifies the proportion of ground-truth trajectories that are successfully tracked for at least 80% of their lifespan. It reflects a tracker's ability to maintain long, continuous trajectories without frequent interruptions. High MT values indicate stable temporal consistency and reliable tracking coverage across sequences. Improving MT generally requires strengthening detection recall and enhancing association stability to reduce trajectory breaks and maintain sustained tracking performance.

4.3.7. ML

The Mostly Lost (ML) metric measures the proportion of ground-truth trajectories that are tracked for less than 20% of their duration. It evaluates how often a tracker fails to capture targets consistently, resulting in substantial temporal discontinuities. Low ML scores indicate effective handling of challenging cases and reduced trajectory loss across frames. Reducing ML typically requires improving the tracker's ability to re-detect objects after occlusions and mitigating association failures that lead to early trajectory termination.

4.4. Ablation Experiments

To systematically validate the independent contributions and synergistic effects of the core modules in TASONet, we conducted ablation experiments based on the ByteTrack framework. The original ByteTrack served as the baseline, while model variants were constructed to isolate the effects of SimAM, LCDAttn, SOE, TFA, TMU, TEM, and their combination (SOE + TEM). All experiments were performed under identical datasets (VisDrone-MOT and UAVDT) and evaluation conditions. For fair comparison and reproducibility, all models were trained using the same optimization settings following the standard YOLOX training protocol. Specifically, the batch size was set to 8, mixed-precision (FP16) training was enabled, and the network was optimized using stochastic gradient descent with the default YOLOX learning rate schedule. The initial learning rate was set to 0.01 and decayed using a cosine annealing strategy over a total of 50 training epochs. The key quantitative results are summarized in Tables 1 and 2.

Table 1. Results of ablation experiments on the VisDrone-MOT dataset.

Method	MOTA ↑	MT ↑	ML ↓	FP ↓	FPR ↓	FN ↓	IDs ↓	IDF1 ↑	FPS ↑
BASE	61.15	732	427	3438	2.79	65,828	4198	77.58	11.78
SimAM	73.12	768	371	2984	2.06	44,050	3792	84.30	12.01
LCDAttn	73.06	738	419	3453	2.37	43,455	4042	84.84	11.91
SOE (SimAM + LCDAttn)	73.37	749	400	3032	2.07	**43,009**	4322	85.14	11.73
TFA	64.66	783	330	1331	1.06	63,170	2329	79.44	**13.34**
TMU	65.68	767	**319**	1284	1.01	61,598	**2015**	80.06	13.18
TEM (TFA + TMU)	65.63	683	417	**1211**	**0.95**	61,729	2060	80.04	13.30
SOE + TEM	**73.52**	**807**	380	2465	1.71	45,007	2609	**85.64**	11.64

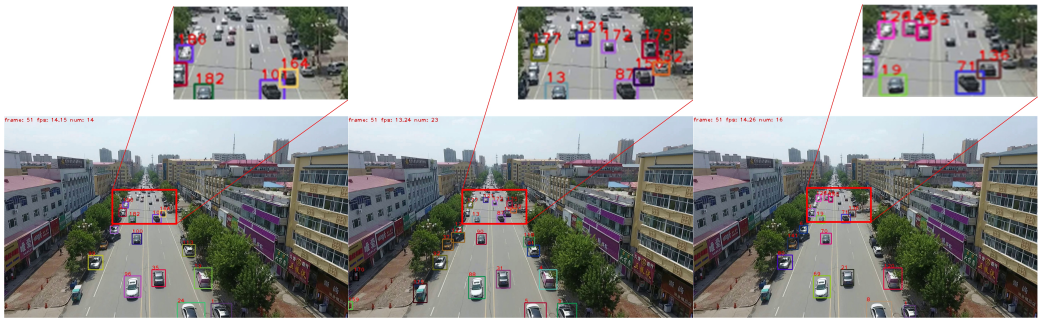
↑ indicates that a higher value is preferable; ↓ indicates that a lower value is preferable. Bold numbers denote the best performance.

The baseline model exhibits clear limitations in handling small objects. On VisDrone-MOT, it obtains a MOTA of 61.15% and an IDF1 of 77.58%, with high FN (65,828) and IDs (4198), indicating insufficient feature discriminability and weak temporal consistency. As illustrated in Figure 3a, the baseline frequently misses distant small objects, which is consistent with its elevated FN (65,828 on VisDrone-MOT; 41,120 on UAVDT). On UAVDT, the baseline reaches 68.33% MOTA with 5883 identity switches, and its trajectories display notable fragmentation (Figure 5a), further confirming the absence of temporal modeling. Although its inference speed (11.78 FPS on VisDrone-MOT and 12.63 FPS on UAVDT) satisfies basic real-time requirements, the accuracy remains inadequate for complex aerial scenarios.

Table 2. Results of ablation experiments on the UAVDT dataset.

Method	MOTA ↑	MT ↑	ML ↓	FP ↓	FPR ↓	FN ↓	IDs ↓	IDF1 ↑	FPS ↑
BASE	68.33	324	52	8646	6.42	41,120	5883	83.50	12.63
SimAM	72.00	383	36	14,192	9.61	28,141	6873	86.31	13.09
LCDAttn	72.05	403	32	15,186	10.22	27,293	6651	86.26	12.62
SOE (SimAM + LCDAttn)	73.31	395	45	14,043	9.35	**25,726**	7148	87.24	11.91
TFA	72.42	397	30	7535	5.39	36,085	4853	85.83	**14.40**
TMU	72.51	424	**18**	8144	5.79	35,226	4933	85.92	14.10
TEM (TFA + TMU)	72.69	**429**	39	8319	5.92	35,262	**4418**	85.84	14.34
SOE + TEM	**75.97**	373	31	**6459**	**4.42**	29,758	6011	**88.51**	12.68

↑ indicates that a higher value is preferable; ↓ indicates that a lower value is preferable. Bold numbers denote the best performance.



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

Figure 3. Visual comparison of distant small object detection under UAV scenarios. (**a**) Baseline ByteTrack frequently misses distant and low-contrast small objects due to weak feature responses. (**b**) SOE significantly improves small object visibility by enhancing local contrast and suppressing background interference through CAM and STEP. (**c**) SOE + TEM preserves the spatial enhancement benefits while further reducing false positives by enforcing temporal consistency across frames.

Individually integrated modules yield distinct improvements, corroborated by the visualization results. For spatial enhancement, both SimAM and LCDAttn strengthen small-object feature responses and significantly reduce FN. SimAM improves MOTA to 73.12% (VisDrone-MOT) and 72.00% (UAVDT), effectively reducing FN to 44,050 and 28,141. LCDAttn achieves the lowest FN among individual spatial modules on VisDrone-MOT (43,455) and a high IDF1 (84.84). Their combination, SOE, further improves feature discrimination, achieving the lowest FN of all models on VisDrone-MOT (43,009) and maintaining competitive accuracy on UAVDT with FN 25,726. Beyond FN reduction, the spatial modules also exhibit measurable improvements in FP control. Compared with the baseline ByteTrack, SimAM decreases the false positive rate (FPR) by suppressing background activations introduced by cluttered scenes, while LCDAttn yields the largest FPR reduction among all single spatial modules by enhancing local contrast and preventing high-frequency noise from being misclassified as objects. When combined, SOE achieves the lowest FPR on VisDrone-MOT, confirming that SimAM and LCDAttn work synergistically to suppress background-induced false alarms. Furthermore, the STEP branch embedded within SOE reinforces this trend by integrating motion-difference cues and confidence-adaptive suppression, which selectively attenuate unstable low-confidence activations. This targeted suppression mechanism directly contributes to lowering FPR, as noisy background responses or jitter-induced fluctuations are filtered before reaching the detection head.

As illustrated in Figure 3, the baseline ByteTrack exhibits frequent missed detections for distant small objects, which is consistent with its high FN count reported in Table 1. These failures mainly occur in low-contrast regions where small targets are easily overwhelmed by background textures. After introducing SOE, small objects become more distinguishable,

as shown in Figure 3b. This improvement can be attributed to the CAM, where SimAM suppresses semantically irrelevant channel responses and LCDAttn enhances local luminance and texture contrast around small objects. In addition, the STEP branch further stabilizes these activations by selectively reinforcing motion consistent regions while attenuating low confidence background responses, leading to a noticeable reduction in missed detections. When TEM is further integrated (Figure 3c), false positives are additionally suppressed through temporal constraints, indicating that temporal modeling effectively regularizes spatial enhancement results across frames.

Figure 4 provides heatmap visualizations of intermediate features before and after SOE processing. Prior to enhancement, small-object regions exhibit low contrast and are easily suppressed by background noise. After SOE enhancement, SimAM amplifies semantically relevant feature channels, while LCDAttn enhances local brightness and texture contrast, producing stronger feature responses in small object regions and effectively suppressing background clutter.

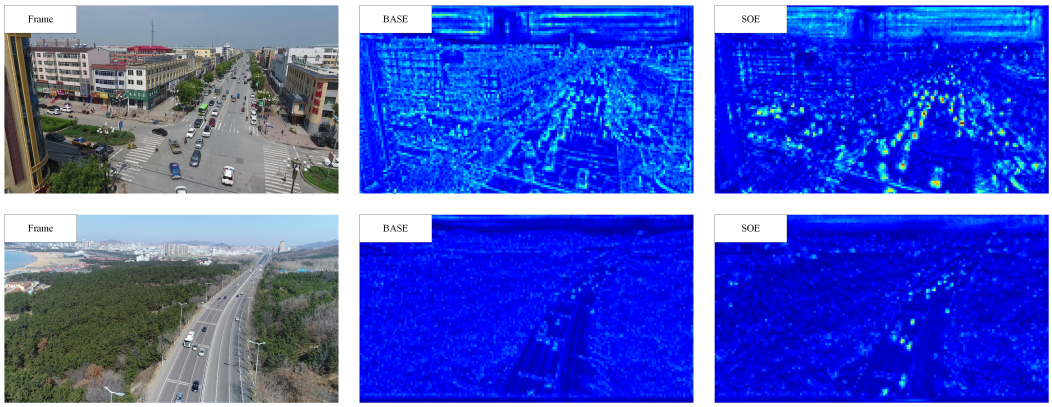
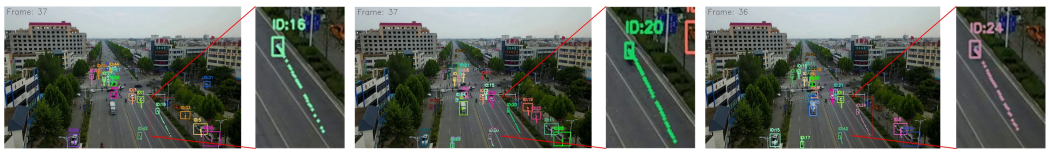


Figure 4. Comparison of intermediate feature heatmaps before and after SOE processing. Before enhancement, small object regions exhibit weak and diffuse activations that are easily submerged by background noise. After SOE, feature responses corresponding to small objects become more concentrated and salient, while background activations are suppressed.

For temporal modeling, TFA and TMU significantly improve trajectory stability and reduce false positives. TMU achieves the lowest ML (319 on VisDrone-MOT; 18 on UAVDT) and the lowest IDs (2015 on VisDrone-MOT). The integrated TEM module further suppresses FP, achieving the lowest FP of all variants on VisDrone-MOT (1211). As shown in Figure 5b, TEM produces smoother and more continuous trajectories than the baseline, particularly under motion blur or brief occlusions, consistent with the reduced IDs and ML. TFA mitigates localization drift for fast-moving targets, while TMU preserves high-confidence memory across frames, jointly contributing to improved MT and reduced ML.



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

Figure 5. Effect of the TEM module on MOT performance. (**a**) Baseline ByteTrack produces fragmented trajectories due to the absence of temporal feature modeling. (**b**) TEM improves trajectory continuity and identity consistency by aligning inter-frame features and leveraging historical target memory. (**c**) TEM improves trajectory continuity and identity consistency by aligning inter-frame features and leveraging historical target memory.

The combination of SOE and TEM yields the best overall performance on both datasets. On VisDrone-MOT, SOE + TEM achieves the highest MOTA (73.52%) and IDF1 (85.64%), with the largest MT (807) and strong suppression of FP and FN. On UAVDT, it similarly obtains the highest MOTA (75.97%) and IDF1 (88.51%), along with the lowest FP (6459). Visually, Figure 3c shows that SOE + TEM not only inherits SOE's ability to detect distant small objects, but also substantially reduces false positives through temporal consistency constraints, matching the FP reduction trends. Figure 5c demonstrates that SOE + TEM achieves precise small-object detection with stable, continuous trajectories and minimal identity switches. Notably, its FPS (11.64 on VisDrone-MOT; 12.68 on UAVDT) remains comparable to the baseline, validating the lightweight nature of the design.

In summary, the ablation results and corresponding visualizations confirm that each core module contributes independently: SOE enhances small-object feature representation to reduce FN, while TEM strengthens temporal coherence to suppress FP and IDs. Their combination effectively balances spatial enhancement and temporal stability, achieving joint improvements in accuracy, identity consistency, and real-time performance.

4.5. Comparison to State-of-the-Art Methods

To further validate the overall superiority and generalization capability of the proposed framework, we conducted a comprehensive comparison with several representative state-of-the-art MOT methods, including BoTSORT, OCSORT, DeepOCSORT, BoostTrack, and Improved Association using the original ByteTrack as the baseline. All methods were evaluated under identical experimental settings, ensuring consistent detection inputs, hardware configurations, and frame rates for fairness. The quantitative results on UAVDT and VisDrone datasets are summarized in Tables 3 and 4.

Table 3. Results of UAVDT comparative experiments.

Method	MOTA ↑	FP + FN ↓	IDs ↓	IDF1 ↑	FPS ↑
ByteTrack	68.33	49,766	5883	83.50	12.63
BoTSORT	64.20	59,444	3468	79.64	**13.71**
OCSORT	59.45	68,767	2499	75.68	17.42
DeepOCSORT	60.70	66,438	2639	76.70	13.68
BoostTrack	50.02	86,120	**1727**	67.55	13.43
ImprAssoc	64.19	55,208	7729	81.37	11.71
Ours	**75.97**	**36,217**	6011	**88.51**	12.68

↑ indicates that a higher value is preferable; ↓ indicates that a lower value is preferable. Bold numbers denote the best performance.

As shown in Table 3, on the UAVDT dataset, the proposed method achieves remarkable improvements across all key evaluation metrics. Compared with the baseline ByteTrack, our method increases MOTA from 68.33 to 75.97, representing a relative gain of 11.18 percent, reduces FP + FN from 49766 to 36217, corresponding to a 27.23 percent decrease, and improves IDF1 from 83.50 to 88.51, marking a 6.00 percent increase. While maintaining a comparable real-time speed of 12.68 FPS compared with 12.63 FPS for ByteTrack, our method demonstrates competitive performance against other advanced trackers. Compared with BoTSORT, which achieves 64.20 MOTA, our method achieves an 18.33 percent improvement. Against OCSORT with 59.45 MOTA, the improvement reaches 27.79 percent. Even when compared with DeepOCSORT and ImprAssoc, which achieve 60.70 and 64.19 MOTA respectively, our method leads by 15.27 percent and 18.35 percent. The significant reduction in FP + FN and stable IDs performance highlight the enhanced detection robustness and identity preservation capability, which is especially beneficial for small and fast-moving targets in UAV scenarios.

For the VisDrone dataset, our method exhibits consistent superiority. Compared with ByteTrack, which achieves 61.15 MOTA, our method reaches 73.52 MOTA, representing a 20.23 percent relative improvement. FP + FN is reduced from 69,266 to 47,472, a decrease of 31.46 percent, and IDF1 increases from 84.52 to 88.83, a gain of 5.10 percent, while maintaining a real-time speed of 11.64 FPS. When benchmarked against other methods, our approach outperforms BoTSORT with 52.53 MOTA by 39.96 percent, OCSORT with 47.95 MOTA by 53.33 percent, DeepOCSORT with 58.79 MOTA by 25.05 percent, and ImprAssoc with 67.73 MOTA by 8.55 percent. On the challenging VisDrone dataset with dense and diverse targets, our method's ability to handle complex scenes is evident from the substantial reduction in FP + FN and the high IDF1 score, indicating robust temporal consistency and identity association.

Table 4. Results of VisDrone comparative experiments.

Method	MOTA ↑	FP + FN ↓	IDs ↓	IDF1 ↑	FPS ↑
ByteTrack	61.15	69,266	4198	77.58	11.78
BoTSORT	52.53	89,668	**806**	75.10	12.51
OCSORT	47.95	96,717	1649	65.64	**15.92**
DeepOCSORT	58.79	75,849	2167	74.92	12.33
BoostTrack	42.68	107,375	1021	60.36	12.39
ImprAssoc	67.73	56,527	4497	82.45	11.32
Ours	**73.52**	**47,472**	2609	**85.64**	11.64

↑ indicates that a higher value is preferable; ↓ indicates that a lower value is preferable. Bold numbers denote the best performance.

To intuitively assess the tracking performance, we analyze the visualization results in Figure 6. The ground truth trajectory plot serves as the reference for accurate object paths. Compared with other algorithms, ByteTrack shows relatively fragmented trajectories, especially for distant or small objects, indicating potential identity switches or missed detections. BoTSORT and OCSORT display somewhat more coherent trajectories, but there are noticeable gaps in coverage for some lanes and small objects, reflecting limitations in handling dense and small objects. DeepOCSORT and BoostTrack exhibit inconsistent trajectory continuity with occasional breaks or misalignments, which can be attributed to insufficient temporal feature integration or computational overhead affecting real-time detection. ImprAssoc displays moderate trajectory integrity but still lags behind in covering all targets comprehensively, especially in crowded areas. Our method achieves trajectories that are much closer to the ground truth in terms of continuity, coverage, and consistency. It effectively tracks targets across the entire scene, including small and fast-moving objects, with minimal fragmentation. This visual consistency aligns with the quantitative improvements, confirming that the integrated modules enhance both spatial feature representation and temporal motion modeling, leading to more reliable and accurate tracking trajectories.

In summary, across both UAVDT and VisDrone datasets, as well as through extensive visual validation, the proposed framework demonstrates superior tracking accuracy, strong identity consistency, and real-time performance. These results validate that integrating small object enhancement and temporal motion modeling at the detector level is a promising solution for complex MOT scenarios, effectively outperforming state-of-the-art methods that rely solely on association strategies or downstream motion modeling. Moreover, the key hyperparameters in our framework are determined according to task-specific characteristics and empirical observations on the validation set, and the overall performance remains stable within reasonable parameter ranges, indicating that the proposed method does not depend on finely tuned hyperparameters.

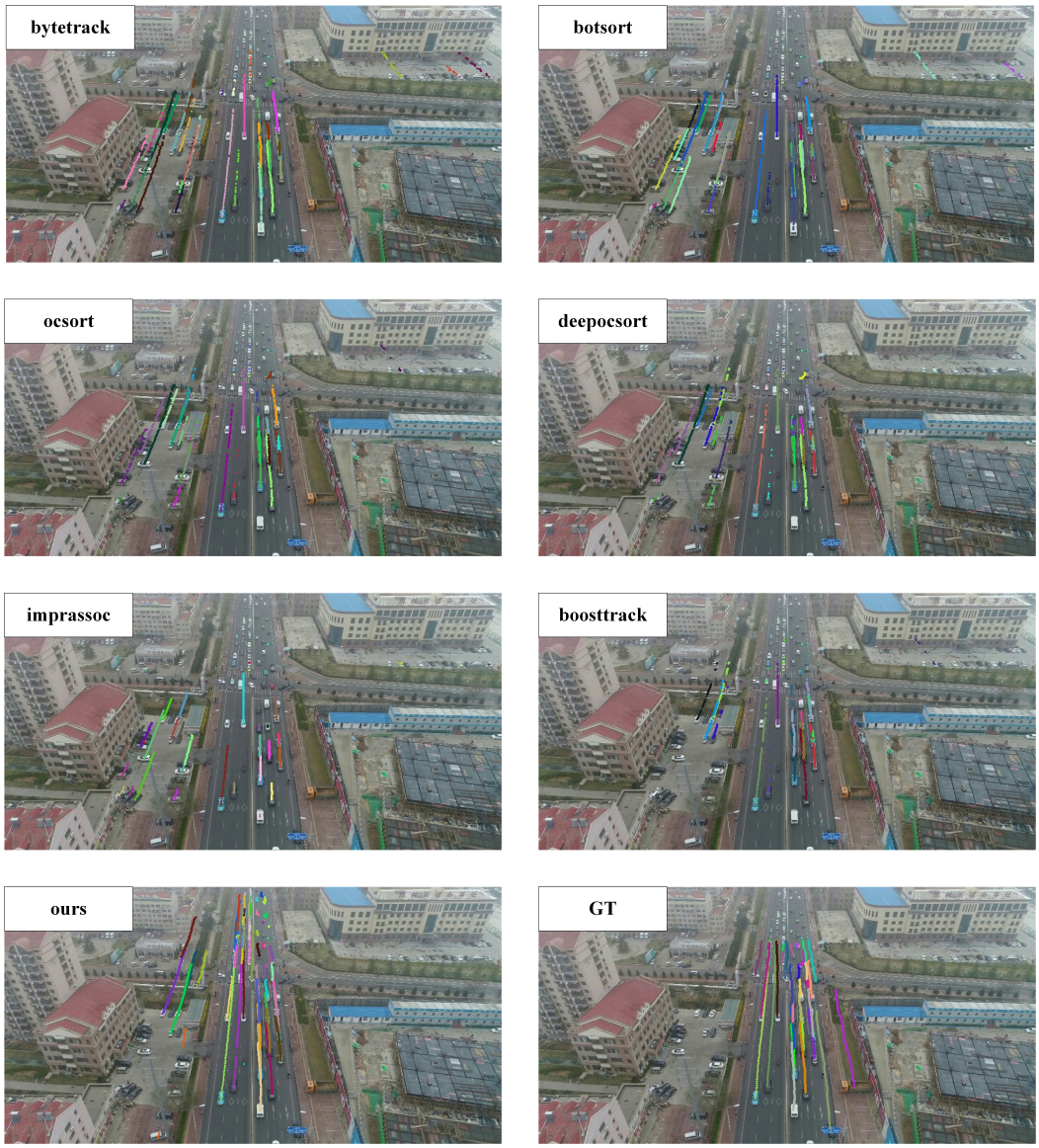


Figure 6. Qualitative comparison of tracking trajectories produced by different MOT algorithms on UAV scenarios. Compared with association-focused methods such as BoTSORT and OCSORT, which exhibit trajectory fragmentation or incomplete coverage for small and distant targets, the proposed method achieves trajectories that more closely align with ground truth in terms of continuity and coverage.

4.6. Computational Efficiency Evaluation

To comprehensively assess the computational overhead introduced by each module, we report the model parameters and FLOPs in Table 5. All experiments are conducted on a workstation equipped with an NVIDIA RTX 1080Ti GPU and an Intel Core i7-6850K CPU. As shown in Table 5, the baseline network contains 12.52 M parameters and 6.0 G FLOPs. Integrating SimAM does not increase the parameter count while slightly raising FLOPs to 6.47 G. The LCDAttn module introduces a small overhead, and the combination of SimAM and LCDAttn in the SOE module results in 12.59 M parameters and 6.97 G FLOPs.

For temporal modeling, TFA and TMU individually add moderate overhead, and their integration as the TEM module yields 12.81 M parameters and 6.15 G FLOPs. Combining SOE and TEM leads to the full TASONet, which contains 12.88 M parameters and 7.12 G FLOPs, only a 2.88% increase over the baseline. This demonstrates that TASONet preserves high computational efficiency while offering enhanced performance.

Table 5. Computational efficiency comparison of TASONet modules.

Method	Params (M)	FLOPs (G)
BASE	12.52	6.00
SimAM	12.52	6.47
LCDAttn	12.59	6.50
SOE (SimAM + LCDAttn)	12.59	6.97
TFA	12.76	6.10
TMU	12.57	6.05
TEM (TFA + TMU)	12.81	6.15
SOE + TEM	12.88	7.12

5. Conclusions

This study presents an improved ByteTrack framework for MOT, targeting the limitations in small object representation, temporal modeling, and robustness against occlusion and background interference. The objective was to enhance small object tracking accuracy while maintaining real-time performance for UAV surveillance and intelligent transportation applications.

The proposed framework integrates three core components: the SOE module, the TEM module, and a detector-level motion integration mechanism. The SOE module strengthens small-object features via global-local attention and a high-resolution detection branch; the TEM module (TFA + TMU) improves temporal consistency and enables track recovery after occlusion; the motion integration mechanism embeds inter-frame motion cues directly into the detector, reducing false negatives and enhancing feature continuity. Experimental results show the SOE and TEM modules increase MOTA from 61.15% (VisDrone-MOT) and 68.33% (UAVDT) to 73.52% and 75.97% respectively, while significantly reducing false negatives and mostly lost targets and retaining real-time performance. Compared with representative state-of-the-art methods, including BoTSORT, OCSORT, DeepOCSORT, BoostTrack, and Improved Association, the proposed framework achieved MOTA values of 73.52% and 75.97% and IDF1 scores of 88.83% and 88.51% on the VisDrone and UAVDT datasets, respectively, confirming its superior accuracy and efficiency.

Theoretically, this work unifies small object enhancement and temporal modeling within the detection stage, providing an integrated tracking paradigm beyond conventional association based designs. Practically, its lightweight and adaptable architecture (12.88 M parameters, 2.88% increase over baseline) ensures reliable performance in UAV based surveillance and intelligent transportation scenarios where small object tracking is critical for real-time decision making. Future work will focus on adaptive background suppression, robustness under adverse environmental conditions, and multimodal data fusion to further enhance detection reliability in complex real-world environments.

Author Contributions: Conceptualization, R.M. and C.L.; methodology, R.M., Q.S. and X.L.; software, R.M.; validation, Z.T.; formal analysis, R.M. and C.L.; investigation, R.M. and Z.T.; resources, Q.S.; data curation, R.M. and C.L.; writing—original draft preparation, R.M.; writing—review and editing, Q.S., Z.T. and X.L.; visualization, R.M. and C.L.; project administration, X.L.; funding acquisition, X.L. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Major Horizontal Project No. KM20230861.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data are contained within the article.

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

References

1. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You Only Look Once: Unified, Real-Time Object Detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016.
2. Redmon, J.; Farhadi, A. YOLOv3: An Incremental Improvement. *arXiv* **2018**, arXiv:1804.02767. [CrossRef]
3. Bochkovskiy, A.; Wang, C.Y.; Liao, H.Y.M. YOLOv4: Optimal Speed and Accuracy of Object Detection. *arXiv* **2020**, arXiv:2004.10934. [CrossRef]
4. Wang, C.Y.; Bochkovskiy, A.; Liao, H.Y.M. YOLOv7: Trainable Bag-of-Freebies Sets New State-of-the-Art for Real-Time Object Detectors. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 18–22 June 2023; pp. 7464–7475.
5. Ge, Z.; Liu, S.; Wang, F.; Li, Z.; Sun, J. YOLOX: Exceeding YOLO Series in 2021. *arXiv* **2021**, arXiv:2107.08430. [CrossRef]
6. Lin, T.Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft COCO: Common Objects in Context. In *Proceedings of the Computer Vision–ECCV 2014*; Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T., Eds.; Springer: Cham, Switzerland, 2014; pp. 740–755.
7. Wen, L.; Zhu, P.; Du, D.; Bian, X.; Ling, H.; Hu, Q.; Zheng, J.; Peng, T.; Wang, X.; Zhang, Y.; et al. VisDrone-MOT2019: The Vision Meets Drone Multiple Object Tracking Challenge Results. In *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*; IEEE: Piscataway, NJ, USA, 2019; pp. 189–198. [CrossRef]
8. Hu, J.; Shen, L.; Albanie, S.; Sun, G.; Wu, E. Squeeze-and-Excitation Networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 7132–7141.
9. Woo, S.; Park, J.; Lee, J.Y.; Kweon, I.S. CBAM: Convolutional Block Attention Module. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 3–19.
10. Wang, Q.; Wu, B.; Zhu, P.; Li, P.; Zuo, W.; Hu, Q. ECA-Net: Efficient Channel Attention for Deep Convolutional Neural Networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 14–19 June 2020; pp. 11534–11542.
11. Shi, X.; Huang, Z.; Bian, W.; Li, D.; Zhang, M.; Cheung, K.C.; See, S.; Qin, H.; Dai, J.; Li, H. VideoFlow: Exploiting Temporal Cues for Multi-frame Optical Flow Estimation. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Paris, France, 1–6 October 2023; pp. 12469–12480.
12. Chen, Y.; Zhu, D.; Shi, W.; Zhang, G.; Zhang, T.; Zhang, X.; Li, J. MFCFlow: A Motion Feature Compensated Multi-Frame Recurrent Network for Optical Flow Estimation. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), Waikoloa, HI, USA, 3–7 January 2023; pp. 5068–5077.
13. Ferede, F.A.; Balasubramanian, M. SSTM: Spatiotemporal recurrent transformers for multi-frame optical flow estimation. *Neurocomputing* **2023**, *558*, 126705. [CrossRef]
14. Bewley, A.; Ge, Z.; Ott, L.; Ramos, F.; Upcroft, B. Simple online and realtime tracking. In Proceedings of the 2016 IEEE International Conference on Image Processing (ICIP), Phoenix, AZ, USA, 25–28 September 2016; pp. 3464–3468. [CrossRef]
15. Wojke, N.; Bewley, A.; Paulus, D. Simple online and realtime tracking with a deep association metric. In Proceedings of the 2017 IEEE International Conference on Image Processing (ICIP), Beijing, China, 17–20 September 2017; pp. 3645–3649. [CrossRef]
16. Zhang, Y.; Wang, C.; Wang, X.; Zeng, W.; Liu, W. FairMOT: On the Fairness of Detection and Re-identification in Multiple Object Tracking. *Int. J. Comput. Vis.* **2021**, *129*, 3069–3087. [CrossRef]
17. Zhang, Y.; Sun, P.; Jiang, Y.; Yu, D.; Weng, F.; Yuan, Z.; Luo, P.; Wang, X. ByteTrack: Multi-object Tracking by Associating Every Detection Box. In *Proceedings of the Computer Vision–ECCV 2022*; Lecture Notes in Computer Science; Avidan, S., Brostow, G., Cissé, M., Farinella, G., Hassner, T., Eds.; Springer: Cham, Switzerland, 2022; Volume 13682. [CrossRef]
18. Aharon, N.; Orfaig, R.; Bobrovsky, B.Z. BoT-SORT: Robust Associations Multi-Pedestrian Tracking. *arXiv* **2022**, arXiv:2206.14651.
19. Cao, J.; Pang, J.; Weng, X.; Khirodkar, R.; Kitani, K. Observation-centric sort: Rethinking sort for robust multi-object tracking. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 18–22 June 2023; pp. 9686–9696.
20. Maggiolino, G.; Ahmad, A.; Cao, J.; Kitani, K. Deep OC-Sort: Multi-Pedestrian Tracking by Adaptive Re-Identification. In *2023 IEEE International Conference on Image Processing (ICIP)*; IEEE: Piscataway, NJ, USA, 2023; pp. 3025–3029. [CrossRef]
21. Stanojevic, V.; Todorovic, B. BoostTrack: Boosting the similarity measure and detection confidence for improved multiple object tracking. *Mach. Vis. Appl.* **2024**, *35*, 53. [CrossRef]
22. Jiang, J.; Wang, Z.; Zhao, M.; Li, Y.; Jiang, D. SAM2MOT: A Novel Paradigm of Multi-Object Tracking by Segmentation. *arXiv* **2025**, arXiv:2504.04519.
23. Ji, D.; Zhu, L.; Gao, S.; Zhu, Q.; Zhao, Y.; Xu, P.; Ding, Y.; Lu, H.; Ye, J.; Wu, F.; et al. View-Centric Multi-Object Tracking with Homographic Matching in Moving UAV. *IEEE Trans. Geosci. Remote Sens.* 2026, *Early Access*. [CrossRef]
24. Fu, Z.; Han, X.; Jia, M.; Zeng, Q.; Zahng, M.; Wang, C.; Meng, W.; Zhang, X. DMTrack: Deformable State-Space Modeling for UAV Multi-Object Tracking with Kalman Fusion and Uncertainty-Aware Association. *arXiv* **2025**, arXiv:2510.17860.

25. Ma, J.; Luo, H.; Chen, Q.; Qi, Y.; Sun, Y.; Beheshti, A.; Zhang, J.; Yang, M.H. Tracking the Unstable: Appearance-Guided Motion Modeling for Robust Multi-Object Tracking in UAV-Captured Videos. *arXiv* **2025**, arXiv:2508.01730.
26. Tan, M.; Pang, R.; Le, Q.V. EfficientDet: Scalable and Efficient Object Detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 14–19 June 2020; pp. 10781–10790.
27. Tian, Z.; Shen, C.; Chen, H.; He, T. FCOS: A Simple and Strong Anchor-Free Object Detector. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *44*, 1922–1933. [CrossRef] [PubMed]
28. Howard, A.; Sandler, M.; Chu, G.; Chen, L.C.; Chen, B.; Tan, M.; Wang, W.; Zhu, Y.; Pang, R.; Vasudevan, V.; et al. Searching for MobileNetV3. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*; IEEE: Piscataway, NJ, USA, 2019; pp. 1314–1324.
29. Ding, X.; Zhang, X.; Ma, N.; Han, J.; Ding, G.; Sun, J. RepVGG: Making VGG-Style ConvNets Great Again. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 19–25 June 2021; pp. 13733–13742.
30. Yang, L.; Zhang, R.Y.; Li, L.; Xie, X. SimAM: A Simple, Parameter-Free Attention Module for Convolutional Neural Networks. In *Proceedings of the 38th International Conference on Machine Learning*; Proceedings of Machine Learning Research; Meila, M., Zhang, T., Eds.; Curran Associates, Inc.: Red Hook, NY, USA, 2021; Volume 139, pp. 11863–11874.
31. Sadoughi, N.; Li, X.; Vajpayee, A.; Fan, D.; Shuai, B.; Santos-Villalobos, H.; Bhat, V.; MV, R. MEGA: Multimodal Alignment Aggregation and Distillation For Cinematic Video Segmentation. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Paris, France, 1–6 October 2023; pp. 23331–23340.
32. Eom, C.; Lee, G.; Lee, J.; Ham, B. Video-based person re-identification with spatial and temporal memory networks. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 11–17 October 2021; pp. 12036–12045.
33. Du, D.; Qi, Y.; Yu, H.; Yang, Y.; Duan, K.; Li, G.; Zhang, W.; Huang, Q.; Tian, Q. The unmanned aerial vehicle benchmark: Object detection and tracking. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 370–386.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.