ARTICLE

AdaptiveTrack: An Environment-Aware and Confidence-Refined Framework for Online Multi-Object Tracking

Said Baz Jahfar Khan[1], Peng Zhang[2,3,*], Mian Muhammad Kamal[4,*], Husam S. Samkari[5,6], Mohammed F. Allehyani[5] and Hala Mostafa[7]

[1]School of Software Engineering, Northwestern Polytechnical University, Xi'an, Shaanxi 710072, China
[2]School of Computer Science, Northwestern Polytechnical University, Xi'an, Shaanxi 710072, China
[3]Ningbo Institute, Northwestern Polytechnical University, Beilun, Ningbo 315048, China
[4]School of Electronic and Communication Engineering, Quanzhou University of Information Engineering, Quanzhou 362000, China
[5]Department of Electrical Engineering, University of Tabuk, Tabuk 47713, Saudi Arabia
[6]Artificial Intelligence and Sensing Technologies Research Center, University of Tabuk, Tabuk 47713, Saudi Arabia
[7]Department of Information Technology, College of Computer and Information Sciences, Princess Nourah bint Abdulrahman University, P.O. Box 84428, Riyadh 11671, Saudi Arabia
*Corresponding Author: Peng Zhang. Email: zh0036ng@nwpu.edu.cn;
Mian Muhammad Kamal. Email: mianmuhammadkamal@qzuie.edu.cn

Received: 28 February 2026; Accepted: 22 May 2026

ABSTRACT: Multi-object tracking (MOT) remains challenging in conditions involving occlusion, small objects, rapid motion, and crowding, wherein the accuracy of detection and the quality of association degrade simultaneously. We propose AdaptiveTrack, an online MOT framework featuring a closed-loop, confidence-aware association and recovery design: CSI-IoU adapts spatial overlap based on confidence and scale, EAMO refines similarity through density, velocity, and scale cues, and DCR updates detection confidence utilizing association context before NMS and assignment. A lightweight continuity module additionally preserves identities during missed detections. On MOT17/MOT20, AdaptiveTrack achieves HOTA 67.33 and 66.73, MOTA 82.55 and 78.30, and IDF1 83.20 and 82.57, operating at 23.5 FPS.

KEYWORDS: Multi-object tracking; YOLOX; association; occlusion; trajectory; crowded conditions

1 Introduction

Multi object tracking (MOT) is a vital task in computer vision consisting of detecting objects in video sequences and associating them across frames to maintain consistent trajectories with unique identities [1]. Maintaining temporally consistent identities is essential in applications such as autonomous navigation, surveillance, wildlife monitoring, human-robot interaction, and sports analysis, where future decisions rely on dependable trajectory continuity. Recent application-specific MOT studies also include autonomous pipeline deposit tracking [2]. Efficient multi-object tracking requires accurate object detection for target localization and efficient association to link detections into coherent tracklets, notwithstanding challenges such as occlusion, small object sizes, rapid motion, backdrop clutter, and changing illumination [3,4]. These features frequently result in missed detections, identity switches, and trajectory fragmentation, compromising tracking accuracy in complex environments. MOT methods are broadly categorized into

offline and online approaches. Offline methods process entire video sequences to improve tracking accuracy but often come with high computational costs [1]. In contrast, online methods operate on frames sequentially, enabling real-time tracking suitable for time-sensitive applications [5].

Despite steady progress, online MOT remains fragile when detection uncertainty and association ambiguity occur at the same time. Small objects frequently result in low-confidence detections due to insufficient pixel coverage, whereas occlusions cause missed detections, hence interrupting tracking continuity [6]. Fast motion and complex backgrounds further exacerbate identity switches, as traditional similarity metrics fail to adapt to dynamic scene variations [4]. The significant reliance on detection quality results in decreased tracker effectiveness when detector failures occur, especially for small or occluded objects.Recent work has also explored feature backtracking to improve robustness under severe occlusion [7]. Recent methods improve either association robustness or trajectory recovery, for example, through Siamese similarity learning [8] or Gaussian-smoothed interpolation [3]. But they normally address these stages separately and therefore remain vulnerable when small-object uncertainty, heavy occlusion, and real-time constraints coincide.

To address these limitations, we propose AdaptiveTrack, an online multi-object tracking framework designed to combine association, confidence refinement, and recovery inside a unified feedback-driven pipeline. As illustrated in Fig. 1, the method is designed to revise ambiguous detections utilizing association evidence before assignment and to maintain track continuity throughout times of temporary detection unreliability.

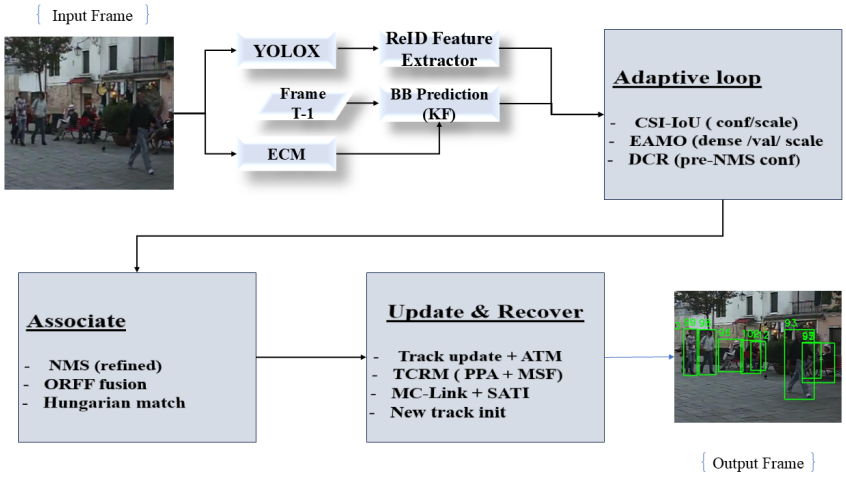

Figure 1: Overall architecture of the proposed AdaptiveTrack framework for multi-object tracking.

- CSI-IoU: Confidence- and size-aware IoU for more reliable detection–tracklet matching, especially for small/ambiguous targets.
- EAMO: Density-, velocity-, and scale-adaptive similarity refinement to reduce identity switches in complex scenes.
- DCR: Dynamic confidence refinement using spatial, motion, and temporal cues to stabilize tracking under occlusion.
- TCRM: Track recovery via velocity-scaled proposals and precise assignment to maintain trajectory continuity.

Unlike previous MOT methods that mainly improve overlap matching, multi-stage association, or recovery separately [4,6,9], AdaptiveTrack integrates these processes through a closed-loop framework. CSI-IoU

changes spatial overlap based on confidence and scale, EAMO adjusts matching scores using density, velocity, and size metrics, while DCR improves detection confidence before the NMS and assignment by utilizing the existing association context. In comparison to StrongSORT [3], BoostTrack++ [10], and BoT-SORT [11], the main difference lies not in the reliance on a singular cue, but in the explicit link between confidence estimation and association.

2 Related Work

Recent advancements in online MOT have mostly come from three areas: stronger detectors, accurate association functions, and post-hoc recovery of missed targets [3,11,12]. These directions address related but not identical failure modes, and their assumptions often break simultaneously in crowded environments, under heavy occlusion, or for small objects. Consequently, we analyze previous research to identify the limitations of these assumptions.

2.1 Detection Foundations

Object detection is fundamental to MOT, providing the bounding box predictions needed to initiate and maintain tracklets [3]. However, challenges such as small object size, occlusion, and motion blur often result in missed or low-confidence detections, thereby compromising tracking reliability [12]. To address these issues, recent MOT frameworks have integrated advanced detectors optimized for both accuracy and speed.

More recent frameworks such as StrongSORT [3], BoostTrack [12], and BoostTrack++ [10] employ YOLOX [13], valued for its real-time performance and improved handling of small objects. However, even strong real-time detectors remain unreliable for small or occluded targets, so downstream tracking stages often inherit low-confidence or missing observations instead of correcting them.

2.2 Data Association Strategies

Conventional methods use Kalman filters for motion prediction and Intersection over Union (IoU) for associating detections with tracklets . StrongSORT [3] integrates motion-based similarity produced by Kalman filters with appearance-based similarity obtained from Re-ID embeddings, hence improving matching accuracy in dynamic environments. Expansion IoU (EIoU) [9] broadens the matching area to reduce mismatches resulting from erratic motion, while Cascaded Buffered IoU (C-BIoU) [4] utilizes adaptive buffers in a cascaded matching framework to improve accuracy for quickly moving or occluded objects. BoostTrack and BoostTrack++ [10,12] use a one-stage association method with confidence enhancement, incorporating IoU and Mahalanobis distance to improve matching and reduce identity switches. BoostTrack++ now includes soft buffered IoU, improving accuracy in crowded environments. Despite these advancements, many association schemes still rely on fixed cue combinations whose behavior does not change with scene density, motion variation, or object size. This limitation drives our implementation of adaptive similarity refinement and confidence-scaled overlap, which are designed to react distinctly when ambiguity results from crowding, rapid motion, or small-object uncertainty.

2.3 Track Recovery Mechanisms

Missed detections caused by occlusion, small object sizes, or motion blur disrupt tracklet continuity, leading to fragmented trajectories in MOT. SMILEtrack [8] combines a robust detector with a Siamese network–based Similarity Learning Module to improve cross-frame object association. AFLink [3] uses simple motion-based predictions to link tracklets, providing a lightweight and efficient solution, though it

may lack robustness in complex scenarios with significant occlusions. Gaussian-Smoothed Interpolation (GSI) [3] applies Gaussian process regression to model non-linear motion, enabling accurate recovery of missing objects at minimal computational cost. These methods enhance recovery, but most of them rely primarily on either motion continuity or appearance consistency. This approach becomes limiting when detections are insufficient, short gaps have to be associated, and the confidence in associations is ambiguous. This conclusion motivates the combination of track maintenance, proposal-level recovery, interpolation, and confidence refinement in a common online framework.

3 Proposed Model

Figure 1 depicts the overall framework of AdaptiveTrack, and Algorithm 1 elaborates its procedural implementation, a novel MOT framework that incorporates adaptive feature modulation and robust association techniques. AdaptiveTrack utilizes a real-time detector (YOLOX [13]) based on the pretrained model from ByteTrack [14]

3.1 Confidence-Scaled IoU Metric (CSI-IoU)

Standard IoU metric becomes unreliable for small, partially occluded, or slightly misaligned objects, as even valid matches may demonstrate minimal pixel overlap. To reduce this sensitivity, we propose CSI-IoU, which adjusts the detection and tracking boxes based on confidence and size before overlap is evaluated:

$$\text{CSI-IoU}(B'_n, K''_m) = \text{IoU}\Big(B'_n \to \sigma_{b'_n},\ K''_m \to \sigma_{k''_m}\Big), \tag{1}$$

Here, $\phi_{b'_n}$ and $\phi_{k''_m}$ are confidence scores from the YOLOX detector and tracking module, $\bar{A}_{B'_n}$ and $\bar{A}_{K''_m}$ are the areas of detection and tracklet boxes, and $\bar{A}_{\max}$ is the maximum area in the frame.

with scaling factors:

$$\sigma_{b'_n} = \text{clip}\left(\sigma_{\min}, \sigma_{\max}, \frac{1 - \phi_{b'_n} \exp\left(\frac{\bar{A}_{B'_n}}{\bar{A}_{\max}}\right)}{2.5}\right), \tag{2}$$

$$\sigma_{k''_m} = \text{clip}\left(\sigma_{\min}, \sigma_{\max}, \frac{1 - \phi_{k''_m} \exp\left(\frac{\bar{A}_{K''_m}}{\bar{A}_{\max}}\right)}{2.5}\right). \tag{3}$$

where $\sigma_{\min}$ and $\sigma_{\max}$ are the minimum and maximum bounds for the scaling factor, ensuring that the scaling remains within a reasonable range. $\phi_{b'_n}$ and $\phi_{k''_m}$ are the confidence scores from the YOLOX detector and tracking module, respectively. $\bar{A}_{B'_n}$ and $\bar{A}_{K''_m}$ represent the areas of the detection and tracklet bounding boxes, and $\bar{A}_{\max}$ denotes the maximum area within the frame.

Additionally, we compute the center-distance $ED(B'_n, K''_m)$ and apply CSI-IoU only when the distance is small relative to the track scale:

$$\frac{ED(B'_n, K''_m)}{\sqrt{(K''^w_m)^2 + (K''^h_m)^2}} < \epsilon. \tag{4}$$

We set ϵ via grid search on MOT17 validation and keep it fixed across datasets.

3.2 Environment-Adaptive Matching Optimization (EAMO)

After CSI-IoU provides a scale- and confidence-aware overlap measure, the remaining question is how much that match should be trusted under different scene conditions. EAMO addresses this issue by changing the basic score through three related factors: local congestion, motion-dependent distance reliability, and scale-sensitive form consistency.

3.2.1 Scene-Density IoU Boost (SDIB)

In crowded environments, the IoU metric loses its discriminative power because overlap may reflect nearby distractors instead of true correspondence. SDIB decreases this effect by adjusting the similarity score according to local scene density.

$$S^{\mathrm{EAMO}} = S_{\mathrm{init}} + \Delta S^{\mathrm{SDIB}} + \Delta S^{\mathrm{VMDC}} + \Delta S^{\mathrm{SDSC}}, \tag{5}$$

$$\Delta S^{\mathrm{SDIB}} = \eta_{\rho} \cdot \frac{\rho_{B'_n}}{\max(\mathrm{IoU}_{\mathrm{scene}}) + \epsilon_0}, \tag{6}$$

where S_{init} is the initial similarity (IoU or distance), $\rho_{B'_n} = \max_k \mathrm{IoU}(B'_n, B'_k)$ measures local crowding/occlusion, $\max(\mathrm{IoU}_{\mathrm{scene}})$ is the maximum overlap in the frame used for normalization, and $\epsilon_0 = 10^{-6}$.

3.2.2 Velocity-Modulated Distance Correction (VMDC)

Mahalanobis distance is useful for motion-based gating; however, its reliability inevitably fluctuates between tracks with differing velocities. VMDC, therefore, adjusts the distance parameter according to tracklet velocity, verifying that slow-moving objects are not penalized aggressively as fast-moving ones.

$$\Delta S^{\mathrm{VMDC}} = \eta_{\mathrm{VM}} \cdot \frac{\mathrm{MD}(B'_n, K''_m)}{1 + \exp\left(V_{K''_m} / V_{\mathrm{avg}}\right)}. \tag{7}$$

where $\mathrm{MD}(B'_n, K''_m)$ is capped at 10.5, $V_{K''_m}$ is the tracklet velocity from the Kalman filter, V_{avg} is the frame's average velocity, and $\eta_{\mathrm{VM}} = 0.02$. VMDC prioritizes slow-moving objects.

3.2.3 Size-Driven Shape Calibration (SDSC)

Shape similarity is valuable for tracking; however, unadjusted measures disregard scale variations. The SDSC includes object size in the evaluation of form similarity, enhancing associations—particularly for small objects—by adjusting shape features according to scale.

$$\delta_{n,m} = \frac{|B'^{w}_n - K''^{w}_m|}{\max(B'^{w}_n, K''^{w}_m)} + \frac{|B'^{h}_n - K''^{h}_m|}{\max(B'^{h}_n, K''^{h}_m)}, \tag{8}$$

$$S_{\mathrm{shape}}(B'_n, K''_m) = \exp(-1.5 \cdot \delta_{n,m}) \cdot \frac{\sqrt{\bar{A}_{K''_m}}}{\bar{A}_{\max} + 1}, \tag{9}$$

$$\Delta S^{\mathrm{SDSC}} = \eta_{\mathrm{SD}} \cdot S_{\mathrm{shape}}(B'_n, K''_m). \tag{10}$$

where $\eta_{\mathrm{SD}} = 0.1$, and $\bar{A}_{\max}$ is the maximum tracklet area. SDSC enhances small object associations.

Parameter Selection and Sensitivity

The hyperparameters introduced in CSI-IoU and EAMO (e.g., η_{VM}, η_{SD}, similarity thresholds) were established empirically with a grid search on the MOT17 validation set. We found stable performance across a broad range (η_{VM} : 0.01–0.05, η_{SD} : 0.05–0.15, thresholds ±0.05), indicating that AdaptiveTrack is not overly sensitive to specific values. To ensure consistency across datasets, we fixed them to the reported configurations.

3.3 Dynamic Confidence Refinement (DCR)

DCR is implemented before NMS and the assignment process to adjust detector scores according to the existing association evidence. Rather than treating detection confidence as static, it enhances the scores of candidates that retain spatial and temporal consistency with predicted tracks. We implement this refinement using CSI-IoU, EAMO, and a Scene-Aware Temporal Factor (SATF), with $\alpha = 0.5$ obtained through grid search on the MOT17 validation set for all experiments.

$$F_{B'_n}^{\text{SATF}} = \frac{T_i}{T_i + 5} \cdot \left(1 - \nu \cdot \rho_{B'_n}\right), \tag{11}$$

$$\phi^*_{b'_n} = \phi_{b'_n} + \alpha \cdot \text{CSI-IoU}(B'_n, K''_m) \cdot \left(S^{\text{EAMO}}(B'_n, K''_m) + F_{B'_n}^{\text{SATF}}\right). \tag{12}$$

3.3.1 Confidence Boost for Matched Detections (CBMD)

CBMD strengthens the accuracy of detections matched with existing tracklets by integrating CSI-IoU, SDSC, and SATF. This adaptive refinement maintains detection accuracy in challenging environments, including occlusion, crowding, or changes in appearance. We reuse the Scene-Aware Temporal Factor $F_{B'_n}^{\text{SATF}}$, where $\rho_{B'_n} = \frac{N_{\text{close}}}{N_{\text{max}}}$, N_{close} is the number of detections within 50 pixels, N_{max} is the 95th percentile of N_{close}, and $\nu = 0.12 \cdot \rho^*$ with ρ^* as the mean density. The boosted confidence is:

$$\phi^*_{b'_n} = \phi_{b'_n} + \alpha \cdot \text{CSI-IoU}(B'_n, K''_m) \cdot \left(S_{\text{shape}}(B'_n, K''_m) + F_{B'_n}^{\text{SATF}}\right). \tag{13}$$

3.3.2 Confidence Recovery for Unmatched Detections (CRUD)

CRUD restores confidence for detections not linked to any tracklet by leveraging motion consistency (via Mahalanobis distance) and SATF. This approach salvages valid but temporarily missed targets, reducing false negatives in dynamic or partially occluded scenes.

$$M_{B'_n} = \frac{\text{MD}(B'_n, K''_m)}{1 + \exp(V_{K''_m}/V_{\text{avg}})}, \tag{14}$$

$$\phi^*_{b'_n} = \phi_{b'_n} + \alpha \cdot \text{CSI-IoU}(B'_n, K'''_m) \cdot \left(1 - \frac{M_{B'_n}}{\sigma_{\text{MD}}} + F_{B'_n}^{\text{SATF}}\right), \tag{15}$$

where K'''_m has the highest CSI-IoU (>0.3), σ_{MD} is the standard deviation of Mahalanobis distances, and NMS with a 0.55 threshold removes redundancies.

3.4 Track Continuity and Recovery Module (TCRM)

TCRM mitigates trajectory breaks caused by occlusions and missed detections. It integrates Adaptive Track Maintenance (ATM) for persistence handling, Precise Proposal Assignment (PPA) for accurate matching, and Motion-Shape Fusion (MSF) for robust association in complex scenes.

3.4.1 Adaptive Track Maintenance (ATM)

ATM preserves tracks that temporarily disappear due to occlusion or missed detections by integrating detection confidence with a motion-based decay function. This ensures the maintenance of reliable tracks for a sufficient duration to allow recovery while reducing excessive false positives.

$$R_n = \phi_{K_n''} \cdot \exp(-\Delta t \cdot \sqrt{V_{K_n''}/V_{\text{avg}}}), \tag{16}$$

where $\phi_{K_n''}$ is the last detection confidence, Δt is the number of lost frames, and tracks persist if $R_n > 0.5$.

3.4.2 Precise Proposal Assignment (PPA)

PPA matches detections within a motion-scaled area, enhancing association accuracy. The detection exhibiting the highest CSI-IoU:

$$Z_n = [x_n \pm \theta \cdot \sqrt{V_{K_n''}}, y_n \pm \theta \cdot \sqrt{V_{K_n''}}], \tag{17}$$

where $\theta = 2.2$. PPA selects the detection with the highest CSI-IoU (>0.4).

Algorithm 1: AdaptiveTrack (very concise)

Require: Frame I_t, detector $\mathcal{Y}$, tracks T_{t-1}, ReID $\mathcal{E}$
Ensure: Updated tracks T_t

$D_t \leftarrow \mathcal{Y}(I_t)$; $F_t^{app} \leftarrow \mathcal{E}(I_t, D_t)$; $\hat{T}_t \leftarrow \text{PredictKalman}(T_{t-1})$
$\rho, V_{\text{avg}}, \bar{A}_{\max}, F_t^{CFAM} \leftarrow \text{SceneStats\&CFAM}(I_t, D_t, \beta{=}0.6)$
$S^{CSI} \leftarrow \text{CSI-IoU}(D_t, \hat{T}_t, ED/\sqrt{(K^w)^2+(K^h)^2}{<}\epsilon)$; $S^{EAMO} \leftarrow \text{EAMO}(S^{CSI}, D_t, \hat{T}_t)$
$\Phi^{DCR} \leftarrow \text{DCR}(D_t, \hat{T}_t, S^{EAMO}, S^{CSI})$; $D_t' \leftarrow \text{ApplyNMS}(D_t, \Phi^{DCR}, 0.55)$
$S^{ORFF} \leftarrow \text{ORFF}(S^{CSI}, F_t^{SATF}, \rho, \gamma{=}0.15)$; $(M, U_T, U_D) \leftarrow \text{Hungarian}(\text{Combine}(S^{EAMO}, S^{ORFF}), 0.4)$
$\tilde{T}_t \leftarrow \text{UpdateTracks}(\hat{T}_t, D_t', F_t^{app}, M)$; $T^{lost} \leftarrow \text{ATM}(U_T, 0.5)$
$T^{rec} \leftarrow \text{SATI}(\text{MC-Link}(\text{TCRM}(T^{lost}, D_t', S^{CSI}), T^{lost}, D_t', F_t^{app}), \tau{=}15, W_s{=}1-\rho)$
$T^{new} \leftarrow \text{InitTracks}(D_t', F_t^{app}, U_D \setminus T^{rec})$; $T_t \leftarrow \text{Merge}(\tilde{T}_t \setminus U_T, T^{rec}, T^{new})$
return T_t

Here K^w and K^h denote the width and height of the predicted track box from the Kalman filter.

3.4.3 Motion-Shape Fusion (MSF)

MSF strengthens similarity scoring by integrating CSI-IoU with motion scaling and short-term interpolation. This integration improves alignment during fast or irregular object motions.

$$S_{\text{MSF}}(B_n', K_m'') = \text{CSI-IoU}(B_n', K_m'') \cdot \left(1 + \exp(V_{K_m''}/V_{\text{avg}} - 1)\right), \tag{18}$$

combined with 3-frame interpolation weighted by R_n.

3.5 Motion-Context Linker (MC-Link)

MC-Link improves tracklet association through motion-driven spatiotemporal analysis. It adjusts a temporal window (10–50 frames) based on object speed and density, using a lightweight CNN to extract motion features. A multilayer perceptron with focal loss evaluates association confidence, while spatial constraints combined with linear assignment optimize global linking.

3.6 Scale-Aware Trajectory Interpolator (SATI)

Trajectory gaps caused by missed detections lead to fragmented tracks that obstruct analysis. SATI uses Gaussian Process Regression with scale-adaptive smoothness, dynamically adjusting interpolation depending on trajectory length and spatial extent for robust gap filling.

$$\mu = \frac{\tau^2}{l_B} \cdot \sqrt{\frac{\bar{A}_{K_m''}}{\bar{A}_{\max} + 1}}, \tag{19}$$

where $\tau = 15$, l_B is trajectory length, and $\bar{A}_{K_m''}$ is the tracklet area. Noise is weighted by $W_s = 1 - \rho_{B_n'}$, and smoothed positions are:

$$P^* = K(T^*, T) \cdot \left[K(T, T) + \sigma^2 \cdot W_s \cdot I\right]^{-1} P, \tag{20}$$

where $K(\cdot, \cdot)$ is the Gaussian kernel.

3.7 Compact Feature Analysis Module (CFAM)

The CFAM focuses on extracting discriminative texture and edge information from small objects using a shallow CNN architecture. CFAM improves tracking performance by efficiently integrating edge and texture cues for stronger small-object representation.

$$F_{B_n'}^{\text{CFAM}} = \frac{1}{1 + \exp(-\beta \cdot (E_{\text{edge}} + E_{\text{texture}}))}, \tag{21}$$

where E_{edge} and E_{texture} are feature strengths, and $\beta = 0.6$. CFAM enhances CSI-IoU and EAMO for small object tracking.

3.8 Occlusion-Resilient Feature Fusion (ORFF)

ORFF integrates overlap and scene-aware temporal cues to maintain reliable tracking under occlusion. ORFF improves association robustness in challenging situations by prioritizing unoccluded objects with prominent features.

$$S_{\text{ORFF}}(B_n', K_m'') = \text{CSI-IoU}(B_n', K_m'') \cdot \left(1 + \gamma \cdot F_{B_n'}^{\text{SATF}} \cdot \exp(-\rho_{B_n'})\right), \tag{22}$$

where $\gamma = 0.15$. ORFF prioritizes unoccluded objects with strong features, improving association robustness.

Justification for CFAM and ORFF

CFAM and ORFF were implemented to address two specific deficiencies in the basic pipeline. CFAM increases local edge and texture cues when objects occupy limited pixels and visual evidence is insufficient. ORFF improves association stability when overlap is ambiguous because of the partial occlusion effect through combining spatial consistency with scene-aware temporal data. Despite their lightweight design,

these modules have a significant function; in the ablation study, they provide supplementary benefits during the establishment of the core association and recovery phases.

4 Experimental

4.1 Datasets and Metrics

We evaluate AdaptiveTrack using the MOT17 [15] and MOT20 [16] benchmarks. MOT17 contains pedestrian sequences captured by both stationary and moving cameras, including 7 training sequences totaling 5316 frames at 14–30 FPS, and 5 test sequences containing 5919 frames. MOT20 contains eight crowded sequences captured under various lighting conditions at 25 FPS (four training sequences totaling 8931 frames; four test sequences totaling 4479 frames).

We report standard MOT metrics: MOTA [17] (errors from false positives, false negatives, and identity switches), IDF1 [18] (identity consistency), and HOTA [19], which together evaluate detection (DetA) and association (AssA) over detection similarity thresholds from 0.05 to 0.95.

4.2 Implementation Details

All experiments were implemented in PyTorch on a workstation with dual NVIDIA RTX 2080 Ti GPUs. We use YOLOX [13] for detection. The ReID model is ResNet-18, fine-tuned for 50 epochs (batch size 32, learning rate 0.0003); total training time is ~48 h.

CFAM is a lightweight three-layer CNN (16/32/64 filters, 3×3 kernels) combining BN and ReLU activation, together with a 1×1 layer for edge and texture cues. MC-Link uses a three-layer MLP (256/128/64) with ReLU activation and a sigmoid output, which optimizes focal loss. Inputs are scaled to 640×640 and normalized to the range $[0, 1]$. AdaptiveTrack runs at 23.5 FPS on MOT17 and 10.6 FPS on MOT20. AdaptiveTrack operates at 23.5 FPS on MOT17 and 10.6 FPS on MOT20 on our dual RTX 2080 Ti workstation; runtime may vary on newer or entry-level hardware.

Computational cost and scalability. Per-module costs are as follows: CSI-IoU/EAMO/DCR compute pairwise detection-track scores in $O(NM)$ time and stores a similarity matrix of size $O(NM)$, where N indicates the number of detections and M represents the number of active tracks; the Hungarian assignment mainly impacts association with a complexity of $O(K^3)$, where $K = \min(N, M)$; recovery modules are bounded by gating/windowing. TCRM (ATM/PPA/MSF) is lightweight with gating, MC-Link is constrained to a window of 10 to 50 frames, and SATI is used only for short gaps. Runtime primarily increases with scene density as N and M increase in crowded sequences (e.g., MOT20).

4.3 Ablation Study

An ablation study was conducted on the MOT17 and MOT20 validation sets to evaluate the contribution of each component in our proposed AdaptiveTrack framework, using the official MOTChallenge evaluation code for accuracy [15,16]. The results from various configurations are presented in Fig. 2 for MOT17 individual results, Table 1 for MOT17 cumulative results, Fig. 3 for MOT20 individual results, and Table 2 for MOT20 cumulative results. The evaluation metrics used include HOTA (Higher-Order Tracking Accuracy), MOTA , IDF1 , and FPS . We evaluate AdaptiveTrack progressively from a baseline YOLOX-based tracking pipeline without CSI-IoU, EAMO, DCR, TCRM, MC-Link, SATI, CFAM, or ORFF. The ablation results indicate a consistent cumulative trend in which the proposed components improve tracking performance on the evaluated split. The cumulative ablation results represent a singular evaluation based on our internal division and should be interpreted as trend evidence for component contribution rather than as a statistical

measure of run-to-run variability. Tables 1 and 2 use an internal 50%/50% split of the official training sets for ablation only and are not directly comparable to Table 3.

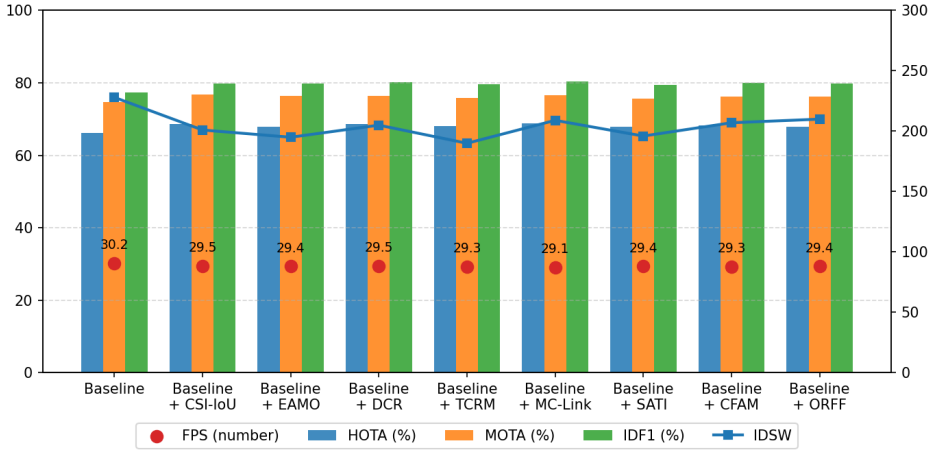

Figure 2: Ablation study: individual component performance on the MOT17 validation set.

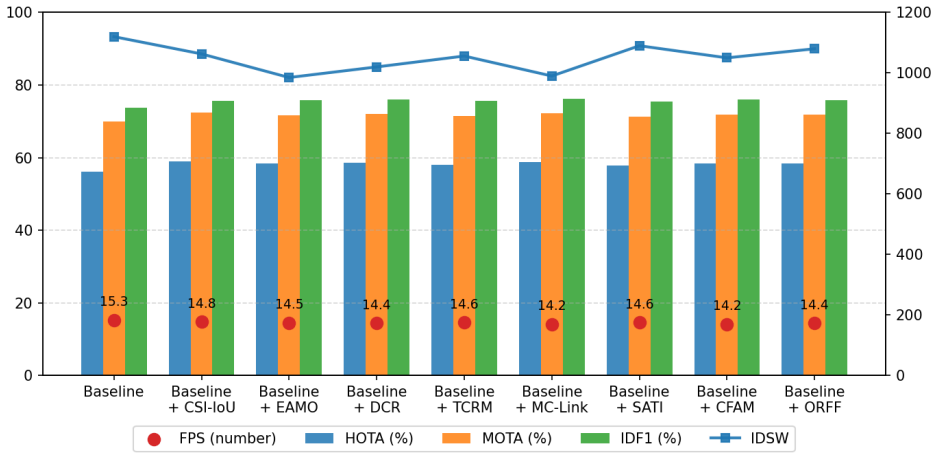

Figure 3: Ablation study: individual component performance on the MOT20 validation set.

Table 1: Cumulative component ablation on MOT17 (Internal 50%/50% split of the official training set).

Configuration	HOTA (%)	MOTA (%)	IDF1 (%)	IDSW
Baseline	66.13	74.80	77.30	227
Baseline + CSI-IoU	68.73	76.70	79.8	201
Baseline + CSI-IoU + EAMO	69.87	77.95	80.00	171
Baseline + CSI-IoU + EAMO + DCR	71.41	79.10	82.50	163
Baseline + CSI-IoU + EAMO + DCR + TCRM	72.35	79.75	83.50	155
Baseline + CSI-IoU + EAMO + DCR + TCRM + MC-Link	74.09	81.10	85.20	129
Baseline + CSI-IoU + EAMO + DCR + TCRM + MC-Link + SATI	74.73	81.55	86.00	110
Baseline + CSI-IoU + EAMO + DCR + TCRM + MC-Link + SATI + CFAM	76.07	82.60	87.30	99
Baseline + CSI-IoU + EAMO + DCR + TCRM + MC-Link + SATI + CFAM + ORFF	77.31	83.55	88.40	82

Table 2: Cumulative component ablation on MOT20 (internal 50%/50% split of the official training set).

Configuration	HOTA (%)	MOTA (%)	IDF1 (%)	IDSW
Baseline	56.17	62.92	73.72	1120
Baseline + CSI-IoU	58.97	72.42	75.62	1063
Baseline + CSI-IoU + EAMO	59.56	73.41	76.79	1025
Baseline + CSI-IoU + EAMO + DCR	61.35	74.80	78.16	945
Baseline + CSI-IoU + EAMO + DCR + TCRM	62.64	75.59	79.03	891
Baseline + CSI-IoU + EAMO + DCR + TCRM + MC-Link	64.63	77.08	80.60	840
Baseline + CSI-IoU + EAMO + DCR + TCRM + MC-Link + SATI	65.72	77.67	81.27	783
Baseline + CSI-IoU + EAMO + DCR + TCRM + MC-Link + SATI + CFAM	67.41	78.86	82.54	731
Baseline + CSI-IoU + EAMO + DCR + TCRM + MC-Link + SATI + CFAM + ORFF	69.10	79.75	83.61	695

Table 3: Comparison of state-of-the-art methods on MOT17 and MOT20 benchmarks.

Method	MOT17				MOT20			
	HOTA	MOTA	IDF1	IDSW	HOTA	MOTA	IDF1	IDSW
StrongSORT++ [3]	64.4	79.6	79.5	1194	62.6	73.8	77.0	770
BoT-SORT [11]	65.0	80.5	80.2	1212	63.3	77.8	77.4	1313
SolidTrack [20]	65.0	80.6	80.0	1306	63.4	77.9	77.5	1228
MotionTrack [21]	65.1	81.1	80.1	1140	62.8	78.0	76.5	1165
SparseTrack [22]	65.1	81.0	80.1	-	63.4	78.2	77.3	-
LG-Track [23]	65.4	81.4	80.4	1125	63.4	77.8	77.4	1161
ConfTrack [24]	65.4	80.0	81.2	1155	64.8	77.2	80.2	702
StrongTBD [25]	65.6	81.6	80.8	954	64.6	78.0	77.0	1101
ImprAsso [26]	66.4	82.2	82.1	924	64.6	78.6	78.8	992
BoostTrack [12]	66.4	80.6	81.8	1086	66.2	77.2	81.5	827
BoostTrack++ [10]	66.6	80.7	82.2	1062	66.4	77.7	82.0	762
FocusTrack [27]	66.91	82.32	82.96	1056	66.5	77.9	82.1	760
HAMOT [5]	67.2	82.5	83.1	1044	66.7	78.2	82.5	746
AdaptiveTrack (Ours)	**67.33**	**82.55**	**83.2**	**1040**	**66.73**	**78.30**	**82.57**	**745**

4.3.1 Individual Component Analysis

Figs. 2 and 3 show the baseline performance and the impact of progressively integrating each AdaptiveTrack component on MOT17 and MOT20. CSI-IoU improves detection-tracklet association by adjusting bounding boxes according to detection confidence and the object's size relative to the largest object in the frame, whereas EAMO boosts the similarity matrix through Scene-Density IoU Boost, Velocity-Modulated Distance Correction, and Size-Driven Shape Calibration. DCR improves confidence estimation using spatial, motion, and temporal cues, while TCRM enhances track persistence and occlusion recovery through Adaptive Track Maintenance, Accurate Proposal Assignment, and Motion-Shape Fusion. MC-Link optimizes association stability with complicated motion, SATI addresses trajectory gaps from missed detections, CFAM adjusts small-object cues utilizing lightweight edge and texture features, and ORFF combines appearance and motion to ensure strong tracking in crowded environments.

4.3.2 Cumulative Component Analysis

Tables 1 and 2 show that the proposed modules are complementary rather than redundant. The initial gains are primarily associated with enhanced local matching and score verification, whereas the later gains come from recovering missed associations and smoothing short gaps. The strong cumulative effect

on MOT20 further illustrates that the complete pipeline is most advantageous when crowd density and occlusion make fixed association cues less reliable.

4.4 Comparison with Other Methods

We conducted a comparison analysis of AdaptiveTrack versus popular online multi-object tracking (MOT) frameworks using the MOT17 and MOT20 benchmarks under the private detection procedure to evaluate its effectiveness. The corresponding quantitative results are reported in Table 3 and are further illustrated graphically in Fig. 4. Tables 1 and 2, by contrast, report ablation results on an internal 50%/50% split of the official training sets.

Fig. 5 presents qualitative comparisons: the first row depicts tracking performance in a low-density pedestrian environment with small targets; the second row illustrates results in a slightly crowded indoor shopping mall; and the final row displays performance in a heavily congested urban street identified by frequent occlusions. These scenarios jointly evaluate the adaptability of each method to various real-world environments.

Compared to current online MOT methods, AdaptiveTrack shows modest yet ongoing improvements in benchmark evaluation criteria. in MOT17, it achieves 67.33 HOTA, 82.55 MOTA, and 83.20 IDF1; in MOT20, it obtains 66.73 HOTA, 78.30 MOTA, and 82.57 IDF1. These improvements are not large in absolute terms, but they are consistent across detection-oriented and identity-oriented measures. This suggests that combining confidence refinement with association results in advantages for both track preservation and identity consistency, rather than only enhancing a single metric.

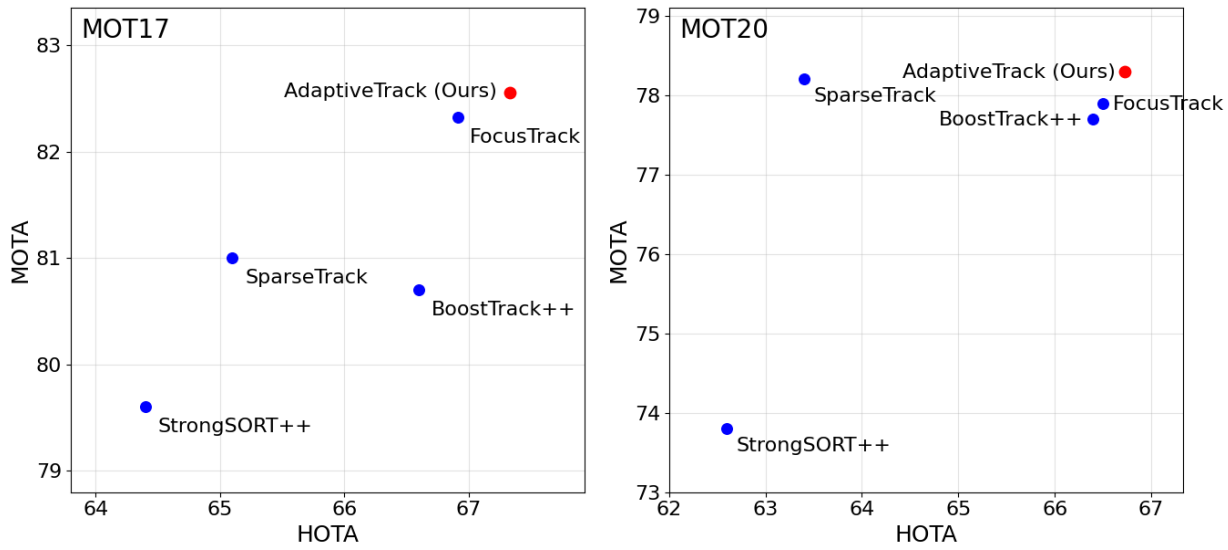

Figure 4: HOTA and MOTA metric results on MOT17 (left) and MOT20 (right) test sets.

Failure cases. AdaptiveTrack may still fail under prolonged full occlusion or when multiple objects show highly similar appearances and motions, causing identity switches. Fig. 6 illustrates a distant object that reappears near the camera after prolonged occlusion with a new ID.

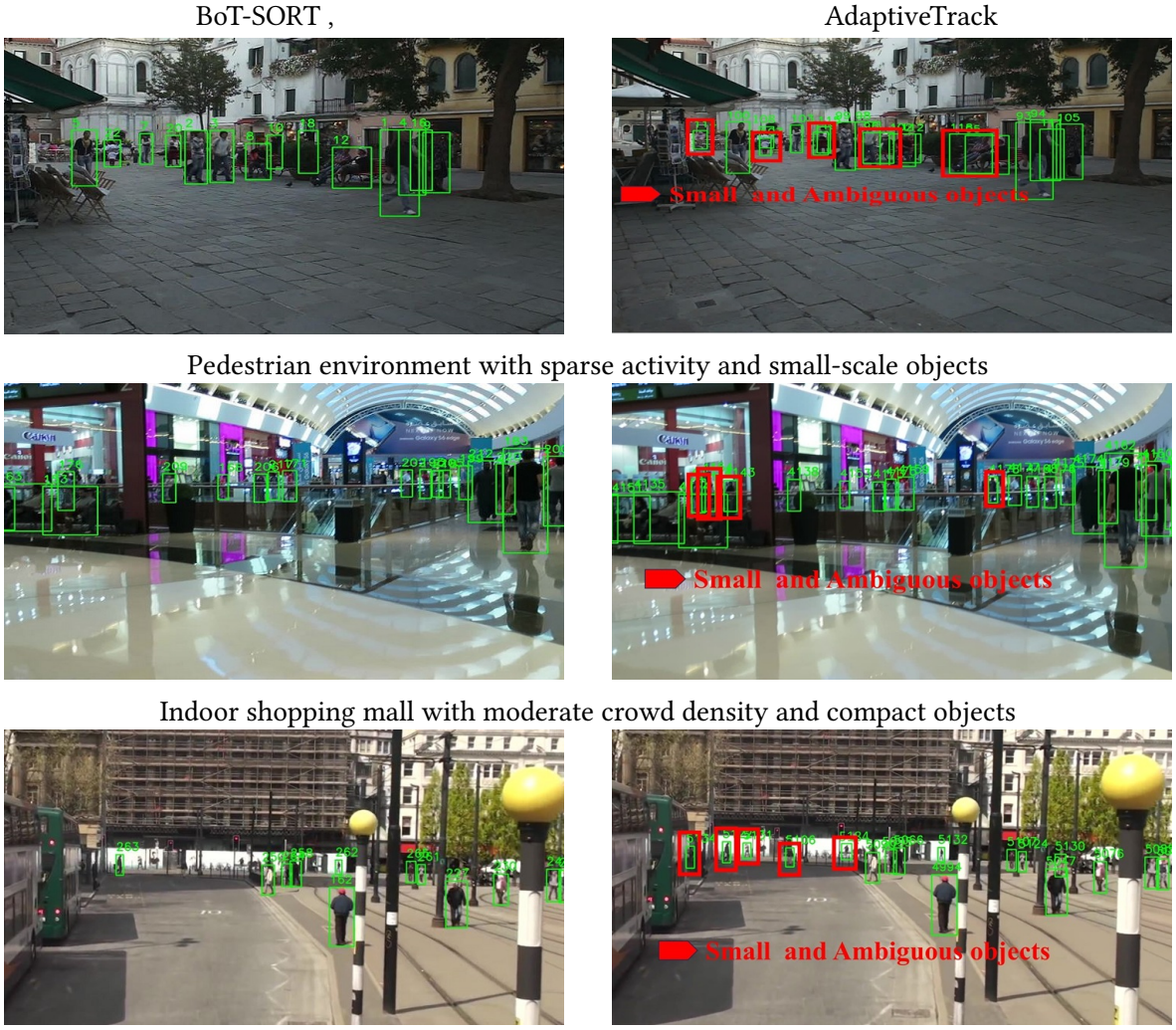

Figure 5: Comparative tracking results across different environments.

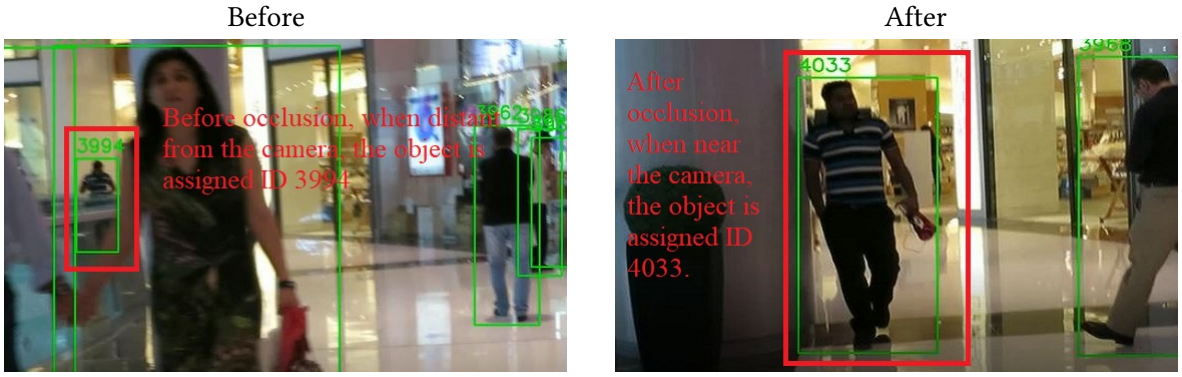

Figure 6: Failure case: the target is initially small and far from the camera, then after prolonged full occlusion reappears near the camera with a different assigned ID.

Sensitivity to detection errors. AdaptiveTrack, being a tracking-by-detection framework, is especially susceptible to persistent false negatives for small or occluded objects, in addition to significant detection

noise that disrupts spatial consistency across frames. DCR and TCRM mitigate short, intermittent misses by refining low-confidence detections and maintaining tracks across brief gaps; however, when the detector repeatedly suppresses a target or produces unstable boxes, recovery becomes unreliable and fragmentation increases.

Dense-crowd limitations. In extremely crowded crowds, significant overlap and merged detections make association inherently ambiguous and increase the number of feasible matches. EAMO and ORFF improve robustness in these situations; however, performance may still decrease when occlusions surpass the effective recovery window or when multiple detections converge into indistinguishable clusters. In addition, runtime escalates with the proliferation of detections and active tracks.

5 Conclusion

AdaptiveTrack is an online MOT framework focused on a closed-loop, confidence-aware association and recovery mechanism. CSI-IoU and EAMO modify spatial similarity according to confidence/scale and scene density/velocity, while DCR updates detection confidence through association context before assignment, differentiating it from conventional fixed-score association frameworks. This improves identity stability under occlusion and small-object ambiguity, including lightweight continuity and recovery, while preserving real-time performance on the tested platform. AdaptiveTrack achieves HOTA 67.33/66.73, MOTA 82.55/78.30, and IDF1 83.20/82.57 on the MOT17/MOT20 datasets at frame rates of 23.5/10.6 FPS. Aside from detection/ReID, the main burden is in pairwise scoring and Hungarian matching, which increases with the number of detections/tracks in crowded environments. Future work will focus on improving robustness under heavy occlusion and rapid camera motion with lower cost in dense scenes.

Acknowledgement: The authors would like to thank all individuals and institutions that contributed to this research.

Funding Statement: Princess Nourah bint Abdulrahman University Researchers Supporting Project number (PNURSP2026R137), Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia.

Author Contributions: Said Baz Jahfar Khan contributed to conceptualization, methodology, software, formal analysis, investigation, data curation, and writing–original draft and review/editing; Peng Zhang contributed to conceptualization, methodology, validation, resources, supervision, project administration, and writing–review/editing; Mian Muhammad Kamal contributed to conceptualization, methodology, formal analysis, investigation, data curation, visualization, and writing–original draft and review/editing; Husam S. Samkari contributed to software, validation, formal analysis, investigation, data curation, and writing–review/editing; Mohammed F. Allehyani contributed to investigation, resources, data curation, visualization, and writing–review/editing; and Hala Mostafa contributed to methodology, validation, formal analysis, writing–original draft and review/editing, project administration, and funding acquisition. All authors reviewed and approved the final version of the manuscript.

Availability of Data and Materials: Data available on request from the authors.

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest to report regarding the present study.

References

1. Khan BJS, Peng Z, Kamal M, Mohamed HG, Kharma QM, Sheraz M, et al. STRACK: robust tracking of small objects in low-light conditions. IEEE Access. 2025;13:151466–78.
2. Wong LNY, Yu S, Wong KYP. Intelligent and autonomous pipeline deposit tracking based on a multi-object tracking framework. Tunn Undergr Space Technol. 2026;171:107463.

3. Du Y, Zhao Z, Song Y, Zhao Y, Su F, Gong T, et al. StrongSORT: make DeepSORT great again. IEEE Trans Multimedia. 2023;25:8725–37.
4. Yang F, Odashima S, Masui S, Jiang S. Hard to track objects with irregular motions and similar appearances? make it easier by buffering the matching space. In: Proceedings of the 2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV); 2023 Jan 2–7; Waikoloa, HI, USA. p. 4799–808.
5. Baz JKS, Zhang P, Kamal MM, Mohamed HG, Sheraz M, Chuah TC. HAMOT: a hierarchical adaptive framework for robust multi-object tracking in complex environments. Comput Model Eng Sci. 2025;145(1):947–69.
6. Vaquero L, Xu Y, Alameda-Pineda X, Brea VM, Mucientes M. Lost and found: overcoming detector failures in online multi-object tracking. arXiv:2407.10151. 2024. [arXiv]
7. Ma J, Zheng Y, Lyu Y, Jia C, Su Y. A Multi-Object Tracking framework with feature backtracking under severe occlusion. Ocean Eng. 2026;351:124175.
8. Wang YH, Hsieh JW, Chen PY, Chang MC, So HH, Li X. SMILEtrack: SiMIlarity LEarning for occlusion-aware multiple object tracking. Proc AAAI Conf Artif Intell. 2024;38(6):5740–8.
9. Huang HW, Yang CY, Sun J, Kim PK, Kim KJ, Lee K, et al. Iterative scale-up ExpansionIoU and deep features association for multi-object tracking in sports. In: Proceedings of the 2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV); 2024 Jan 1–6; Waikoloa, HI, USA. p. 163–72.
10. Stanojević V, Todorović B. BoostTrack++: using tracklet information to detect more objects in multiple object tracking. arXiv:2408.13003. 2024. [arXiv]
11. Aharon N, Orfaig R, Bobrovsky BZ. BoT-SORT: robust associations multi-pedestrian tracking. arXiv:2206.14651. 2022. [arXiv]
12. Stanojević VD, Todorović BT. BoostTrack: boosting the similarity measure and detection confidence for improved multiple object tracking. Mach Vis Appl. 2024;35(3):53.
13. Ge Z. YOLOX: exceeding YOLO series in 2021. arXiv:2107.08430. 2021. [arXiv]
14. Zhang Y, Sun P, Jiang Y, Yu D, Weng F, Yuan Z, et al. ByteTrack: multi-object tracking by associating every detection box. In: Proceedings of the European Conference on Computer Vision (ECCV); 2022 Oct 23–27; Tel Aviv, Israel. p. 1–21.
15. Milan A. MOT16: a benchmark for multi-object tracking. arXiv:1603.00831. 2016. [arXiv]
16. Dendorfer P. MOT20: a benchmark for multi-object tracking in crowded scenes. arXiv:2003.09003. 2020. [arXiv]
17. Bernardin K, Stiefelhagen R. Evaluating multiple object tracking performance: the CLEAR MOT metrics. EURASIP J Image Video Process. 2008;2008(1):246309.
18. Ristani E, Solera F, Zou R, Cucchiara R, Tomasi C. Performance measures and a data set for multi-target, multi-camera tracking. In: Proceedings of the European Conference on Computer Vision (ECCV); 2016 Oct 11–14; Amsterdam, The Netherlands. p. 17–35.
19. Luiten J, Osep A, Dendorfer P, Torr P, Geiger A, Leal-Taixé L, et al. HOTA: a higher order metric for evaluating multi-object tracking. Int J Comput Vis. 2021;129(2):548–78.
20. Wu D, Huang Z, Zhang Y. SolidTrack: a novel method for robust and reliable multi-pedestrian tracking. Electronics. 2025;14(7):1370.
21. Qin Z, Zhou S, Wang L, Duan J, Hua G, Tang W. MotionTrack: learning robust short-term and long-term motions for multi-object tracking. In: Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2023 Jun 17–24; Vancouver, BC, Canada. p. 17939–48.
22. Liu Z, Barzen J, Bechtold M, Dustdar S, Leymann F, Raith P, et al. SparseTrack: multi-object tracking by performing scene decomposition based on pseudo-depth. arXiv:2305.05238. 2023. [arXiv]
23. Meng T, Kaminskyi D, Wittbold F, Dierl S, Howar F, König B, et al. Localization-guided track: a deep association multi-object tracking framework based on localization confidence of detections. arXiv:2303.14111. 2023. [arXiv]
24. Jung H, Kang S, Kim T, Kim H. ConfTrack: kalman filter-based multi-person tracking by utilizing confidence score of detection box. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV); 2024 Jan 3–8, Waikoloa, HI, USA. p. 6583–592.
25. Stadler D. A detailed study of the association task in tracking-by-detection-based multi-person tracking. In: Proceedings of the 2022 joint workshop of fraunhofer IOSB and institute for anthropomatics, vision and fusion laboratory. Karlsruhe, Germany: KIT Scientific Publishing; 2023. p. 59–85.

26. Stadler D, Beyerer J. An improved association pipeline for multi-person tracking. In: Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW); 2023 Jun 17–24; Vancouver, BC, Canada. p. 3170–3179.
27. Khan SBJ, Li C, Zhang P. FocusTrack: enhancing object detection and tracking for small and ambiguous objects. J Vis Commun Image Represent. 2025;111:104549.