# FusionTrack: Collaborative Multi-Object Tracking with Arbitrary Multi-UAV

Xiaohe Li[1,2], Pengfei Li[1,2,3,4], Kaixin Zhang[1,2,3,4], Jiahao Li[1,2,3,4], and Zide Fan[1,2⋆]

[1] Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing, China
[2] Key Laboratory of Target Cognition and Application Technology (TCAT), Beijing, China
[3] University of Chinese Academy of Sciences, Beijing, China
[4] School of Electronic, Electrical and Communication Engineering, University of Chinese Academy of Sciences, Beijing, China
{lixiaohe,fanzd}@aircas.ac.cn
{zhangkaixin25,lijiahao243}@mails.ucas.ac.cn
fly723822@163.com

**Abstract.** Multi-object tracking (MOT) via UAV fleets is crucial for low-altitude applications. However, existing multi-view multi-object tracking (MVMOT) methods and datasets are constrained by fixed camera configurations and limited coverage of complex scenarios, thereby failing to address real-world applications. To fill this gap, we first conduct a systematic study of MOT under arbitrary camera setups and introduce MDMOT, a novel multi-UAV benchmark covering diverse real-world scenarios, exposing practical challenges like random object entry/exit and cross-view appearance inconsistencies. We further propose FusionTrack, an end-to-end MVMOT framework that abandons the traditional decoupled track-then-associate paradigm. It jointly optimizes tracking and association via bidirectional object fusion between a Tracklet Memory Pool and a Trajectory Identity Pool, leveraging spatio-temporal context to enhance representation discriminability. For inference, we design View-aware Hierarchical Clustering with neighbor filtering to ensure intra-view exclusivity and inter-view consistency in cross-view association. Extensive experiments verify that FusionTrack achieves state-of-the-art performance in both single- and multi-view tracking tasks, outperforming existing methods by at least 2.3% in relative improvement on our MDMOT benchmark. Project page: https://github.com/aircas501/FusionTrack

**Keywords:** Multi-View Multi-Object Tracking · Multi-UAV Tracking · Collaborative Tracking · Cross-View Association

## 1 Introduction

Multi-object tracking (MOT) in UAV imagery is a core enabling technology for low-altitude applications including urban governance and public security surveillance. However, single UAV systems suffer from limited field-of-view coverage and inherent ego-motion perturbations, rendering single-view observations insufficient for large-scale

⋆ Corresponding author

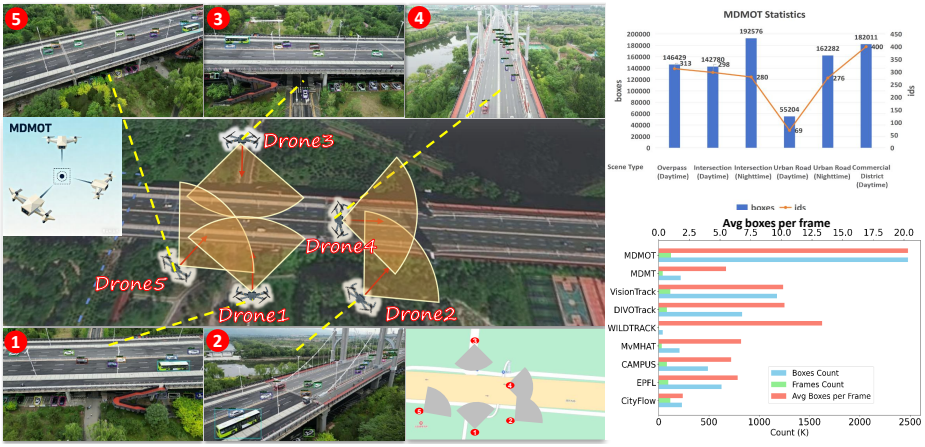



**Fig. 1:** Description of the MDMOT dataset. Left panel: drones spatial distribution in an overpass scenario. Top-right panel: bounding-box distribution and identity frequency statistics across collection scenes. Bottom-right panel: target count comparison between MDMOT and other MV-MOT datasets.

monitoring tasks. These limitations have driven growing research interest in UAV-fleet-based multi-view multi-object tracking (MVMOT) [20], which jointly performs tracklet estimation and cross-view target association to expand spatial coverage and boost tracking robustness via collaborative perception. Nevertheless, MVMOT brings unique under-explored challenges: view-dependent occlusions, unconstrained field-of-view overlap, and severe cross-view appearance discrepancies, all of which critically undermine reliable cross-camera identity association.

Existing MVMOT benchmarks are divided by field-of-view configuration into overlapping and non-overlapping settings [1]. Overlapping benchmarks dominate the field, including ground-based pedestrian datasets (e.g., MvMHAT [9], MMP-Tracking [12]) and drone-based datasets (e.g., DIVOTrack [13], MDMT [20], VisionTrack [7]), while non-overlapping settings are mainly covered by CityFlow [27] and HST [39], with recent works expanding to multi-category tracking [32]. However, nearly all existing benchmarks use static camera setups with fixed view correspondence, failing to model the unconstrained dynamic flight of real-world UAV fleets, where field-of-view coverage and camera positions are unpredictable, and overlapping view groups change dynamically (Fig. 1). Tracking and identity matching under such arbitrary view settings remain inadequately addressed.

To fill this gap, we conduct the first systematic study of MOT under arbitrary views and unconstrained camera configurations, and propose MDMOT, a novel multi-UAV MOT benchmark collected from diverse real-world scenarios. MDMOT covers both overlapping and non-overlapping view configurations with dense targets, and reveals three core challenges for unconstrained UAV MVMOT: (1) Arbitrary target entry/exit invalidating fixed boundary assumptions of conventional methods; (2) Severe motion blur from high-speed UAV/target movement; (3) Drastic cross-view appear-

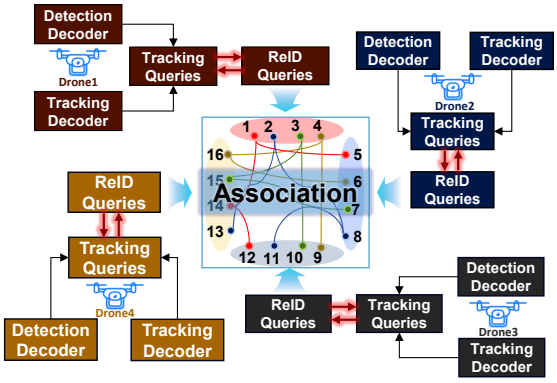



**Fig. 2:** FusionTrack achieves tracking and association within a unified closed-loop formulation.

ance/geometric inconsistencies from varying viewpoints, which cripple reliable cross-view identity association.

Methodologically, most existing MVMOT methods adopt a dominant two-stage track-then-associate paradigm: they first generate per-view tracklets via single-view MOT (e.g., DeepSORT [30], ByteTrack [40]), then link cross-view tracklets via three mainstream strategies: (1) ReID feature similarity matching (e.g., MTMCT [17]), heavily bounded by upstream tracking quality; (2) Cross-view homography estimation via image registration (e.g., SIFT [21], MIA-Net [20]), limited by strict view overlap assumptions, complex pipelines and high computational cost; (3) Matrix factorization-based matching (e.g., NMF [16]), suffering from poor optimization stability and low efficiency. Despite progress in constrained scenarios, these decoupled pipelines suffer from inherent cascaded error accumulation, and fail to fuse tracking and identity matching to leverage cross-view information for correcting tracklet drift and ID switches.

Motivated by the strong visual representation capability of Transformer-based models [38], we propose FusionTrack, an end-to-end MOT framework that supports an arbitrary number of views with unconstrained camera configurations. Unlike fragmented track-and-associate pipelines that "track first and re-identify later," FusionTrack jointly estimates trajectories and performs cross-view identity reasoning within a unified closed-loop formulation, as shown in Fig. 2. For single-view tracking, we maintain temporal continuity and handle cross-view re-entry via a Tracklet Memory Pool, while a dedicated Trajectory Identify Pool enables global identity matching across views. An Object Fusion Module updates current-frame queries by aggregating spatio-temporal context, producing more discriminative and temporally consistent object representations.

Concretely, tracking benefits from ReID-aware features that enforce identity consistency across time and views, suppressing drift under occlusion and viewpoint changes, while ReID benefits from tracking-aware temporal aggregation that provides longer, clearer evidence than isolated detections or short tracklets. Fusing the features of two related tasks and achieving collaborative training can help improve the overall performance. During inference, we apply an efficient post-processing pipeline, called View-aware Hierarchical Clustering, to achieve accurate and scalable cross-view association.

To summarize, our contributions are fourfold:

- To our knowledge, we conduct the first systematic study of arbitrary-view MOT under unconstrained camera configurations, and propose MDMOT, a novel multi-UAV MOT benchmark built on diverse real-world scenarios.
- We propose FusionTrack, an end-to-end multi-view tracking framework for joint tracking and ReID optimization via bidirectional object fusion between Tracklet Memory Pool and Trajectory Identity Pool.
- We design a View-aware Hierarchical Clustering with neighbor filtering, ensuring intra-view exclusivity and inter-view consistency with efficient computation for inference-stage target association.
- Extensive experiments show FusionTrack achieves state-of-the-art performance on multiple single-view tracking and cross-view association benchmarks, outperforming competitors by at least 2.3% on our MDMOT multi-view tracking benchmark.

## 2 RELATED WORK

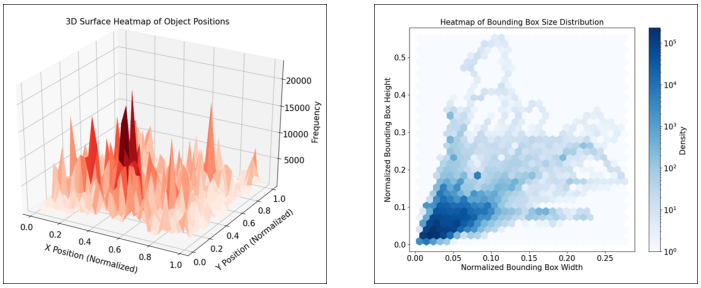
Multi-object tracking (MOT) is a fundamental computer vision task that localizes targets, preserves cross-frame identities, and generates consistent trajectories, with core applications in surveillance, autonomous driving, and UAV perception. Existing MOT research falls into two core branches: single-view MOT for monocular inputs, and multi-view multi-object tracking (MVMOT) leveraging complementary spatio-temporal information from synchronized multi-camera systems.

Single-view MOT is dominated by two paradigms: Tracking-by-Detection (TBD) and Joint Detection and Tracking (JDT). TBD decouples tracking into frame-level detection and data association, benefiting from mature detection techniques but bounded by detector performance. Classic works include DeepSORT [30] (Kalman filtering + appearance/IoU matching) and ByteTrack [40] (two-stage matching for occlusion robustness), with recent advances focusing on multi-granularity cues and reduced handcrafted rules [24, 34]. JDT unifies detection and association into an end-to-end framework to eliminate decoupled error accumulation. FairMOT [41] first balances detection and ReID optimization, while Transformer-based JDT works [26, 38] unify the pipeline via learnable target queries, with recent efforts mitigating training imbalance and enabling heuristic-free learnable association [10, 23]. All single-view methods, however, are inherently limited by narrow FoV and occlusion-induced performance degradation.

MVMOT addresses single-view limitations via cross-camera information fusion for globally consistent trajectories. Traditional MVMOT adopts a two-stage pipeline: per-camera single-view tracking, followed by cross-view target association. Cross-view matching relies heavily on ReID features to handle viewpoint, illumination, and occlusion variations [37], yet state-of-the-art ReID models [15, 36] fail to generalize to unconstrained dynamic scenarios. Existing works improve association via self-supervised learning [9], geometric constraints [20, 25], and domain priors [35], while recent trends shift to end-to-end frameworks for joint single-view feature learning and cross-view association, reducing error accumulation via global modeling and cross-view embedding

**Table 1:** Comparison of MDMOT and other datasets.

| Dataset | Source | Resolution | Scenes | Views | Frames | Boxes | Average Boxes | Moving Cameras | View Distribution |
|---|---|---|---|---|---|---|---|---|---|
| EPFL [8] | Monitor | 360×288 | 5 | 2-4 | 97K | 625k | 6.44 | × | overlap |
| CAMPUS [33] | Monitor | 1920×1080 | 4 | 4 | 83K | 490K | 5.90 | × | overlap,non-overlap |
| MvMHAT [9] | Monitor | 1920×1080 | 1 | 3-4 | 31K | 208K | 6.71 | ✓ | overlap,non-overlap |
| WILDTRACK [3] | Monitor | 1920×1080 | 1 | 7 | 3K | 40K | 13.33 | × | overlap |
| CityFlow [27] | Monitor | 1280×960 | 1 | 46 | 117K | 230K | 1.97 | × | overlap,non-overlap |
| MDMT [20] | Drone | 1920×1080 | 6 | 2 | 40K | 220K | 5.50 | ✓ | overlap |
| DIVOTrack [13] | Mobile,Drone | 3640×2048,1920×1080 | 10 | 3 | 81K | 830K | 10.25 | ✓ | overlap |
| VisionTrack [7] | Drone | 1920×1080 | 15 | 2 | 116K | 1176K | 10.14 | ✓ | overlap |
| MDMOT | Drone | 1920×1080 | 6 | 3-5 | 122K | 2481K | 20.34 | ✓ | overlap,non-overlap |




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

**Fig. 3:** (a) Heatmap of target location distribution: x/y are normalized image coordinates, and z denotes occurrence frequency. (b) Heatmap of box size distribution: x/y represent the relative box width and height.

optimization [4, 6, 7, 13]. Most existing MVMOT methods, however, target fixed camera setups, with no systematic study on arbitrary-view MOT under unconstrained UAV camera configurations.

## 3 MDMOT Benchmark

To advance research on MVMOT in realistic UAV settings, we introduce Multi-Drone Multi-Object Tracking (MDMOT), a multi-scene benchmark captured by a coordinated drone fleet. To our knowledge, MDMOT is the first dataset to support arbitrary camera settings under unconstrained, dynamically changing viewpoints.

*Data Collection* MDMOT is collected using a fleet of five DJI Mini 3 Pro drones. We record 20 synchronized multi-view video sequences at $1920 \times 1080$ resolution, covering 6 real-world urban scenes, including Overpasses, Intersections (Day/Night), Urban Road (Day/Night) and Commercial District. Detailed descriptions of each scene are in the appendix. Each environment contains 3-4 video segments, and 3-5 drones simultaneously capture each segment. More descriptions of MDMOT are provided in Appendix 1.

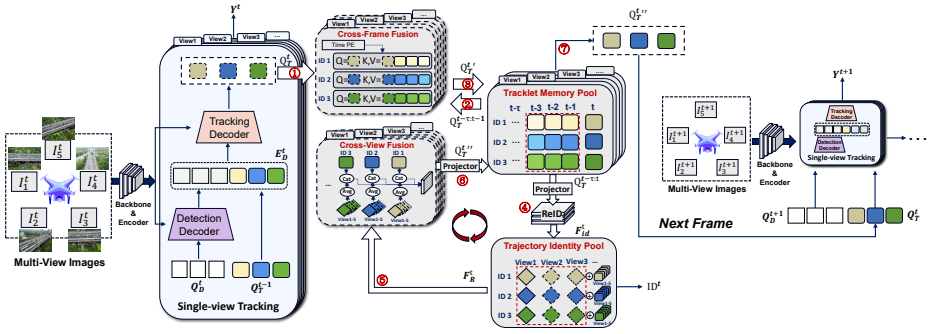



**Fig. 4:** Overview of the proposed FusionTrack framework.

*Statistics* MDMOT targets the inherent spatial and scale variability of drone imagery. Unlike fixed camera datasets with strong spatial priors, UAV mobility yields a broadly uniform target spatial distribution (Fig. 3(a)). Drastic scale variations with flight altitude cause imbalanced bounding box distributions (Fig. 3(b)), aggravating detection and identity preservation difficulties. We split the dataset into 8:1:1 training/validation/test sets for standardized evaluation.

*Comparison with Prior Benchmarks* Table 1 benchmarks MDMOT against state-of-the-art multi-view tracking datasets. Early works (EPFL [8], CAMPUS [33], MvMHAT [9], WILDTRACK [3], CityFlow [27]) focus on fixed-camera pedestrian/vehicle tracking, whose static layouts simplify cross-view association but lack viewpoint diversity for UAV scenarios. Recent dynamic benchmarks (DIVOTrack [13], VisionTrack [7], MDMT [20]) explore drone-based tracking, but suffer from limited views, coverage and real-world diversity. In contrast, MDMOT constructs a 3–5 drone coordinated observation network, with 122k frames and 24.81M annotations collected across diverse urban scenes. It introduces richer cross-view motion variations and realistic challenges (occlusion, rapid viewpoint/illumination shifts), offering a more representative benchmark for open-scene MVMOT evaluation.

## 4 METHOD

Fig. 4 overviews FusionTrack, an end-to-end architecture unifying single-view tracking and cross-view association, departing from conventional track-then-associate paradigms.

### 4.1 Problem Definition

We consider a synchronized multi-view multi-object tracking (MVMOT) system with $C$ cameras. Let $\mathcal{I} = \{I_c^t\}_{c=1,t=1}^{C,T}$ denote input video frames, where $I_c^t$ is the frame from camera $c$ at time $t$. For each camera $c$, single-view tracking (SVT) outputs tracklets $\mathcal{T}^c = \{T_i^c\}_{i=1}^{N_c}$ ($N_c$: target count in view $c$). Each tracklet $T_i^c$ is a temporally ordered state sequence, with state at time $s$ defined as:

$$t_i^{c,s} = \left(B_i^{c,s}, \mathrm{CLS}_i^{c,s}, \mathrm{ID}_i^{c,s}\right), \tag{1}$$

where $B_i^{c,s} = (x, y, w, h)$ is the bounding box (top-left corner $(x, y)$ and its width $w$ and height $h$), $\text{CLS}_i^{c,s}$ the object category, and $\text{ID}_i^{c,s}$ the target identity. MVMOT aims to maintain intra-view temporal trajectory consistency and inter-view identity association.

### 4.2 Joint Training of Tracking and Association

Building on query-based Transformer tracking paradigms [11,38], FusionTrack extends single-view query tracking to multi-view settings. We first introduce the SVT backbone, which generates track queries for subsequent cross-view refinement.

**Single-View Tracking** For each frame $I_c^t$, we adopt a DETR-style architecture [11]: a ResNet-50 backbone [14] extracts features, contextualized by a Transformer encoder to produce $F_c^t$.

*Detection Decoder* Given fixed learnable detection queries $Q_D \in \mathbb{R}^{N_D \times d}$ ($N_D$: max detection count, $d$: feature dimension), the decoder outputs detection embeddings via cross-attention over $F_c^t$:

$$E_D^t = \text{DetectionDecoder}(Q_D^t, F_c^t). \tag{2}$$

*Tracking Decoder* To maintain temporal consistency, the tracking decoder concatenates previous active track queries $Q_T^{t-1} \in \mathbb{R}^{N_T \times d}$ and $E_D^t \in \mathbb{R}^{N_D \times d}$, then decodes with $F_c^t$ to update track queries:

$$Q_T^t = \text{TrackingDecoder}\big(\text{Concat}\{Q_T^{t-1}, E_D^t\}, F_c^t\big). \tag{3}$$

Updated embeddings are used for box regression and next-step track query initialization. Tracking supervision follows standard DETR [11] bipartite matching for one-to-one assignment.

### Bidirectional Fusion of Tracking and Association

*(1) Re-identification Feature Extraction.* We maintain a Tracklet Memory Pool (TMP) to buffer recent track queries within a temporal window $\tau$: $\{Q_T^{t-\tau}, \ldots, Q_T^t\}$, preserving temporal cues for cross-view association.

Since tracking queries focus on short-term continuity while association requires identity-discriminative embeddings, we introduce a lightweight linear projection $\phi_1$ to decouple the tracking feature space from the identity space, projecting TMP queries to $\{R_T^{t-\tau}, \ldots, R_T^t\} = \phi_1(\{Q_T^{t-\tau}, \ldots, Q_T^t\})$. A multi-layer FFN $f$ then aggregates these projected queries to extract ReID features $F_{id}^t \in \mathbb{R}^d$:

$$F_{id}^t = f(\text{concat}(\{R_T^{t-\tau}, \ldots, R_T^t\})). \tag{4}$$

*(2) Momentum Update in the Trajectory Identity Pool.* We further maintain a Trajectory Identity Pool (TIP) to cache cross-view identity representations. After identity prediction, we update stored features via a momentum rule for temporal stability:

$$F_R^t = \mu F_R^{t-1} + (1-\mu) F_{id}^t, \tag{5}$$

where $\mu \in [0, 1)$ is the momentum coefficient, and $F_R^{t-1}$ is the cached feature of the same identity from the previous timestep. During training, we fuse $F_R^{t-1}$ and $F_{id}^t$ based on cross-view ground truth.

### 4.3 Object Fusion Module

To leverage spatio-temporal context and multi-view features, we propose an Object Fusion Module (OFM) that refines track queries via two-step cross-frame and cross-view aggregation, as shown in Fig. 4.

**Cross-Frame Fusion** Before inserting $Q_T^t$ into the Tracklet Memory Pool (TMP), we perform cross-frame temporal fusion via CrossAttention to preserve both short-term continuity and long-term memory. We attend to past $\tau$ queries from TMP $\{Q_T^{t-\tau}, \ldots, Q_T^{t-1}\}$ with temporal positional encoding:

$$\begin{aligned} {Q_T^t}' = \mathrm{CrossAttn}\Big(&Q = Q_T^t,\, K, V = \{Q_T^{t-\tau}, \ldots, Q_T^t\}, \\ &PE = \mathrm{Pos}(t-\tau{:}t)\Big), \end{aligned} \tag{6}$$

where missing frames are represented by zero vectors. To prioritize recent observations, we apply a time-decayed modulation to scaled dot-product logits:

$$\begin{aligned} \mathrm{CrossAttn}(Q, K, V) &= \mathrm{Softmax}\left(\frac{QK^\top}{\sqrt{d_k}} \odot W\right) \cdot V, \\ W &= \exp(-\alpha \cdot \Delta t), \end{aligned} \tag{7}$$

where $\Delta t \in [0, \tau]^{1\times\tau}$ is the temporal distance to frame $t$, and $\alpha = 0.1$ controls the decay rate.

**Cross-View Fusion** To reduce viewpoint discrepancies and enable bidirectional fusion, we aggregate multi-view identity features from the Trajectory Identity Pool (TIP) and project them back to refine track queries. Specifically, we average TIP's multi-view identity features $F_R^t$, then concatenate with the track query:

$${Q_T^t}'' = Q_T^t + \phi_2(\mathrm{Concat}(Q_T^t, \mathrm{Avg}(F_R^t|_1^C))), \tag{8}$$

where $\phi_2$ is a linear projection opposite to $\phi_1$, mapping identity features back to the tracking space.

This forms a closed loop: temporally coherent track queries enable cross-view association via TIP, while identity-consistent TIP features enhance tracking under occlusion/viewpoint changes. Bidirectional TMP-TIP coupling thus enables joint end-to-end optimization of tracking and association.

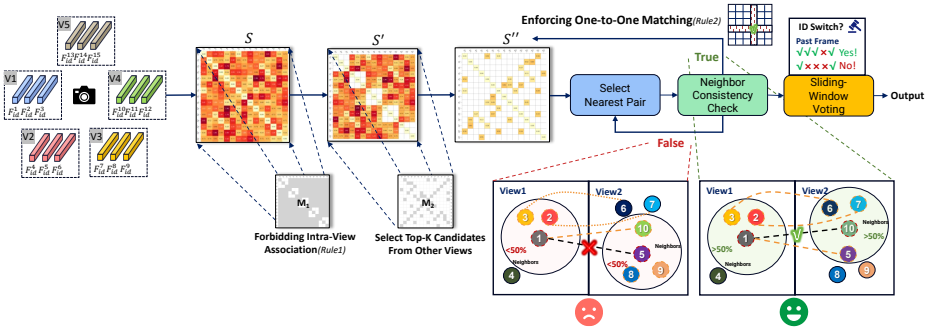



**Fig. 5:** The process of View-aware Hierarchical Clustering.

### 4.4 Joint Loss Function

To enable joint optimization of tracking and ReID, we employ a combined objective with an uncertainty-aware weighting scheme [41]:

$$\mathcal{L}_{\text{total}} = e^{-\omega_1}\mathcal{L}_T + e^{-\omega_2}\mathcal{L}_R + \omega_1 + \omega_2, \tag{9}$$

where $\omega_1, \omega_2$ are learnable scalars adaptively balancing the two tasks.

The tracking loss consists of:

$$\mathcal{L}_T = \mathcal{L}_{cls} + \mathcal{L}_{reg} + \mathcal{L}_{giou}, \tag{10}$$

where $\mathcal{L}_{cls}$ is focal loss, $\mathcal{L}_{reg}$ is $\ell_1$ regression loss, and $\mathcal{L}_{giou}$ is Generalized IoU loss.

The ReID loss is:

$$\mathcal{L}_R = \mathcal{L}_{id} + \mathcal{L}_{trip} + \mathcal{L}_{rc}, \tag{11}$$

where $\mathcal{L}_{id}$ is identity classification cross-entropy, $\mathcal{L}_{trip}$ is triplet loss for discriminative embeddings. To constrain the projection between tracking-query and ReID embedding spaces, we introduce a bidirectional reconstruction regularizer $\mathcal{L}_{rc}$ based on MSE of forward-inverse mappings to prevent collapse:

$$\mathcal{L}_{rc} = \frac{1}{N}((Q_T^{t\,\prime} - \phi_2(\phi_1(Q_T^{t\,\prime})))^2 + (R_T^t - \phi_1(\phi_2(R_T^t)))^2. \tag{12}$$

which stabilizes feature transformation and enhances cross-view identity robustness.

We adopt a staged training schedule: first pre-train the single-view tracking branch with $\mathcal{L}_T$ to obtain stable features, then introduce $\mathcal{L}_R$ and $\mathcal{L}_{rc}$ after a fixed number of epochs.

### 4.5 View-aware Hierarchical Clustering

To achieve reliable trajectory association across arbitrary views and avoid local optimal matching during inference, we propose view-aware hierarchical clustering (VHC) with neighbor filtering, following two core rules:

**Rule 1:** Forbid intra-view association by setting distances between same-view objects to $+\infty$.

**Rule 2:** Matched target pairs cannot re-match with objects from each other's views.

As shown in Fig. 5, VHC proceeds as follows: First, compute a cosine similarity matrix $S \in \mathbb{R}^{N_f \times N_f}$ ($S_{i,j} = \cos(F_{id}^i, F_{id}^j)$) for all current-frame object queries, and apply an intra-view mask $M_1$ to restrict clustering to cross-view scenarios (Rule 1). Second, use a mutual top-$k$ ($k = 5$) neighbor filtering strategy to generate a sparse candidate graph $M_2$, eliminating ambiguous associations and improving efficiency. Third, introduce a Neighborhood Filtering Mechanism (NFM) leveraging local spatial neighborhood stability: for the nearest candidate match $(i, j)$, retain it only if the matching proportion between spatial neighbor sets $\mathcal{N}(i)$ and $\mathcal{N}(j)$ exceeds threshold $\delta$, filtering spurious matches with similar appearance but inconsistent context. Fourth, apply a top-down association process to avoid multiple matches (Rule 2). Finally, integrate a Sliding-Window Voting (SWV) mechanism to leverage historical association results: an identity switch is only permitted if the same switch is supported by more than half of the past frames within a window size empirically set to 5, ensuring temporal consistency and preventing noisy ID switches.

VHC achieves $\mathcal{O}(N_f^2 \log N_f)$ time complexity via a priority queue-accelerated agglomerative pipeline, where the primary computational bottleneck lies in nearest candidate match search and cluster aggregation. Masks $M_1$ and $M_2$ make the similarity matrix highly sparse, significantly reducing the actual computational overhead. The detailed inference process and full strategies are provided in Appendix 2.

# 5 EXPERIMENTS

## 5.1 Implementation Details

FusionTrack uses ResNet-50 [14] initialized with DAB-Deformable-DETR [19] COCO [18] weights. All experiments are conducted in PyTorch1.13 on 4×NVIDIA A800 GPUs, trained for 100 epochs via Adam ($lr = 10^{-5}$) with a batch size of 2. A progressive schedule mitigates early cross-view instability: (1) single-view tracking only for 20 epochs; (2) additionally introduce the ReID module for another 20 epochs; (3) full joint refinement in the remaining epochs. All baselines are retrained on MDMOT per open-source settings. We adopt CVMA/CVIDF1 (multi-view extensions of MOTA/IDF1 from DIVOTrack [13]) and standard community metrics for single-view tracking. Additional details can be found in Appendix 3.

## 5.2 Multi-View Tracking Results

**Results on our MDMOT** Table 2 (first column) evaluates multi-view tracking on MDMOT against representative baselines: ReID-based methods (OSNet [42], Strong [22], AGW [36]) and multi-view trackers (CityTrack [35], MvMHAT [9], CrossMOT [13], GMT [7]). GMT is the strongest prior work (79.9 CVMA, 73.8 CVIDF1). Our FusionTrack achieves 81.8 CVMA and 75.5 CVIDF1, outperforming GMT by 1.9 and 1.7 points, respectively. Gains on both association accuracy (CVMA) and identity preservation (CVIDF1) validate its reliable cross-view matching.

**Table 2:** Comparison of multi-view tracking performance on MDMOT, CAMPUS, WILDTRACK, MvMHAT and DIVOTrack. The best two results are shown in red and blue.

| | MDMOT | | CAMPUS | | WILDTRACK | | MvMHAT | | DIVOTrack | |
|---|---|---|---|---|---|---|---|---|---|---|
| Methods | CVMA↑ | CVIDF1↑ | CVMA↑ | CVIDF1↑ | CVMA↑ | CVIDF1↑ | CVMA↑ | CVIDF1↑ | CVMA↑ | CVIDF1↑ |
| OSNet | 50.1 | 45.5 | 58.8 | 47.8 | 10.8 | 18.2 | 92.6 | 87.7 | 33.0 | 44.9 |
| Strong | 52.3 | 51.4 | 63.4 | 55.0 | 28.6 | 41.6 | 49.0 | 55.1 | 39.1 | 44.7 |
| AGW | 49.3 | 49.9 | 60.8 | 52.8 | 15.6 | 23.8 | 92.5 | 86.6 | 15.6 | 23.8 |
| CT | - | - | 63.7 | 55.0 | 19.0 | 42.0 | 46.7 | 53.5 | 64.9 | 65.0 |
| MGN | - | - | 63.3 | 56.1 | 32.6 | 46.2 | 92.3 | 87.4 | 33.5 | 39.4 |
| Citytrack | 57.3 | 56.1 | - | - | - | - | - | - | - | |
| MvMHAT | 78.8 | 63.7 | 56.0 | 55.6 | 10.3 | 16.2 | 70.1 | 68.4 | 61.1 | 62.6 |
| CrossMOT | 78.5 | 72.9 | 65.6 | 61.2 | 42.3 | 56.7 | 92.3 | 87.4 | 72.4 | 71.1 |
| GMT | 79.9 | 73.8 | 66.4 | 66.8 | 61.7 | 72.0 | 94.1 | 95.6 | 74.5 | 73.2 |
| FusionTrack | 81.8 | 75.5 | 68.6 | 67.5 | 62.5 | 72.3 | 94.8 | 95.9 | 74.8 | 77.5 |

**Table 3:** Comparison of SVT results on MDMOT. The best two results among the Transformer-based methods are shown in red and blue.

| Methods | MOTA↑ | MOTP↑ | IDF1↑ | MT↑ | ML↓ | HOTA↑ | FP↓ | FN↓ | IDS↓ |
|---|---|---|---|---|---|---|---|---|---|
| *CNN-based*: | | | | | | | | | |
| DeepSORT | 84.12 | 84.45 | 88.32 | 520 | 50 | 79.70 | 45324 | 28532 | 280 |
| CenterTrack | 88.36 | 87.60 | 92.41 | 740 | 29 | 83.33 | 37698 | 21913 | 248 |
| FairMOT | 88.96 | 87.32 | 91.80 | 709 | 27 | 86.30 | 37072 | 19876 | 103 |
| TraDes | 86.78 | 86.66 | 91.38 | 733 | 41 | 81.40 | 43589 | 24189 | 221 |
| ByteTrack | 86.38 | 87.94 | 91.96 | 703 | 48 | 86.90 | 34506 | 29468 | 135 |
| OC-SORT | 89.42 | 88.73 | 92.15 | 768 | 18 | 85.12 | 32105 | 18542 | 102 |
| HybridSORT | 89.81 | 90.58 | 93.28 | 783 | 22 | 86.10 | 30183 | 18375 | 98 |
| *Transformer-based*: | | | | | | | | | |
| TransTrack | 84.17 | 86.42 | 89.37 | 498 | 47 | 82.35 | 35542 | 32584 | 202 |
| MOTR | 87.03 | 86.82 | 91.13 | 634 | 57 | 81.42 | 38945 | 27102 | 103 |
| MeMOTR | 87.57 | 87.69 | 91.97 | 719 | 21 | 83.94 | 31984 | 24980 | 92 |
| COMOT | 86.11 | 88.44 | 92.05 | 721 | 43 | 84.72 | 35826 | 29103 | 101 |
| MOTIP | 88.10 | 88.32 | 91.93 | 602 | 25 | 84.26 | 32174 | 24125 | 90 |
| FusionTrack | 88.75 | 89.22 | 93.06 | 776 | 19 | 84.88 | 30512 | 22328 | 81 |

**Results on Other Datasets** Table 2 (the four columns on the right) reports multi-view tracking performance on four public benchmarks: CAMPUS, WILDTRACK, MvMHAT, and DIVOTrack. We compare FusionTrack against representative baselines, including ReID backbones (OSNet [42], Strong [22], AGW [36], CT [29], MGN [28]) and multi-view tracking methods (MvMHAT [9], CrossMOT [13], GMT [7]). FusionTrack achieves the best performance on all four datasets. On CAMPUS, it reaches 68.6/67.5 (CVMA/CVIDF1), improving over the strongest prior method GMT by 2.2 CVMA and 0.7 CVIDF1. On WILDTRACK, FusionTrack attains 62.5/72.3, yielding 0.8 and 0.3 gains. On MvMHAT, it further improves to 94.8/95.9 (+0.7 and +0.3). On DIVOTrack, FusionTrack obtains 74.8/77.5 and delivers the largest identity-consistency gain, surpassing GMT by 0.3 CVMA and 4.3 CVIDF1. These results demonstrate FusionTrack's strong generalizability across diverse multi-camera settings.

### 5.3 Single-view Tracking Results on our MDMOT

Table 3 reports single-view MOT results on MDMOT, comparing FusionTrack with CNN-based trackers (DeepSORT [30], CenterTrack [5], FairMOT [41], TraDes [31], ByteTrack [40], OC-SORT [2], HybridSORT [34]) and Transformer-based trackers (TransTrack [26], MOTR [38], MeMOTR [11], COMOT [23], MOTIP [10]) (top two Transformer results highlighted per table note). As a state-of-the-art Transformer tracker, FusionTrack achieves 88.75 MOTA, 89.22 MOTP, 93.06 IDF1, with the highest HOTA (84.88) and MT (776), and lowest ML (19). It also records the fewest FP (30512), FN (22328), and IDS (81) among Transformer-based methods, outperforming rivals like COMOT and MOTIP on both accuracy (MOTA/MOTP) and identity (IDF1/IDS) metrics. This validates its effectiveness in enhancing trajectory tracking via multi-view information fusion.

### 5.4 Ablation Studies

We ablate key training/inference components on MDMOT (Table 4). Training-stage ablations evaluate cross-frame fusion, cross-view fusion, and the reconstruction loss, while inference-stage ablations examine the association strategy (VHC vs. Hungarian matching with pairwise union), the Sliding-Window Voting (SWV), and the neighborhood filtering mechanism (NFM).

**Training-stage ablation** The upper half of Table 4 shows the base model achieves 76.5 CVMA and 66.8 CVIDF1. Cross-frame fusion brings significant gains, boosting performance to 78.8/72.2, as temporal cue aggregation greatly enhances identity consistency. Adding cross-view fusion further raises it to 79.5/74.3, confirming the value of multi-view complementarity. Finally, incorporating reconstruction loss delivers the optimal performance (81.8/75.5), as this extra constraint enhances feature learning and cross-view association.

**Inference-stage ablation** The lower half of Table 4 shows VHC alone performs poorly (77.1/72.3), emphasizing the challenge of reliable association without extra constraints. Sliding-window voting (SWV) drastically boosts performance to 79.2/74.1, confirming temporal voting stabilizes associations. Combining SWV with NFM yields further gains: Hungarian matching + SWV + NFM achieves 78.5/73.3, and VHC + SWV + NFM delivers optimal results (81.8/75.5). This indicates NFM effectively suppresses noisy matches, and VHC benefits more from neighborhood-aware filtering under multi-view ambiguity. The full configuration validates all proposed modules' effectiveness.

### 5.5 Case Analysis

To intuitively compare performance on MDMOT, we visualize results of FusionTrack, CrossMOT [13], MvMHAT [9], and GMT [7] (Fig. 6). Rows and columns represent different views and methods, respectively, with blue and green arrows indicating correct matches and red arrows indicating errors. Views 1/3/4 partially overlap, while View 2

**Table 4:** Ablation study of training stage.

| Training stage | | | Inference stage | | | Metrics | |
|---|---|---|---|---|---|---|---|
| Cross-frame | Cross-view | Reconstruction Loss | Association Association | SWV | NFM | CVMA↑ | CVIDF1↑ |
| | | | VHC | ✓ | ✓ | 76.5(-5.3) | 66.8(-8.7) |
| ✓ | | | VHC | ✓ | ✓ | 78.8(-3.0) | 72.2(-3.3) |
| ✓ | ✓ | | VHC | ✓ | ✓ | 79.5(-2.3) | 74.3(-1.2) |
| ✓ | ✓ | ✓ | VHC | | | 77.1(-4.7) | 72.3(-3.2) |
| ✓ | ✓ | ✓ | VHC | ✓ | | 79.2(-2.6) | 74.1(-1.4) |
| ✓ | ✓ | ✓ | Hungarian Matching | ✓ | ✓ | 78.5(-3.3) | 73.3(-2.2) |
| ✓ | ✓ | ✓ | VHC | ✓ | ✓ | **81.8** | **75.5** |

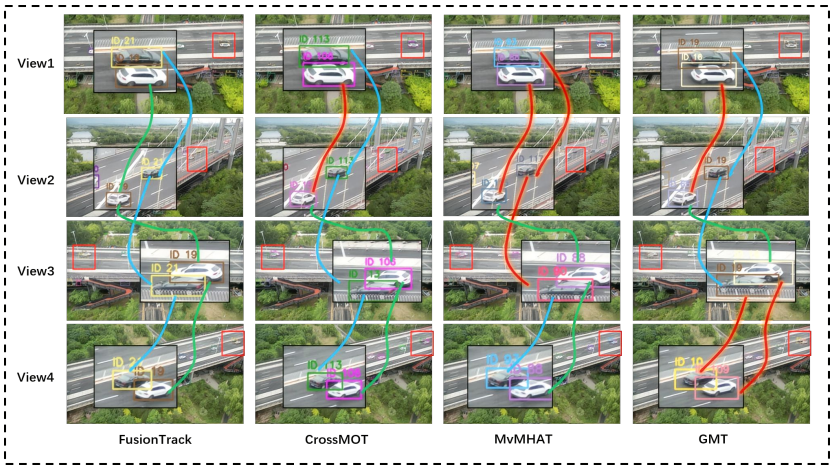



**Fig. 6:** Illustrative results of associated tracking on MDMOT.

is independent of the others. The visualization results clearly show that FusionTrack significantly reduces cross-view association errors compared to baselines, maintaining correct identity matching across overlapping/non-overlapping views. For example, all baselines failed to associate the white car in non-overlapping View 2, and GMT suffered severe errors between Views 3/4 due to vehicle occlusion. This advantage stems from OFM and VHC, which jointly enhance the extraction of high-quality cross-view identity features.

### 5.6 Computational Efficiency

Inference was performed on a single NVIDIA A800 80G GPU. The FusionTrack model achieves an average frame rate of 3.9 fps across five scenes. As shown in Table 5, single-view tracking across different views runtime dominates the total network latency. Parallel tracking processing will be explored in future work to further improve real-time performance.

**Table 5:** Model parameter size and real-time performance.

| | Backbone | SVTransformer | OFM | VHC |
|---|---|---|---|---|
| Params | 25.8M | 85M | 4.72M | 1.6M |
| Flops | 4.1G | 23G | 1.41G | 0.8G |

### 5.7 Qualitative Results

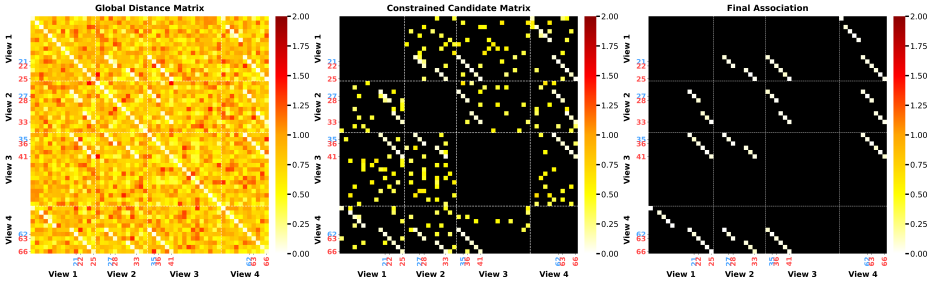
To intuitively explain FusionTrack's fewer cross-view association errors, we provide distance-matrix heatmaps and t-SNE visualizations in Fig. 7 and Fig. 8, offering evidence from both "association process" and "feature distribution" perspectives. These images are produced simultaneously with the example scenario in Fig. 6, further demonstrating the effectiveness of our method.

In heatmaps (Fig. 7), light-to-dark colors indicate near-to-far distances. Black entries denote invalid pairs (infinite distances due to view and top-$k$ constraints). We highlight target ID ♯21 and visually similar distractors (♯22, ♯25) to illustrate candidate suppression.

Specifically, (1) in the raw distance matrix, multiple similar candidates simultaneously exhibit small distances, leading to clear ambiguities in cross-view matching. (2) After applying the view constraint and the top-$k$ constraint, a large number of implausible candidates are assigned infinite distances (black), substantially shrinking the candidate space and effectively reducing interference from similar distractors. (3) The final association result matrix becomes sparser and closer to a correct one-to-one structure, thanks to the effective filtering of easily confused target features by NFM.

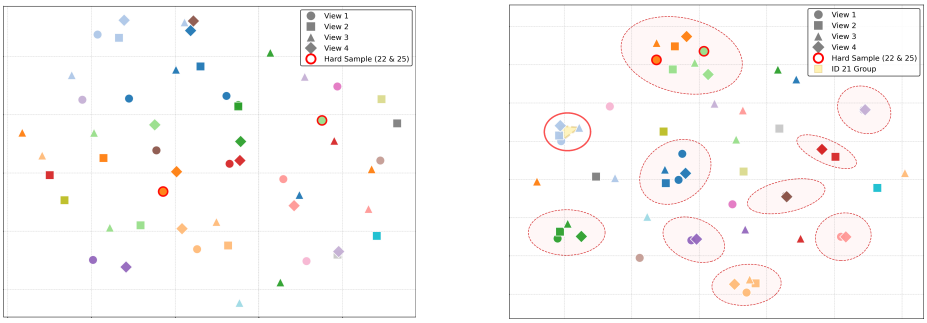
We visualize query and ReID feature distributions via t-SNE to analyze discriminability in Fig. 8. From query features to ReID features, targets with different identities become more clearly separated; combined with the post-processing steps, our method achieves more accurate associations (e.g., for the two visually similar white vehicles with ID 22 and 25 in View 1).

Other experimental results can be found in Appendix 3.




**(a)** Raw distance matrix. **(b)** After applying $M_1$ and $M_2$ filtering in VHC. **(c)** Final association matrix.

**Fig. 7:** Distance-matrix heatmaps in the association process.




**(a)** t-SNE of query features **(b)** t-SNE of ReID features

**Fig. 8:** t-SNE visualization of feature distributions.

## 6 Conclusion

We explore the problem of multi-object tracking from arbitrary viewpoints in drone swarms for the first time, and propose a new benchmark named MDMOT. It covers multiple real-world scenarios and provides a reference for future community research. We further propose FusionTrack, an end-to-end framework that fuses tracking and cross-view ReID via bidirectional feature fusion. Extensive experiments demonstrate that FusionTrack achieves SOTA performance on multiple benchmarks. Future research may extend the framework to handle more complex dynamic UAV fleets and further enhance robustness to extreme environmental perturbations.

## Acknowledgements

This work has been supported by the Strategic Priority Research Program of Chinese Academy of Sciences (Grant No. XDA0310502), Future Star of Aerospace Information Research Institute, Chinese Academy of Sciences, China (Grant No. E3Z10701), National Key Laboratory of Space Target Awareness, China (Grant No. STA2025ZCA0102).

## References

1. Amosa, T.I., Sebastian, P., Izhar, L.I., Ibrahim, O., Ayinla, L.S., Bahashwan, A.A., Bala, A., Samaila, Y.A.: Multi-camera multi-object tracking: A review of current trends and future advances. Neurocomputing **552**, 126558 (2023)
2. 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. pp. 9686–9696 (2023)
3. Chavdarova, T., Baqué, P., Bouquet, S., Maksai, A., Jose, C., Lettry, L., Fua, P., Van Gool, L., Fleuret, F.: The wildtrack multi-camera person dataset. arXiv preprint arXiv:1707.09299 (2017)
4. Chen, W., Cao, L., Chen, X., Huang, K.: An equalized global graph model-based approach for multicamera object tracking. IEEE Transactions on Circuits and Systems for Video Technology **27**(11), 2367–2381 (2016)

5. Duan, K., Bai, S., Xie, L., Qi, H., Huang, Q., Tian, Q.: Centernet: Keypoint triplets for object detection. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 6569–6578 (2019)
6. Fan, H., Qiao, Y., Zhen, Y., Zhao, T., Fan, B., Wang, Q.: All-day multi-camera multi-target tracking. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 16892–16901 (2025)
7. Fan, H., Zhao, T., Wang, Q., Fan, B., Tang, Y., Liu, L.: Gmt: A robust global association model for multi-target multi-camera tracking. arXiv preprint arXiv:2407.01007 (2024)
8. Fleuret, F., Berclaz, J., Lengagne, R., Fua, P.: Multicamera people tracking with a probabilistic occupancy map. IEEE transactions on pattern analysis and machine intelligence **30**(2), 267–282 (2007)
9. Gan, Y., Han, R., Yin, L., Feng, W., Wang, S.: Self-supervised multi-view multi-human association and tracking. In: Proceedings of the 29th ACM international conference on multimedia. pp. 282–290 (2021)
10. Gao, R., Qi, J., Wang, L.: Multiple object tracking as id prediction. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 27883–27893 (2025)
11. Gao, R., Wang, L.: Memotr: Long-term memory-augmented transformer for multi-object tracking. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9901–9910 (2023)
12. Han, X., You, Q., Wang, C., Zhang, Z., Chu, P., Hu, H., Wang, J., Liu, Z.: Mmptrack: Large-scale densely annotated multi-camera multiple people tracking benchmark. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 4860–4869 (2023)
13. Hao, S., Liu, P., Zhan, Y., Jin, K., Liu, Z., Song, M., Hwang, J.N., Wang, G.: Divotrack: A novel dataset and baseline method for cross-view multi-object tracking in diverse open scenes. International Journal of Computer Vision **132**(4), 1075–1090 (2024)
14. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
15. He, S., Luo, H., Wang, P., Wang, F., Li, H., Jiang, W.: Transreid: Transformer-based object re-identification. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 15013–15022 (2021)
16. He, Y., Wei, X., Hong, X., Shi, W., Gong, Y.: Multi-target multi-camera tracking by tracklet-to-target assignment. IEEE Transactions on Image Processing **29**, 5191–5205 (2020)
17. Hsu, H.M., Cai, J., Wang, Y., Hwang, J.N., Kim, K.J.: Multi-target multi-camera tracking of vehicles using metadata-aided re-id and trajectory-based camera link model. IEEE Transactions on Image Processing **30**, 5198–5210 (2021)
18. 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: Computer vision–ECCV 2014: 13th European conference, zurich, Switzerland, September 6-12, 2014, proceedings, part v 13. pp. 740–755. Springer (2014)
19. Liu, S., Li, F., Zhang, H., Yang, X., Qi, X., Su, H., Zhu, J., Zhang, L.: Dab-detr: Dynamic anchor boxes are better queries for detr. arxiv 2022. arXiv preprint arXiv:2201.12329
20. Liu, Z., Shang, Y., Li, T., Chen, G., Wang, Y., Hu, Q., Zhu, P.: Robust multi-drone multi-target tracking to resolve target occlusion: A benchmark. IEEE Transactions on Multimedia **25**, 1462–1476 (2023)
21. Low, D.G.: Distinctive image features from scale-invariant keypoints. Journal of Computer Vision **60**(2), 91–110 (2004)
22. Luo, H., Gu, Y., Liao, X., Lai, S., Jiang, W.: Bag of tricks and a strong baseline for deep person re-identification. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops. pp. 0–0 (2019)

23. Luo, W., Zhong, Y., Gan, Y., Ma, L., et al.: Co-mot: Boosting end-to-end transformer-based multi-object tracking via coopetition label assignment and shadow sets. In: The Thirteenth International Conference on Learning Representations (2025)
24. Qin, Z., Wang, L., Zhou, S., Fu, P., Hua, G., Tang, W.: Towards generalizable multi-object tracking. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18995–19004 (2024)
25. Samplawski, C., Fang, S., Marlin, B.M.: End-to-end differentiable multi-view tracking: Architecture and fine-tuning experiments. In: 2025 28th International Conference on Information Fusion (FUSION). pp. 1–9. IEEE (2025)
26. Sun, P., Cao, J., Jiang, Y., Zhang, R., Xie, E., Yuan, Z., Wang, C., Luo, P.: Transtrack: Multiple object tracking with transformer. arXiv preprint arXiv:2012.15460 (2020)
27. Tang, Z., Naphade, M., Liu, M.Y., Yang, X., Birchfield, S., Wang, S., Kumar, R., Anastasiu, D., Hwang, J.N.: Cityflow: A city-scale benchmark for multi-target multi-camera vehicle tracking and re-identification. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8797–8806 (2019)
28. Wang, G., Yuan, Y., Chen, X., Li, J., Zhou, X.: Learning discriminative features with multiple granularities for person re-identification. In: Proceedings of the 26th ACM international conference on Multimedia. pp. 274–282 (2018)
29. Wieczorek, M., Rychalska, B., Dkabrowski, J.: On the unreasonable effectiveness of centroids in image retrieval. In: Neural Information Processing: 28th International Conference, ICONIP 2021, Sanur, Bali, Indonesia, December 8–12, 2021, Proceedings, Part IV 28. pp. 212–223. Springer (2021)
30. Wojke, N., Bewley, A., Paulus, D.: Simple online and realtime tracking with a deep association metric. In: 2017 IEEE international conference on image processing (ICIP). pp. 3645–3649. IEEE (2017)
31. Wu, J., Cao, J., Song, L., Wang, Y., Yang, M., Yuan, J.: Track to detect and segment: An online multi-object tracker. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 12352–12361 (2021)
32. Xu, M., Zhu, Y., Jiang, H., Li, J., Shen, Z., Wang, S., Huang, H., Wang, X., Yang, Q., Zhang, H., et al.: Mitracker: Multi-view integration for visual object tracking. arXiv preprint arXiv:2502.20111 (2025)
33. Xu, Y., Liu, X., Liu, Y., Zhu, S.C.: Multi-view people tracking via hierarchical trajectory composition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4256–4265 (2016)
34. Yang, M., Han, G., Yan, B., Zhang, W., Qi, J., Lu, H., Wang, D.: Hybrid-sort: Weak cues matter for online multi-object tracking. In: Proceedings of the AAAI conference on artificial intelligence. vol. 38, pp. 6504–6512 (2024)
35. Yang, X., Ye, J., Lu, J., Gong, C., Jiang, M., Lin, X., Zhang, W., Tan, X., Li, Y., Ye, X., et al.: Box-grained reranking matching for multi-camera multi-target tracking. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 3096–3106 (2022)
36. Ye, M., Shen, J., Lin, G., Xiang, T., Shao, L., Hoi, S.C.: Deep learning for person reidentification: A survey and outlook. IEEE transactions on pattern analysis and machine intelligence **44**(6), 2872–2893 (2021)
37. You, S., Yao, H., Xu, C.: Multi-target multi-camera tracking with optical-based pose association. IEEE Transactions on Circuits and Systems for Video Technology **31**(8), 3105–3117 (2020)
38. Zeng, F., Dong, B., Zhang, Y., Wang, T., Zhang, X., Wei, Y.: Motr: End-to-end multiple-object tracking with transformer. In: European conference on computer vision. pp. 659–675. Springer (2022)

39. Zhang, X., Yu, H., Qin, Y., Zhou, X., Chan, S.: Video-based multi-camera vehicle tracking via appearance-parsing spatio-temporal trajectory matching network. IEEE Transactions on Circuits and Systems for Video Technology (2024)
40. Zhang, Y., Sun, P., Jiang, Y., Yu, D., Weng, F., Yuan, Z., Luo, P., Liu, W., Wang, X.: Bytetrack: Multi-object tracking by associating every detection box. In: European conference on computer vision. pp. 1–21. Springer (2022)
41. Zhang, Y., Wang, C., Wang, X., Zeng, W., Liu, W.: Fairmot: On the fairness of detection and re-identification in multiple object tracking. International journal of computer vision **129**, 3069–3087 (2021)
42. Zhou, K., Yang, Y., Cavallaro, A., Xiang, T.: Omni-scale feature learning for person re-identification. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 3702–3712 (2019)