ModTrack: Sensor-Agnostic Multi-View Tracking via Identity-Informed PHD Filtering with Covariance Propagation

Aditya Iyer*, Jack Roberts*, and Nora Ayanian

Department of Computer Science, Brown University, Providence, RI, USA
{aditya_iyer, jack_roberts, nora_ayanian}@brown.edu
*Equal contribution.

Abstract. Multi-View Multi-Object Tracking (MV-MOT) aims to localize and maintain consistent identities of objects observed by multiple sensors. This task is challenging, as viewpoint changes and occlusion disrupt identity consistency across views and time. Recent end-to-end approaches address this by jointly learning 2D Bird's Eye View (BEV) representations and identity associations, achieving high tracking accuracy. However, these methods offer no principled uncertainty accounting and remain tightly coupled to their training configuration, limiting generalization across sensor layouts, modalities, or datasets without retraining. We propose ModTrack, a modular MV-MOT system that matches end-to-end performance while providing cross-modal, sensor-agnostic generalization and traceable uncertainty. ModTrack confines learning methods to just the *Detection and Feature Extraction* stage of the MV-MOT pipeline, performing all fusion, association, and tracking with closed-form analytical methods. Our design reduces each sensor's output to calibrated position-covariance pairs $(\mathbf{z}, R)$; cross-view clustering and precision-weighted fusion then yield unified estimates $(\hat{\mathbf{z}}, \hat{R})$ for identity assignment and temporal tracking. A feedback-coupled, identity-informed Gaussian Mixture Probability Hypothesis Density (GM-PHD) filter with HMM motion modes uses these fused estimates to maintain identities under missed detections and heavy occlusion. ModTrack achieves 95.5 IDF1 and 91.4 MOTA on *WildTrack*, surpassing all prior modular methods by over 21 points and rivaling the state-of-the-art end-to-end methods while providing deployment flexibility they cannot. Specifically, the same tracker core transfers unchanged to *MultiviewX* and *RadarScenes*, with only perception-module replacement required to extend to new domains and sensor modalities.

Keywords: Multi-view tracking · Bird's-eye-view · Sensor-agnostic fusion · PHD filter · Uncertainty propagation · Random finite sets

1 Introduction

Multi-view multi-object tracking (MV-MOT) underpins applications from surveillance to collaborative robotics. This task is challenging because objects observed

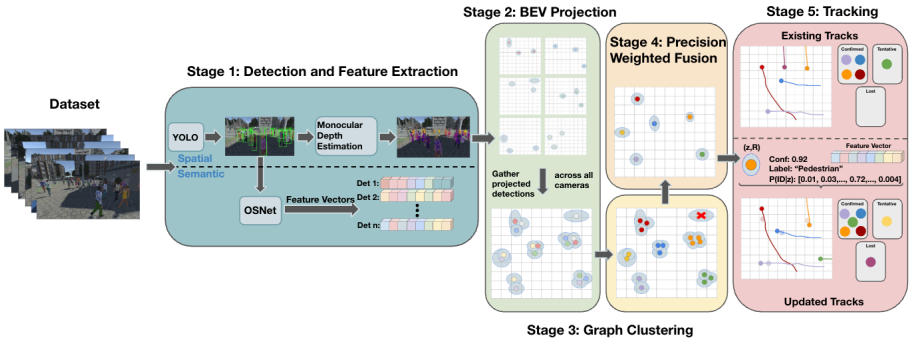


Fig. 1: ModTrack pipeline. Per-camera detections and appearance features (Stage 1) are projected to BEV with Jacobian-propagated covariance (Stage 2), associated via χ^2 graph clustering (Stage 3), and precision-weighted into fused $(\hat{\mathbf{z}}, \hat{R})$ pairs (Stage 4). An identity-informed GM-PHD filter (Stage 5) maintains persistent identities, with track predictions feeding back to guide identity assignment (dashed). Neural modules (blue) are confined to Stage 1; all downstream stages are purely analytical.

from different viewpoints vary significantly in appearance, scale, and visibility, while occlusion and missed detections disrupt identity consistency across both views and time.

Recent end-to-end approaches [22–24, 27] achieve high accuracy by jointly learning Bird's Eye View (BEV) representations and data association, but they provide no interpretable uncertainty estimates and remain tightly coupled to the sensor configuration seen during training, making adaptation to new sensors or modalities difficult without retraining or architectural modification.

Classical random finite set (RFS) filters provide principled multi-target estimation with closed-form covariance propagation. However, the standard PHD filter does not maintain persistent identities [14]. Labeled extensions [26] recover identities but require combinatorial hypothesis enumeration, and existing multi-view RFS trackers do not propagate measurement covariances from perception through cross-view fusion and tracking, limiting uncertainty interpretability and downstream calibration [16].

We propose ModTrack, a modular MV-MOT system that rivals end-to-end accuracy while providing interpretable, traceable uncertainty, by confining learning to just the *Detection and Feature Extraction* stage of the pipeline (Figure 1) and performing all fusion, association, and tracking with closed-form analytical methods. The tracking backend uses the GM-PHD predict-update recursion as an analytical backbone for state estimation and component management, augmented with deterministic identity assignment and lifecycle logic for persistent track continuity [25]. Supplementary Material X validates that these modifications improve identity consistency while preserving state-estimation accuracy. ModTrack adapts to new sensor configurations at inference time seamlessly, avoiding the retraining or engineering accomodations typically required

by end-to-end approaches. It also extends prior RFS trackers by operating directly in calibrated BEV space with metric covariance propagation, narrowing the accuracy gap to learned methods in high-clutter scenes. Across domains, all sensor measurements are represented as calibrated position-covariance pairs $(\mathbf{z}, R)$, so the same analytical core runs unchanged on WildTrack [6] (7 real cameras), MultiviewX [11] (6 synthetic cameras), and RadarScenes [21] (4 automotive radars), requiring only perception-module replacement. On WildTrack, ModTrack achieves 95.5 IDF1 and 91.4 MOTA, surpassing all prior modular methods by over 21 points and rivaling state-of-the-art end-to-end methods.

The contributions of this work are threefold.

1. **Identity-informed GM-PHD filter with closed-loop propagation.** We augment the GM-PHD filter with persistent identity labels, HMM motion modes, and appearance features. Track predictions feed back as association priors into each PHD update, yielding a +8.9 IDF1 gain over spatial information-only association.
2. **Analytical covariance chain.** A fully interpretable uncertainty pipeline propagates monocular depth variance through Jacobian BEV projection, χ^2-calibrated graph clustering, and common-mode-aware precision-weighted fusion, with every covariance traceable to physical sensor characteristics.
3. **Sensor-agnostic deployment flexibility.** The identical tracker core operates unchanged on WildTrack, MultiviewX, and RadarScenes with only perception-module replacement, and we validate robustness to sensor dropout over all $\binom{7}{k}$ camera subsets of WildTrack.

2 Related Work

End-to-End and Learned Multi-View Tracking. Early multi-camera systems used single-view trackers [7] independently per camera and reconciled identities across views using appearance cues and epipolar constraints. Learned re-identification embeddings improved cross-view matching [28,35]. More recent approaches perform joint reasoning directly in BEV. EarlyBird [22] projects per-camera features to a shared ground plane before joint detection and tracking, while later work introduces hierarchical graph reasoning (MCBLT [27]), sparse cross-view transformers (SCFusion [24]), view-aware attention mechanisms [2], and trajectory-level motion–appearance aggregation (MVTrajecter [30], TrackTacular [23]). Related formulations treat multi-view association as a learned graph problem (UMPN [9]) or support variable camera configurations for BEV detection (MIC-BEV [34]). Despite strong performance, these methods typically embed sensor geometry and fusion within learned parameters and do not expose explicit uncertainty estimates.

Random Finite Set Tracking. Random finite set (RFS) formulations provide a principled Bayesian framework for multi-target estimation under clutter and missed detections. The GM-PHD filter [25] offers closed-form multi-target state estimation with linear complexity in the number of mixture components, but does not maintain persistent identity labels. The GLMB filter [26] restores

labeled tracking with full Bayesian data association, yet requires enumerating association hypotheses whose number grows combinatorially. The LMB filter [19] offers a tractable approximation but, like GLMB, carries only statistical labels without appearance integration or feedback from track state to measurement assignment. MV-GLMB-OC [16] extends GLMB to MV-MOT in pedestrian settings, incorporating occlusion reasoning across cameras, but operating in image-space bounding boxes and is unable to propagate metric covariances in a unified world frame. While RFS-based methods provide rigorous uncertainty handling, they have not been combined with BEV-based geometric fusion or scalable identity reasoning as used in contemporary learned MV-MOT systems.

ModTrack bridges learned BEV-based tracking and classical RFS filtering. By combining precision-weighted multi-view fusion with an RFS-style identity-informed GM-PHD backend, it preserves explicit probabilistic state estimates while remaining competitive with modern learned MV-MOT systems.

3 Methodology

3.1 Detection and Feature Extraction

ModTrack's detection and feature extraction stage is sensor-agnostic: any front-end that produces calibrated BEV position-covariance pairs $(\mathbf{z}, R)$ can drive the identical downstream tracking pipeline. We describe two perception modules in this work, a monocular vision front-end and an analytical radar front-end, to validate this claim across modalities. Empirical validation on additional modalities remains future work.

Camera Front-End (WildTrack, MultiviewX). At each time step, YOLOv11x [12] predicts bounding boxes, class labels, and confidence scores for each camera. For each detection, we estimate metric depth by precision-weighting two complementary geometric cues: a ground-plane "footpoint" back-projection and a bounding-box height prior under the pinhole camera model. We clarify "footpoint" back-projection does not require explicit footpoint localization but rather approximates the footpoint as the lowest central point of the bounding box. A learned monocular depth network [17] serves as an outlier check, inflating the fused variance when it confidently disagrees with the geometric estimate. The resulting depth $\hat{d}$ and variance σ_d^2 are passed to the BEV projection stage (Section 3.2); full derivations and parameter choices are provided in Appendices A and F, respectively.

OSNet-AIN [37] then processes each bounding box and computes feature vectors for each detection. These serve as appearance/semantic identity cues used to augment the PHD filter.

Radar Front-End (RadarScenes). Each radar measurement provides range r and azimuth θ with known uncertainties $(\sigma_r, \sigma_\theta)$, where θ is measured from

the forward (longitudinal) sensor axis so that $x_s = r\sin\theta$ and $y_s = r\cos\theta$ in the sensor-local frame. The Jacobian of this polar-to-Cartesian transform,

$$J_{\text{radar}} = \begin{bmatrix} \sin\theta & r\cos\theta \\ \cos\theta & -r\sin\theta \end{bmatrix}, \tag{1}$$

yields sensor-local Cartesian covariance $R_s = J_{\text{radar}}\,\text{diag}(\sigma_r^2, \sigma_\theta^2)\,J_{\text{radar}}^\top$. A rotation R_ψ (ego-vehicle yaw plus sensor mount angle) transforms to the global BEV frame:

$$R_{\text{BEV}} = R_\psi\, R_s\, R_\psi^\top + R_{\text{pose}}, \tag{2}$$

Object class and confidence scores are derived from Doppler velocity and radar cross-section via ad-hoc thresholds, replacing YOLO with zero learned parameters. DBSCAN pre-clustering [10] groups per-point detections into object-level measurements. This produces the same $(\mathbf{z}, R)$ interface without any semantic information.

3.2 Ground-Plane Projection with Jacobian Uncertainty Propagation

This subsection applies to the camera front-end (Section 3.1); the radar front-end (Section 3.1) produces $(\mathbf{z}, R)$ pairs directly and bypasses ground-plane projection.

Given a detection at pixel (u, v) in camera c with estimated depth $\hat{d}$ and variance σ_d^2, we back-project $\mathbf{r}_c = K^{-1}[u, v, 1]^\top$, map the ray in the camera frame into the world frame at the depth $\hat{d}$, and orthogonally project it onto the ground plane:

$$\mathbf{X}_w = R_c\,(\hat{d}\,\mathbf{r}_c) + \mathbf{t}_c, \quad \mathbf{X}_{\text{proj}} = \mathbf{X}_w + \left(d_{\text{plane}} - \mathbf{n}^\top \mathbf{X}_w\right)\mathbf{n}, \tag{3}$$

and define the BEV position $\mathbf{z}_{\text{BEV}} = \mathbf{X}_{\text{proj},\,1:2}$.

We then propagate the scalar depth uncertainty through the projection using the Jacobian with respect to $\hat{d}$, producing a rank-1 BEV covariance term $R_{\text{indep}} = \sigma_d^2\, J\, J^\top$. The full derivation, including the closed-form Jacobian and the role of the horizontal-plane assumption, is given in Supplementary Material (Supp. Material) A.1. We model the BEV measurement covariance as the sum of the depth-propagated term and an isotropic calibration/ego-pose term, $R_{\text{BEV}} = R_{\text{indep}} + R_{\text{pose}}$, where $R_{\text{pose}} = \sigma_{\text{pose}}^2 I_2$ and σ_{pose} is set per dataset (Supp. Material Table 6); Supp. Material A.1 gives the eigenvalue flooring used to avoid overconfident gating. This defines the *sensor-agnostic* $(\mathbf{z}, R)$ *interface* for the camera front-end.

3.3 χ^2 Graph Clustering

In order to group $(\mathbf{z}, R)$ pairs from different sensors associated with the same object we construct an association graph where edge weights encode the χ^2 consistency (survival) probability that two detections originate from the same

target. For detections p_i and p_j from distinct sensors, the Mahalanobis distance under summed covariances is

$$d_{ij}^2(p_i, p_j) = (\mathbf{z}_i - \mathbf{z}_j)^\top (R_i + R_j)^{-1} (\mathbf{z}_i - \mathbf{z}_j). \tag{4}$$

Under the null hypothesis that both observations originate from the same physical location, d_{ij}^2 follows a χ^2 distribution with 2 degrees of freedom. The resulting χ^2 consistency score is $P_{\text{same}}(p_i, p_j) = 1 - F_{\chi_2^2}(d_{ij}^2)$, where $F_{\chi_2^2}$ is the χ^2 CDF. Edges with $P_{\text{same}} > \tau_P$ (corresponding to the $\chi^2_{2,0.99}$=9.21 gate at τ_P=0.01) are retained; a hard Euclidean cutoff τ_{euc} prevents transitive chaining. Following ByteTrack-style [33] cascaded confidence processing, we first cluster high-confidence esimations, then cluster low-confidence estimations in a second pass only for tracks left unmatched after the first pass. Because transitive closure can place multiple estimates from the same sensor into one component, we additionally split each estimate into sensor-unique groups (at most one estimate per sensor, selecting higher-confidence estimates first) before downstream fusion. This is based on the assumption that the object detector does not produce duplicate detections for a single object. We retain only connected components supported by at least two distinct sensors; singleton components are discarded before fusion. This is a pedestrian dataset specific design choice. We observed that the vast majority of single-view clusters were false positives due to the highly overlapping camera coverage making single-view detections rare. For RadarScenes all DBSCAN clusters, including singletons, reach the PHD tracker.

3.4 Precision-Weighted Multi-View Fusion

For each cluster with M sensor measurements $\{(\mathbf{z}_m, R_m)\}_{m=1}^M$, we decompose each covariance into an independent term and a shared common-mode term (calibration/ego-pose), $R_m = R_{\text{indep},m} + R_{\text{pose}}$, and fuse only the independent components (when the shared term is available and consistent across measurements; otherwise R_{pose}=0):

$$P_{\text{indep}}^{-1} = \textstyle\sum_{m=1}^M R_{\text{indep},m}^{-1}, \tag{5}$$

$$\hat{\mathbf{z}}_{\text{fused}} = P_{\text{indep}} \textstyle\sum_{m=1}^M R_{\text{indep},m}^{-1}\, \mathbf{z}_m, \tag{6}$$

$$P_{\text{fused}} = P_{\text{indep}} + R_{\text{pose}}. \tag{7}$$

The fused estimate $(\hat{\mathbf{z}}, \hat{R}) \triangleq (\hat{\mathbf{z}}_{\text{fused}}, P_{\text{fused}})$ remains order-independent; importantly, only the independent perceptual uncertainty shrinks with additional sensors, while the shared calibration/ego-pose term is added back once and therefore does not spuriously vanish. A confidence-weighted appearance feature is pooled across cluster members when semantic information is available:

$$\mathbf{f}_{\mathcal{C}} = \frac{\sum_{m=1}^M \beta_m \cdot \hat{\mathbf{f}}_m}{\left\| \sum_{m=1}^M \beta_m \cdot \hat{\mathbf{f}}_m \right\| + \epsilon}, \tag{8}$$

where $\hat{\mathbf{f}}_m = \mathbf{f}_m / \|\mathbf{f}_m\|$ is the L_2-normalized appearance feature and β_m is the detection confidence.

3.5 Identity-Informed GM-PHD Filter and Closed-Loop Identity Assignment

GM-PHD State Definition and Tracking Loop. We build on the GM-PHD predict-update recursion [25] as the backbone for state estimation and component management, augmenting it with persistent identity labels, appearance features, HMM motion modes, and lifecycle heuristics for practical identity maintenance. These pragmatic extensions prioritize robust identity tracking over strict adherence to the RFS-theoretic PHD recursion. The j-th Gaussian component at time step k is:

$$\mathcal{C}_k^{(j)} = \{w_k^{(j)},\ \mathbf{m}_k^{(j)},\ P_k^{(j)},\ \mathrm{ID}_k^{(j)},\ s_k^{(j)},\ \mathbf{f}_k^{(j)}\}, \tag{9}$$

where w is the PHD weight, $\mathbf{m}{=}[x, y, \dot{x}, \dot{y}]^\top$ is the BEV state, $P{\in}\mathbb{R}^{4\times4}$ is the state covariance, ID is the persistent identity, $s{\in}\{1, 2, 3\}$ is the HMM motion mode, and $\mathbf{f}$ is the pooled appearance feature.

We distinguish between *components* and *tracks*: a component j is a single Gaussian hypothesis (typically corresponding to one motion-mode hypothesis), while a track t denotes a persistent identity label and may be represented by multiple components $\mathcal{J}_k(t) \triangleq \{j : \mathrm{ID}_k^{(j)} = t\}$ within a frame.

At each frame, the method follows the same loop: (i) export one representative GM-PHD component state per track to the fusion stage, (ii) assign each fused cluster both a hard identity and a soft identity distribution using that exported pool, (iii) predict the existing GM-PHD components forward by one time step and append identity-informed birth components, (iv) perform track-level association and Kalman update, and then (v) prune, merge, and apply lifecycle transitions before exposing the next frame's track pool.

Closed-Loop Identity Prior. The fusion stage and the PHD filter communicate in a *closed loop*. Before processing new measurements, the filter exports one representative component per visible track, $\mathcal{T}_k = \{(t, \mathbf{m}_k^{(j_t^\star)}, P_k^{(j_t^\star)}, s_k^{(j_t^\star)}, \mathbf{f}_k^{(j_t^\star)})\}_{t=1}^{T_k}$, where $j_t^\star = \arg\max_{j\in\mathcal{J}_k(t)} w_k^{(j)}$. For each fused estimate $(\hat{\mathbf{z}}_i, \hat{R}_i)$ and representative component t, a gated Mahalanobis cost drives Hungarian matching [13]:

$$\mathrm{Cost}_{i,t} = \begin{cases} d_{i,t}^2, & d_{i,t}^2 < \tau_{\mathrm{gate}} \\ +\infty, & \text{otherwise} \end{cases}, \mathrm{Cost}_{\mathrm{new}} = \tau_{\mathrm{new}} \tag{10}$$

Here $d_{i,t}^2$ is computed as in Equation (4) with $p_i \triangleq (\hat{\mathbf{z}}_i, \hat{R}_i)$ and $p_t \triangleq (\mathbf{z}_t, R_t)$, where $(\mathbf{z}_t, R_t)$ is the predicted BEV position and covariance from the $\mathbf{m}, P$ entries of the representative component for track t.

A heuristic spatial identity score provides a soft preference toward nearby tracks. Rather than a full Bayesian posterior, we apply the same χ_2^2 gate as Equation (10): the soft distribution is formed only over tracks with $d_{i,t}^2 < \tau_{\mathrm{gate}}$ and then renormalized:

$$g_{i,t} = \exp\left(-\frac{d_{i,t}^2}{2\,\sigma_{\mathrm{spatial}}^2}\right), \quad \pi_{i,t}^{\mathrm{geo}} = \frac{\exp(g_{i,t}/\tau_{\mathrm{geo}})}{\sum_{t'}\exp(g_{i,t'}/\tau_{\mathrm{geo}})}. \tag{11}$$

where $\sigma_{\text{spatial}} = 1.0$ is a fixed spatial bandwidth and τ_{geo} controls the sharpness of the softmax. $\pi_{i,t}^{\text{geo}}$ denotes a normalized spatial proximity score for assigning measurement i to track t. The hard Hungarian identity initializes the candidate persistent label for the fused cluster, while the soft distribution re-enters the tracker through the association cost in Equation (17). The Hungarian-assigned identity receives nonzero mass before renormalization, to ensure each measurement provides a valid distribution.

Prediction and Identity-Informed Birth. Three motion modes share the linear state-transition structure $\mathbf{x}_{k+1} = F_s\,\mathbf{x}_k + \mathbf{w}_k$ with $\mathbf{w}_k \sim \mathcal{N}(0, Q_s)$: *stationary* ($s$=1) damps velocity, *constant velocity* (s=2) uses the standard CV transition, and *maneuvering* (s=3) uses the same transition with elevated process noise $Q_3 \gg Q_2$ [5]. Mode transitions follow a class-specific Markov chain $\Pi_{ij}^{(c)} = P(s_{k+1}{=}j \mid s_k{=}i)$.

Prediction and birth are executed as a single stage. We first propagate the surviving components through the HMM motion model,

$$w_{k|k-1}^{(j,s')} = p_S^{(c)} \cdot \Pi^{(c)}_{s_k^{(j)},s'} \cdot w_k^{(j)}, \tag{12}$$

$$\mathbf{m}_{k|k-1}^{(j,s')} = F_{s'}\,\mathbf{m}_k^{(j)}, \tag{13}$$

$$P_{k|k-1}^{(j,s')} = F_{s'}\,P_k^{(j)}\,F_{s'}^{\top} + Q_{s'}^{(c)}. \tag{14}$$

and then append identity-aware birth components from the current fused clusters whose hard identity assignment does not match an existing predicted track and whose confidence exceeds the birth threshold τ_{birth},

$$\mathbf{m}_{\text{birth}}^{(s)} = \begin{bmatrix} \hat{\mathbf{z}}_{\mathcal{C}} \\ \mathbf{0} \end{bmatrix},\; P_{\text{birth}}^{(s)} = \begin{bmatrix} \hat{R}_{\mathcal{C}} & 0 \\ 0 & (\sigma_v^{(c)})^2 I_2 \end{bmatrix},\; w_{\text{birth}}^{(s)} = \beta_{\max} \cdot \bar{\Pi}_{1,s}, \tag{15}$$

where $\sigma_v^{(c)}$ is the class-specific birth velocity prior, $\bar{\Pi}_{1,:}$ is the stationary-mode prior, and $\beta_{\max}$ is the maximum detection confidence in the cluster. Multiple mode hypotheses are retained through association and update, then collapsed during post-update component management.

Track-Level Association and Update. At this point, the predicted pool contains both propagated hypotheses from existing tracks and newly initialized tentative birth hypotheses. These predicted components are associated against the current fused measurements from the same frame. Association is performed at the *track level* rather than the component level so that multiple HMM modes of the same identity do not compete for different measurements within the same frame.

Cost construction. For each track component j and measurement i, the base cost is the Gaussian negative log-likelihood:

$$c_{ij}^{\text{base}} = \tfrac{1}{2}\big(d_{ij}^2 + \log|S_k^{(j,i)}|\big) - \log p_D^{(c)}, \tag{16}$$

where $d_{ij}^2 = (\mathbf{z}_k^{(i)} - H\mathbf{m}_{k|k-1}^{(j)})^\top (S_k^{(j,i)})^{-1} (\mathbf{z}_k^{(i)} - H\mathbf{m}_{k|k-1}^{(j)})$ and $S_k^{(j,i)} = HP_{k|k-1}^{(j)} H^\top + R_k^{(i)}$ is the innovation covariance (H=$[I_2 \mid 0]$). We omit the constant $\log(2\pi)$ term, since it is a fixed offset on all match costs but would otherwise distort match-vs-miss comparisons. Pairs failing the same χ_2^2 gate as Section 3.3 are treated as infeasible (c_{ij}=$+\infty$). Three additional terms modify the cost:

$$c_{ij} = c_{ij}^{\text{base}} - \lambda_{\text{assoc}} \cdot \pi_{i,t}^{\text{geo}} - \mu_{\text{sem}} \cdot \cos(\mathbf{f}_k^{(j)}, \mathbf{f}_k^{(i)}) + \psi_{\text{turn}}(j, i), \tag{17}$$

where λ_{assoc} is the identity boost, μ_{sem} is the appearance similarity boost, and ψ_{turn} is a speed-adaptive turn penalty.

Track-level grouping. Components are grouped by identity; for each track t and measurement i, the grouped cost is $c_{t,i} = \min_{j \in \mathcal{J}_k(t)} c_{ij}$. We then solve a gated track-level Hungarian assignment over the confirmed-track subset, while tentative/lost hypotheses (including newly birthed tracks) are updated from the remaining measurements in a second pass; the full augmented assignment matrix and second-pass details are given in Supp. Material B.

Given the resulting assignments, matched components receive a Kalman update:

$$K_k^{(j,i)} = P_{k|k-1}^{(j)} H^\top \left(S_k^{(j,i)}\right)^{-1}, \tag{18}$$

$$\mathbf{m}_k^{(j)} = \mathbf{m}_{k|k-1}^{(j)} + K_k^{(j,i)} (\mathbf{z}_k^{(i)} - H\mathbf{m}_{k|k-1}^{(j)}), \tag{19}$$

$$P_k^{(j)} = (I - K_k^{(j,i)} H)\, P_{k|k-1}^{(j)}\, (I - K_k^{(j,i)} H)^\top + K_k^{(j,i)}\, R_k^{(i)}\, K_k^{(j,i)\top}. \tag{20}$$

These matched components also receive a weight boost $w \leftarrow \min(w + w_{\text{boost}}, 1)$; unmatched components receive a miss penalty $w \leftarrow w \cdot (1 - p_D)$ and are not updated.

Component Management and Track Lifecycle. After the PHD update, low-weight components are pruned, nearby components with the same identity may be merged, and only the dominant motion-mode component is retained per identity, yielding one representative component per identity for the next frame.

Tracks evolve through Tentative, Confirmed, and Lost states to stabilize persistent IDs and support re-identification after brief occlusion. Exact merge conditions, lifecycle thresholds, deletion rules, and the detailed re-identification score are provided in Supp. Material B. After these transitions, confirmed tracks are extracted and the representative components form the track pool used for the next frame.

4 Results and Analysis

4.1 Datasets

WildTrack [6]: 7 synchronized calibrated HD cameras (1920×1080, 2 fps) covering a 12×36 m outdoor area with ∼20 pedestrians per frame; 400 frames total,

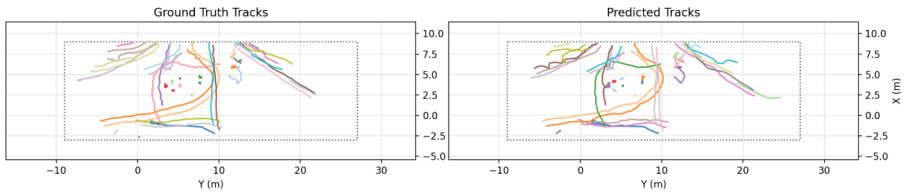


Fig. 2: Predicted BEV tracks on WildTrack (colored by identity). Ground truth (left) vs. ModTrack (right).

360/40 train/test split (90%/10%).
MultiviewX [11]: 6 synthetic cameras (1920×1080, 2 fps) covering a 16×25 m area with ∼40 pedestrians per frame; same split.
RadarScenes [21]: 158 driving sequences from four 77 GHz radars (72 868 frames). Each point carries range, azimuth, Doppler velocity, and radar cross-section.

4.2 Evaluation Metrics

We report MOT metrics [4] (MOTA, MOTP) and identity-aware IDF1 [20] with a 1.0 m positive assignment threshold for the camera datasets. We further report GOSPA [18] (p=2, c=1, α=2), which jointly penalizes localization error, missed targets, and false alarms. For RadarScenes we report LSTQ [3].

4.3 Implementation Details

YOLOv11x [12] provides monocular object detection; the Lift depth network [17] (EfficientNet-B0, 41 bins) provides an outlier check for fused geometric depth estimates; OSNet-AIN [37] trained with triplet loss via torchreid [36] provides appearance features. All perception modules are trained on the target dataset using the same 90%/10% train/test split (e.g., 360/40 frames for WildTrack); Supp. Material E summarizes the exact data preparation, losses, and optimizer settings used for the YOLO, Lift, and OSNet-AIN front-ends. All clustering, tracking, and calibration hyperparameters are detailed in Tables 6 and 7.

4.4 Camera Tracking Results

Figure 2 shows qualitative tracking results of ModTrack on the test set of WildTrack. Table 1 compares ModTrack against end-to-end and modular baselines on WildTrack and MultiviewX. Among modular methods, ModTrack surpasses all prior work by +21.2 IDF1 and +21.7 MOTA on WildTrack. GOSPA scores of 1.20 and 1.83 reaffirms strong tracking performance.

The best end-to-end method [30] exceeds ModTrack by 1.0 IDF1 and 2.9 MOTA on WildTrack; in exchange, ModTrack provides sensor-agnostic deployment without retraining and physically traceable uncertainty. As for MultiviewX, we investigated the larger gap between ModTrack and other methods (86.3 vs.

Table 1: Comparison of tracking performance on WildTrack and MultiviewX datasets.

	WildTrack				**MultiviewX**			
Method	IDF1	MOTA	MOTP	GOSPA↓	IDF1	MOTA	MOTP	GOSPA↓
End-to-end (fixed config.)								
EarlyBird [22]	92.3	89.5	86.6	–	82.4	88.4	86.2	–
Ali et al. [1]	92.0	87.4	–	–	–	–	–	–
MCBLT [27]	95.6	92.6	**94.3**	–	–	–	–	–
TrackTacular [23]	95.3	91.8	85.4	–	85.6	92.4	80.1	–
SCFusion [24]	95.9	92.4	86.3	–	85.0	92.5	85.4	–
Alturki et al. [2]	96.1	92.7	88.8	–	85.7	91.3	86.9	–
MVTrajecter [30]	**96.5**	**94.3**	93.0	–	**85.8**	**92.8**	**95.0**	–
Learned (Non-Interpretable)								
DMCT [31]	77.8	72.8	79.1	–	–	–	–	–
ReST [8]	86.7	84.9	84.1	–	–	–	–	–
UMPN [9]	96.3	93.9	86.9	–	–	–	–	–
Modular (flexible config.)								
KSP-DO [6]	73.2	69.6	61.5	–	–	–	–	–
MV-GLMB-OC [16]	74.3	69.7	73.2	–	–	–	–	–
ModTrack (Ours)	**95.5**	**91.4**	**87.2**	**1.20**	**86.2**	**86.3**	**84.0**	**1.83**
ModTrack (Oracle det.)	97.1	94.7	94.0	0.68	88.2	93.6	88.8	1.25

92.8 MOTA) with a sweep over $\tau_{\text{euc}} \in \{0.10, 0.25, 0.50, 0.75\}$ which revealed a structural tradeoff: increasing the gate reduces FN clusters (682 to 56) but increases merge errors (225 to 481), with no threshold resolving both. This suggests a structural limitation rather than a tuning issue that emerged with MultiviewX due to its high 2× higher pedestrian density. However, ModTrack leads all methods in IDF1 on MultiviewX, the primary tracking consistency metric, despite the higher crowd density.

Oracle detection analysis. To isolate the quality of the tracker backend from upstream perception noise, we replace the output of the monocular object detector with ground-truth bounding boxes; we refer to this configuration as *oracle detection*. Under oracle detection, ModTrack achieves 97.1/94.7/94.0 (IDF1/MOTA/MOTP) on WildTrack, which would be SOTA. On MultiviewX, oracle performance reaches 88.2/93.6 (IDF1/MOTA), closing the MOTA gap with end-to-end methods entirely (93.6 vs. 92.8). These results establish that the residual gap under standard operation originates entirely in the upstream perception modules. Because ModTrack's perception modules are replaceable, improvements in detection or depth estimation directly improve tracking without modifying the tracker.

4.5 Ablation Studies

Table 2 compares ModTrack's three matching modes on the two camera datasets: geometric-only association (*Spatial*), appearance-only association (*Semantic*),

Table 2: Ablation of matching modes on WildTrack and MultiviewX. Joint mode uses spatial clustering and reserves semantic features for the tracking stage.

	WildTrack				**MultiviewX**			
Mode	IDF1	MOTA	MOTP	GOSPA	IDF1	MOTA	MOTP	GOSPA
Spatial only	86.6	89.5	82.4	1.22	84.2	84.9	83.7	1.86
Semantic only	49.6	41.8	79.6	2.60	65.1	56.7	79.6	3.00
Joint (full)	**95.5**	**91.4**	**87.2**	**1.20**	**86.2**	**86.3**	**84.0**	**1.83**

Table 3: Sensor dropout robustness (WildTrack, joint mode). Mean±std over all $\binom{7}{k}$ subsets. GOSPA (p=2, c=1 m, α=2): lower is better.

	Cameras available					
	7	6	5	4	3	2
IDF1	95.5	91.7±2.9	85.1±5.3	74.5±9.7	57.8±15.7	31.6±19.5
MOTA	91.4	87.7±2.0	81.1±4.7	70.2±9.0	51.7±16.2	25.5±18.3
MOTP	87.2	85.9±0.9	83.8±2.9	82.4±3.0	80.7±4.2	73.2±18.2
GOSPA	1.2	1.3±0.1	1.5±0.2	1.9±0.3	2.3±0.4	2.9±0.4

and the combined *Joint* configuration. Spatial-only achieves strong MOTA (89.5), confirming that calibrated geometric reasoning alone provides robust association. The +8.9 IDF1 improvement from joint mode on WildTrack (+2.0 on MultiviewX) comes entirely from better identity preservation through appearance-informed tracking. Supp. Material Section H ablates over other key design choices.

4.6 Deployment Robustness: Sensor Dropout and Addition

Table 3 quantifies ModTrack's sensor dropout robustness by exhaustively evaluating all $\binom{7}{k}$ camera subsets on WildTrack (120 configurations).

In real-world scenarios where sensor dropout or addition can happen instantaneously, our system adapts seamlessly and effectively, while end-to-end methods in Table 1 would need additional engineering accommodations or retraining. With only 5 of 7 cameras, ModTrack achieves 85.1 IDF1 and 81.1 MOTA, surpassing MV-GLMB-OC (74.3/69.7) *with all 7 cameras* by over 10 points. Additionally, MOTP remains above 80 with just 3 cameras, confirming graceful degradation across all metrics as camera coverage reduces.

The increasing standard deviation for IDF1 at lower camera counts (±2.9 at k=6 vs. ±15.7 at k=3) quantifies which camera placements are critical, directly informing redundancy planning.

Figure 3 visualizes how each additional sensor contributes measurable information through precision-weighted fusion, with shrinking covariance ellipses quantifying the reduction in localization uncertainty that underlies the performance trends in Table 3.

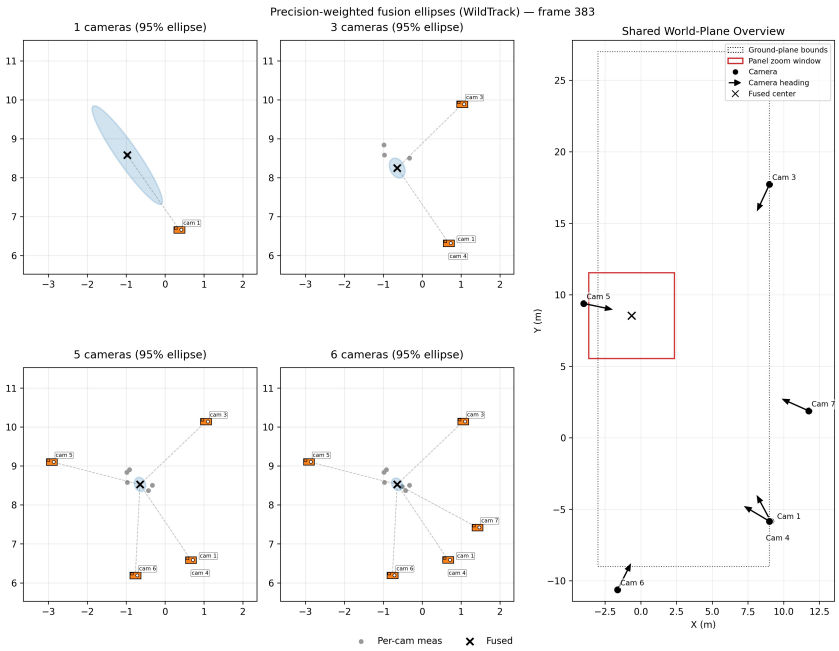


Fig. 3: Precision-weighted fusion on WildTrack (frame 383). Each panel shows BEV covariance ellipses (95% confidence) as cameras are incrementally added. The rightmost panel shows the 6-camera shared world-plane.

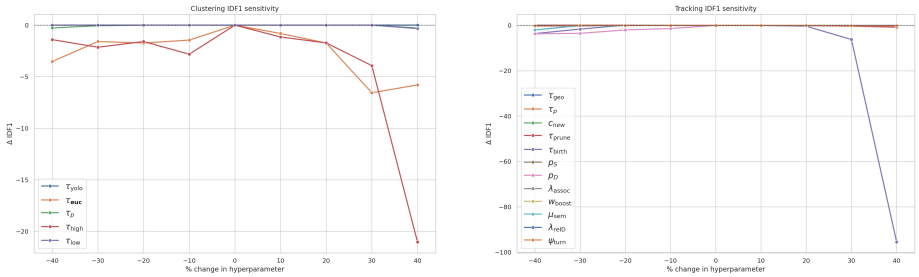
Table 4: Comparison on RadarScenes dataset.

Approach	LSTQ	S_{cls}	S_{assoc}	GOSPA
MOT [29]	42.4	19.4	92.7	–
CA-Net [15]	34.8	13.0	92.7	–
Radar Tracker [32]	66.8	92.7	48.2	–
ModTrack + Doppler perception	42.1	70.0	26.9	2.50
ModTrack + Oracle perception	76.9	99.1	60.9	1.09

4.7 Radar Tracking Results

The $(\mathbf{z}, R)$ interface extends naturally to any sensor producing a calibrated BEV position and covariance. We validate this by applying the identical tracker core to automotive radar on RadarScenes [21], replacing only the perception front-end (Section 3.1).

Table 4 shows ModTrack achieves 42.1 LSTQ with a zero-training Doppler perception front-end. Providing the ModTrack with Oracle ground-truth labels raises LSTQ to 76.9 and surpasses the best performing method reported in the literature (Radar Tracker) by 10.1 points. The entire gap is in perception: a stronger front-end translates directly into better tracking without any changes to the tracker.



(a) Clustering hyperparameters. **(b)** Continuous tracking hyperparameters.

Fig. 4: IDF1 sensitivity analysis (WildTrack, joint mode). Each parameter is swept individually; all others held at defaults. Extended results are in Supp. Material D.

4.8 Runtime Analysis

When tested on WildTrack, the tracking backend accounts for less than 20% of per-frame latency (103 ms vs. 444 ms for neural perception on an NVIDIA RTX 3090 GPU); a full breakdown is provided in Supp. Material C.

4.9 Sensitivity Analysis

To assess the robustness of ModTrack to hyperparameter choices, we perform a one-at-a-time sensitivity analysis on the WildTrack benchmark.

Only three parameters, τ_{euc} (Euclidean gating), τ_{high} (ByteTrack high confidence threshold), and τ_{birth} (minimum cluster confidence for track birth), meaningfully affect IDF1 (Figure 4); all three directly control the FP/FN tradeoff, and all remaining parameters vary by at most 2–3 IDF1 points.

5 Conclusion

ModTrack demonstrates that restricting neural networks to perception while performing downstream reasoning with closed-form analytical methods can match end-to-end performance while providing deployment flexibility, auditable uncertainty, and sensor agnosticism. Across WildTrack, MultiviewX, and RadarScenes, the same identity-informed GM-PHD tracking core operates unchanged, while the closed-loop identity mechanism and sensor-agnostic $(\mathbf{z}, R)$ covariance chain enable robust operation under sensor dropout. Oracle detection experiments confirm that the tracker backend is not the limiting factor: with improved perception, ModTrack's tracking performance scales directly toward SOTA. More broadly, the explicit propagation of calibrated uncertainty throughout the pipeline enables future sensor placement and selection strategies that maximize information gain in detection and tracking systems.

Acknowledgements

This work was supported by the Office of Naval Research (ONR) Award No. N00014-24-1-2662

References

1. Ali, R., Mehltretter, M., Heipke, C.: Integrating Viewing Direction and Image Features for Robust Multi-View Multi-Object 3D Pedestrian Tracking. ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences **X-G-2025**, 47–55 (Jul 2025)
2. Alturki, R., Hilton, A., Guillemaut, J.Y.: Attention-aware multi-view pedestrian tracking (2025), arXiv:2504.03047
3. Aygun, M., Osep, A., Weber, M., Maximov, M., Stachniss, C., Behley, J., Leal-Taixe, L.: 4d panoptic lidar segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5527–5537 (June 2021)
4. Bernardin, K., Stiefelhagen, R.: Evaluating Multiple Object Tracking Performance: The CLEAR MOT Metrics. EURASIP Journal on Image and Video Processing **2008**(1), 246309 (May 2008)
5. Blom, H., Bar-Shalom, Y.: The interacting multiple model algorithm for systems with Markovian switching coefficients. IEEE Transactions on Automatic Control **33**(8), 780–783 (Aug 1988)
6. Chavdarova, T., Baqué, P., Bouquet, S., Maksai, A., Jose, C., Bagautdinov, T., Lettry, L., Fua, P., Van Gool, L., Fleuret, F.: Wildtrack: A multi-camera hd dataset for dense unscripted pedestrian detection. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5030–5039 (2018)
7. Chen, L., Ai, H., Zhuang, Z., Shang, C.: Real-time multiple people tracking with deeply learned candidate selection and person re-identification. In: 2018 IEEE International Conference on Multimedia and Expo (ICME). pp. 1–6 (2018)
8. Cheng, C.C., Qiu, M.X., Chiang, C.K., Lai, S.H.: ReST: A Reconfigurable Spatial-Temporal Graph Model for Multi-Camera Multi-Object Tracking. In: 2023 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 10017–10026. IEEE, Paris, France (Oct 2023)
9. Engilberge, M., Vrkic, I., Grosche, F.W., Pilet, J., Turetken, E., Fua, P.: One Graph to Track Them All: Dynamic GNNs for Single- and Multi-View Tracking (Dec 2025), arXiv:2507.08494
10. Ester, M., Kriegel, H.P., Sander, J., Xu, X.: A density-based algorithm for discovering clusters in large spatial databases with noise. In: Proceedings of the Second International Conference on Knowledge Discovery and Data Mining. p. 226–231. KDD'96, AAAI Press (1996)
11. Hou, Y., Zheng, L., Gould, S.: Multiview detection with feature perspective transformation. In: ECCV (2020)
12. Jocher, G., Qiu, J.: Ultralytics yolo11 (2024), `https://github.com/ultralytics/ultralytics`, accessed: June 24, 2026
13. Kuhn, H.W.: The hungarian method for the assignment problem. Naval Research Logistics Quarterly **2**(1-2), 83–97 (1955)
14. Mahler, R.: Multitarget bayes filtering via first-order multitarget moments. IEEE Transactions on Aerospace and Electronic Systems **39**(4), 1152–1178 (2003)

15. Marcuzzi, R., Nunes, L., Wiesmann, L., Vizzo, I., Behley, J., Stachniss, C.: Contrastive instance association for 4d panoptic segmentation using sequences of 3d lidar scans. IEEE Robotics and Automation Letters **7**(2), 1550–1557 (2022)
16. Ong, J., Vo, B.T., Vo, B.N., Kim, D.Y., Nordholm, S.: A Bayesian Filter for Multi-View 3D Multi-Object Tracking With Occlusion Handling. IEEE Transactions on Pattern Analysis and Machine Intelligence **44**(5), 2246–2263 (May 2022)
17. Philion, J., Fidler, S.: Lift, Splat, Shoot: Encoding Images from Arbitrary Camera Rigs by Implicitly Unprojecting to 3D. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.M. (eds.) Computer Vision – ECCV 2020, vol. 12359, pp. 194–210. Springer International Publishing, Cham (2020), series Title: Lecture Notes in Computer Science
18. Rahmathullah, A.S., García-Fernández, Á.F., Svensson, L.: Generalized optimal sub-pattern assignment metric. In: 2017 20th International Conference on Information Fusion (Fusion). pp. 1–8 (Jul 2017)
19. Reuter, S., Vo, B.T., Vo, B.N., Dietmayer, K.: The Labeled Multi-Bernoulli Filter. IEEE Transactions on Signal Processing **62**(12), 3246–3260 (Jun 2014)
20. Ristani, E., Solera, F., Zou, R., Cucchiara, R., Tomasi, C.: Performance Measures and a Data Set for Multi-target, Multi-camera Tracking. In: Hua, G., Jégou, H. (eds.) Computer Vision – ECCV 2016 Workshops. pp. 17–35. Springer International Publishing, Cham (2016)
21. Schumann, O., Hahn, M., Scheiner, N., Weishaupt, F., Tilly, J.F., Dickmann, J., Wöhler, C.: RadarScenes: A Real-World Radar Point Cloud Data Set for Automotive Applications. In: 2021 IEEE 24th International Conference on Information Fusion (FUSION). pp. 1–8 (Nov 2021)
22. Teepe, T., Wolters, P., Gilg, J., Herzog, F., Rigoll, G.: EarlyBird: Early-Fusion for Multi-View Tracking in the Bird's Eye View. In: 2024 IEEE/CVF Winter Conference on Applications of Computer Vision Workshops (WACVW). pp. 102–111 (Jan 2024)
23. Teepe, T., Wolters, P., Gilg, J., Herzog, F., Rigoll, G.: Lifting multi-view detection and tracking to the bird's eye view. In: 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). pp. 667–676 (2024)
24. Toida, K., Sakai, T., Kato, N., Yokota, K., Nakamura, T., Hotta, K.: Sparse BEV Fusion with Self-View Consistency for Multi-View Detection and Tracking (Sep 2025), arXiv:2509.08421
25. Vo, B.N., Ma, W.K.: The Gaussian Mixture Probability Hypothesis Density Filter. IEEE Transactions on Signal Processing **54**(11), 4091–4104 (Nov 2006)
26. Vo, B.N., Vo, B.T., Phung, D.: Labeled Random Finite Sets and the Bayes Multi-Target Tracking Filter. IEEE Transactions on Signal Processing **62**(24), 6554–6567 (Dec 2014)
27. Wang, Y., Meinhardt, T., Cetintas, O., Yang, C.Y., Pusegaonkar, S.S., Missaoui, B., Biswas, S., Tang, Z., Leal-Taixé, L.: Mcblt: Multi-camera multi-object 3d tracking in long videos. In: 2025 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW). pp. 5304–5313 (2025)
28. Wei, L., Zhang, S., Gao, W., Tian, Q.: Person Transfer GAN to Bridge Domain Gap for Person Re-identification. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 79–88. IEEE, Salt Lake City, UT, USA (Jun 2018)
29. Weng, X., Wang, J., Held, D., Kitani, K.: 3d multi-object tracking: A baseline and new evaluation metrics. In: 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 10359–10366 (2020)

30. Yamane, T., Masumura, R., Suzuki, S., Orihashi, S.: Mvtrajecter: Multi-view pedestrian tracking with trajectory motion cost and trajectory appearance cost. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 13270–13280 (October 2025)
31. You, Q., Jiang, H.: Real-time 3D Deep Multi-Camera Tracking (Mar 2020), arXiv:2003.11753
32. Zeller, M., Herraez, D.C., Behley, J., Heidingsfeld, M., Stachniss, C.: Radar tracker: Moving instance tracking in sparse and noisy radar point clouds. In: 2024 IEEE International Conference on Robotics and Automation (ICRA). pp. 16170–16177 (2024)
33. 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: Avidan, S., Brostow, G., Cissé, M., Farinella, G.M., Hassner, T. (eds.) Computer Vision – ECCV 2022, vol. 13682, pp. 1–21. Springer Nature Switzerland, Cham (2022), series Title: Lecture Notes in Computer Science
34. Zhang, Y., Zheng, Z., Liu, J., Huang, Z., Zhou, Z., Meng, Z., Cai, T., Ma, J.: MIC-BEV: Multi-Infrastructure Camera Bird's-Eye-View Transformer with Relation-Aware Fusion for 3D Object Detection (Oct 2025), arXiv:2510.24688
35. Zheng, L., Shen, L., Tian, L., Wang, S., Wang, J., Tian, Q.: Scalable Person Re-identification: A Benchmark. In: 2015 IEEE International Conference on Computer Vision (ICCV). pp. 1116–1124. IEEE, Santiago, Chile (Dec 2015)
36. Zhou, K., Xiang, T.: Torchreid: A Library for Deep Learning Person Re-Identification in Pytorch (Oct 2019), arXiv:1910.10093
37. Zhou, K., Yang, Y., Cavallaro, A., Xiang, T.: Learning Generalisable Omni-Scale Representations for Person Re-Identification. IEEE Transactions on Pattern Analysis and Machine Intelligence **44**(9), 5056–5069 (Sep 2022)