University of Milano - Bicocca
Department of Informatics, Systems and Communication
Master's degree in Data Science

Design and Evaluation of Multi-Camera Multi-Object Tracking Pipelines in Calibrated Environments

Supervisor: Dr. Marco Buzzelli
Co-supervisor: Prof. Simone Bianco

Master's degree thesis by:
Julius Maliwat
student ID number 864520

Academic Year 2024–2025

Alla mia famiglia

Acknowledgements

This thesis holds particular significance: it marks the conclusion of my academic journey and embodies the commitment I made and kept to complete my master's degree while working full-time. For two years, that meant early mornings, late evenings, and most weekends devoted to study. I did my best to make every hour count, and there was rarely a moment to pause. Reaching the end of this thesis finally lets me stop, look back, and thank the many people who made this journey possible. I could not have done it alone.

To Professor Buzzelli, for his steady guidance throughout the development and writing of this thesis.

To my father, who shaped the person I am and stood by me in every major decision, especially in moments of uncertainty. He always ensured I had the right environment and support to study at my best. To my mother, whose unconditional support and high expectations have always pushed me to grow. To my siblings, who inspire me every day, often without even realising it, and remind me to be the best example I can.

To Nicole, who has been by my side through every high and low. You listened to the post-exam rants, you were there when the thesis felt stuck, you waited for me after late-evening study sessions, and you also shared the successes and the joy of this journey. Your presence and constant support made this path infinitely lighter and truly possible.

Thank you all.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Many operational systems, such as those involved in urban mobility, public space safety, sports analytics, and autonomous platforms, depend on measuring how objects move, and on recognizing identity stability over time. Within the context of computer vision, tracking serves this need: it estimates object trajectories while maintaining consistent identities so that automated counts, flows, and behavioral measures can be trusted.

Single-camera tracking degrades in crowded scenes: occlusions remove evidence, variations in perspective and scale changes confound motion cues, and intermittent misses or short gaps often cause identity switches, i.e., the system mistakenly carries over the previous identity to a different object. Multi-camera multi-object tracking (MC-MOT) addresses these issues by coordinating calibrated, time-synchronized cameras [11, 27]. With overlapping views, what is unseen in one image is often visible in another, and geometry can reconcile observations across views while keeping a shared notion of position. Multi-camera tracking, however, introduces its own demands. Evidence is viewpoint-dependent and must be fused into a coherent representation; identities must remain stable not only over time but also across views with different visibility and appearance; small calibration or timing imperfections can misalign otherwise compatible signals; and overlap creates competing hypotheses that must be resolved across views. In practice, these requirements must be met under low-latency constraints [53].

Historically, many systems for MC-MOT followed a late-fusion pattern: each camera produces detections or short tracklets and cross-view identity is resolved downstream by geometric checks or global association. While modular and often low-latency, late fusion tends to duplicate hypotheses across views, propagate per-view noise, and struggle through extended occlusions [14, 6, 30]. More recent work moves upstream the fusion of information coming from multiple cameras: in the bird's-eye view (BEV), per-view evidence is lifted to a shared metric frame on the ground plane and detection is decided there. Resolving spatial consistency in BEV reduces cross-view conflicts and concentrates the remaining difficulty

on the temporal axis—keeping identities intact over time [53, 2, 57].

This thesis adopts the BEV frame and focuses on pedestrian tracking in dense, calibrated scenes with overlapping views. Given cameras observing an approximately planar ground, the task is to estimate, at each video frame, ground-plane person states (position and identity in BEV) and to maintain those identities over time online, using only past and current video frames with no look-ahead, and with low latency.

Within the BEV frame, identity stability is shaped by a few practical tensions. A first tension concerns where to compute appearance features. BEV is the natural space for localization, geometric gating, and proximity tests, but the projection introduces position-dependent distortions [53, 2]. Learning appearance directly in BEV keeps everything in one space, whereas constructing appearance in image space from multi-view crops preserves visual detail and detector-agnosticity, at the cost of additional processing and runtime overhead.

A second tension concerns maintaining identity continuity when operating online. In BEV, a lightweight state-space model with prediction and geometric gating efficiently narrows candidates and keeps latency low. However, long occlusions and returns in dense areas remain challenging without look-ahead; typical failures include fragmentation and identity switches when a target reappears close to visually similar neighbors [53, 2, 54].

A third tension is portability with respect to the detector. Using features computed inside the detector to represent identity ties the tracker to a specific backbone and training domain, whereas computing appearance with a separate encoder from multi-view crops keeps the tracker detector-agnostic and allows independent tuning of the re-identification component, at the cost of additional runtime overhead [57].

Early fusion in the bird's-eye view stabilizes cross-view geometry, yet a gap remains: to our knowledge, existing MC-MOT methods operating in BEV do not yet provide a fully online, detector-agnostic tracker under realistic latency budgets [53, 54, 57, 2].

Such a tracker should meet two requirements. First, causality: decisions are taken frame by frame with lightweight management of the state of tracked objects and no near-/offline windowing. Second, detector independence for identity cues: identity association must be detector-agnostic, so the tracker can use BEV detections from different detectors without retraining detector-tied components. We consider this gap within the calibrated, time-synchronized, overlapping multi-camera setting over an approximately planar ground with pedestrians, and under online, low-latency constraints.

Goals. Two goals guide this work. The first is to define a BEV tracker that fills the gap stated above. It must be fully online and detector-agnostic—causal frame-by-frame decisions with lightweight state management and no near-/offline windowing, and no reliance on detector-specific identity components or the backbone, so that it can consume BEV

detections from different detectors. Once defined, the tracker is evaluated against representative prior work to position its identity stability; with a fixed detector, we also quantify how effectively detections become identities that remain consistent over time, separating the contribution of association from that of the detector.

The second goal is to understand where and why the tracker makes identity errors in the online regime and to use that diagnosis to guide improvements. The analysis relates switches and dropouts to observable conditions, maps recurring patterns, derives operational guidance, and verifies that incorporating these insights yields measurable benefits.

Contributions. This thesis delivers two elements: (i) a bird's-eye-view tracker that is fully online and detector-agnostic, combining predictive motion with geometric gating in BEV and appearance built from multi-view image crops via back-projection; (ii) an occlusion-aware visibility score used as an operational diagnostic tool to interpret identity failures.

First, the BEV tracker is causal with lightweight state management and makes no use of detector-specific identity components or the backbone, so it can consume BEV detections from different detectors. Association couples prediction and geometric gating in BEV with appearance computed in image space from multi-view crops via back-projection, with a geometric fallback when appearance is unreliable. The practical value is portability across early-fusion detectors and a clean separation between motion and appearance that makes control and diagnosis straightforward. In calibrated, approximately planar scenes with overlapping views, the tracker maintains reliable identities in line with, or above, recent references under comparable settings; our ablations studies indicate that multi-view appearance contributes markedly beyond a motion-only variant, clarifying that identity gains largely come from image-space crops rather than from gating alone.

Second, the occlusion-aware visibility score quantifies, for each target and instant, the effective multi-view availability of visual evidence using scene geometry and mutual-occlusion reasoning. It is used to analyse where and why identity errors arise in the online regime and to derive operational guidance for improving the tracker. In our experiments, identity errors concentrate in low-visibility regimes and recur under crowding, providing a consistent diagnostic signal. We also evaluate a visibility-aware association policy within the tracker that activates the appearance stage only when the estimated visibility exceeds a threshold; otherwise the association defers to the geometric stage. The variant is effective, but the measured gains are modest and performance is overly sensitive to the threshold. We therefore leave it for future development.

1.1 Thesis Outline

The following chapters are organised as follows. In Chapter 2 we review the literature under the adopted operational assumptions: calibrated and time-synchronised cameras, approximately planar scenes, and BEV as the reference frame. We structure prior work along three axes: (i) the fusion stage, late or early in BEV; (ii) the coupling between detection and association, ranging from reuse of detector features through joint models to detector-agnostic tracking; and (iii) the temporal regime, online or near/offline. We close by situating our contribution within that landscape.

In Chapter 3 we present the proposed method and its modules in the order in which they interact: motion in BEV with a Kalman model and geometric gating; appearance built from multi-view image crops via BEV-to-image back-projection; a two-stage online association strategy and track lifecycle management. We also introduce an occlusion-aware visibility score as a diagnostic tool to analyze identity errors.

In Chapter 4 we describe the datasets and a single, reproducible evaluation protocol, then report the main results. First, we position the tracker with respect to representative references under comparable conditions. Second, with a fixed detector, we measure how effectively detections become consistent identities over time. We include ablations that isolate the contributions of motion and appearance, a diagnostic analysis that relates error patterns to the visibility score, and an evaluation of a visibility-aware variant with a brief discussion of its dependence on context and on the activation threshold.

In Chapter 5 we summarise the main takeaways, discuss limitations that emerge from the study, and outline directions for future work suggested by the diagnostic analysis.

Appendix A collects the formal problem setting, global notation, and the geometric background used throughout the thesis. Appendix B reports additional diagnostic figures.

Chapter 2

Related Work

This chapter surveys multi-camera multi-object tracking (MC-MOT) under calibrated and synchronized setups that observe approximately planar scenes, with pedestrians as the primary target class and bird's-eye view (BEV) as the unifying reference frame. The working assumptions are accurate camera calibration, tight temporal alignment across views, and an approximately planar ground.

We structure the literature along three recurring axes: where fusion occurs (post hoc late versus early in BEV), how tightly detection and association are coupled (from detector-agnostic tracking, through reuse of detector features, to fully joint models), and the temporal regime of association (online/causal versus near- or offline global optimization). Geometry is transversal to the three axes: it is explicit in some methods via planar homographies, epipolar constraints, triangulation and back projection, and implicit in others through calibration and BEV lifting.

We follow a historical-to-conceptual arc: model-based geometric and occupancy foundations, cross-view association under late fusion, and modern early-fusion detection with BEV tracking. The chapter is organized as follows. Section 2.1 revisits geometric and occupancy-based foundations that anticipated BEV reasoning. Section 2.2 covers cross-view association and late fusion, separating geometry-driven and global methods. Section 2.3 presents early-fusion detection and BEV tracking, within which we discuss three tracking families (reuse of detector features, joint models, and detector-agnostic tracking). This taxonomy frames the trade-offs between modularity, efficiency, and robustness to occlusion, and motivates the design choices developed later, where we propose a detector-agnostic, online BEV tracker (see Chapter 3).

2.1 Geometric and Occupancy Foundations

This section revisits classical, model-based approaches that anticipated modern BEV methods. The common idea is to fuse multi-view evidence directly on the ground plane—via probabilistic occupancy maps and multi-view geometry (planar homographies, epipolar constraints)—to reason about visibility and occlusions without explicit cross-view identity matching [19, 48, 18]. These works provide the conceptual rationale for BEV but are not used directly in our pipeline; formal BEV mapping and assumptions are given in Appendix A.3.

Shared ground-plane reference. In calibrated multi-camera setups with overlapping fields of view, warping per-view evidence onto a common ground-plane reference (BEV; see Appendix A.3) reduces perspective distortions and makes observations directly comparable across views. This metric space is natural for proximity tests, gating, and motion priors; it also exposes visibility constraints that are cumbersome in per-view coordinates. Early multi-camera systems already reasoned on the ground plane by accumulating per-view evidence and tracking there, showing that many occlusions resolve more cleanly when the unit of reasoning is the metric ground position rather than per-view bounding boxes.

Evidence-level BEV occupancy (POM/CRF). Occupancy-based methods fuse per-view foreground segmentation (e.g., silhouettes) onto a ground-plane grid to estimate per-cell occupancy probabilities, treat occlusions generatively, and recover trajectories via probabilistic filtering or Viterbi-style dynamic programming [19]. Probabilistic Occupancy Maps (POM) instantiate this evidence-level fusion, while Conditional Random Fields (CRFs) extend the idea by enforcing multi-view consistency and simple occlusion models through an energy defined on the plane, optimized iteratively [48]. Conceptually, decisions are taken in BEV at the evidence level, not by late merging per-view detections; this anticipated modern BEV fusion. By construction, these classical BEV evidence-fusion methods are not tracking-by-detection: they do not rely on per-view detector outputs, and trajectories are recovered by temporal inference on occupancy rather than by linking detection instances.

Multi-view geometry as a structural prior. Planar homographies align per-camera evidence onto the ground plane, while epipolar constraints prune geometrically inconsistent correspondences [18]. Several systems accumulate "support" on the plane via homographic warps, detect local maxima as ground-plane person locations, and track them with simple motion models and appearance cues [49]. Geometry thus acts as a structural prior that stabilizes evidence fusion under strong viewpoint changes and partial occlusions, reducing reliance on per-view box matching.

Practical limitations and motivation. Occupancy/geometry pipelines depend on foreground segmentation quality (illumination changes, ghost/spurious occupancy), offer limited discriminative appearance (coarse grids, no robust re-ID), and are sensitive to calibration errors and off-plane bias. Identity maintenance degrades when occupancy peaks merge or when spurious support accumulates. These limits motivate the paradigms reviewed next: (i) cross-view association under late fusion, which first produces per-view detections/tracklets and then reconciles identities across cameras, and (ii) early-fusion detection with BEV tracking, which decides directly in BEV using richer (learned) features and maintains identities in the shared metric space [3].

2.2 Cross-View Association and Late Fusion

In calibrated, tightly synchronized multi-camera setups over an approximately planar scene, per-view hypotheses (either frame-level detections or short local tracklets) are projected to a shared ground-plane/BEV reference so that evidence becomes directly comparable across views. Cross-view identity is then resolved after this per-camera stage. Both late-fusion families fall under tracking-by-detection: detections (or short per-view tracklets) are produced first and cross-view identity is then resolved by an external association module. Two complementary families dominate. *Geometry-driven* methods enforce multi-view consistency via epipolar constraints and ground-plane homographies, optionally validating with triangulation/back-projection; they operate online with low latency and typically consume synchronized per-view detections/tracklets [32, 62]. *Global association* methods cast identity as an optimization over space–time–view—using network flow or k-shortest paths [5, 6], (lifted) multicut with long-range constraints and re-identification [51, 42], higher-order hypergraphs on tracklets [59], or long-window tracklet clustering with re-ID in offline settings [30, 61]. Some graph-based models remain causal at run time while being global in formulation [14]. Importantly, these families are not mutually exclusive: global formulations often encode geometric consistency, and geometry-driven pipelines may still rely on global solvers or causal graph updates [42, 14]. Across both families, the outcome is a set of cross-view correspondences (defining identities) and, when estimated, a consolidated ground/3D position for temporal management. Geometry is preferable with accurate calibration and substantial field-of-view overlap; global optimizers are advantageous under partial overlap, crowding, detector noise, or extended observation gaps.

2.2.1 Per-View Detections and Local Tracklets

Multi-camera tracking under late/hybrid fusion starts from *per-camera* hypotheses. Each view supplies a stream of observations that must remain reliable under density and partial occlusions, as cross-view fusion will later reconcile identities. In practice, pipelines either

expose frame-level detections directly or add a short-horizon smoothing stage to form per-view tracklets before any cross-view reasoning.

Detectors. Each camera provides frame-level detections (bounding boxes with confidence after NMS) computed by modern person detectors. Rather than surveying architectures, we frame the families most commonly used in this role: two-stage detectors for accuracy [46]; one-stage detectors for latency and throughput [38, 45, 21]; and anchor-free designs [55], with class-imbalance mitigation as a standard reference [37]. In crowded scenes, pre-training or fine-tuning on pedestrian-centric datasets is common to improve recall under heavy occlusion [64, 50]. In practice, recall-oriented settings (relaxed score thresholds, soft-NMS [10]) are often preferred at this stage, with downstream association expected to suppress residual false positives.

Tracklet Formation. When short-term temporal stability is desirable before cross-view fusion, each camera may run a lightweight online tracker to form per-view tracklets. Canonical representatives are SORT [9] as a motion-centric baseline, DeepSORT [60] when appearance embeddings help curb identity switiches, and ByteTrack [65] when preserving low-confidence detections during associaton improves robustness under partial occlusions. The intent is pragmatic: deliver short, reliable segments that stabilize local dynamics—using conservative activation/termination policies—without attempting to define a global multi-camera identity.

In what follows, cross-view fusion will consume either lists of frame-level detections or sets of per-view tracklets, depending on the family considered (see Sections 2.2.2 and 2.2.3).

2.2.2 Cross-View Association by Geometric Consistency

Multi-camera identity can be resolved by enforcing geometric consistency on hypotheses already estimated per view (typically short tracklets or local tracks). For each hypothesis in one view, candidates in the other views are screened for compatibility with the calibrated multi-view geometry and the ground-plane assumption, yielding cross-view correspondences and, when needed, a consolidated ground/3D position.

Geometric Matching Across Views. At the object level, candidate correspondences are first reduced by *epipolar* constraints, then verified on the *ground plane* via a planar homography to the shared reference, checking spatial agreement in BEV; combining epipolar and homography cues increases specificity in maximum-likelihood association schemes [62]. In overlapping setups, visibility priors (field-of-view boundaries and their intersections) further gate feasible matches across cameras [32]. When available, a final geometric validation

(e.g., triangulation or back-projection checks) consolidates the ground/3D estimate, and only geometrically coherent matches are retained. The solution regime is typically online/-greedy (frame-by-frame or with short look-ahead), with tight time alignment and simple BEV distance gating [32, 62].

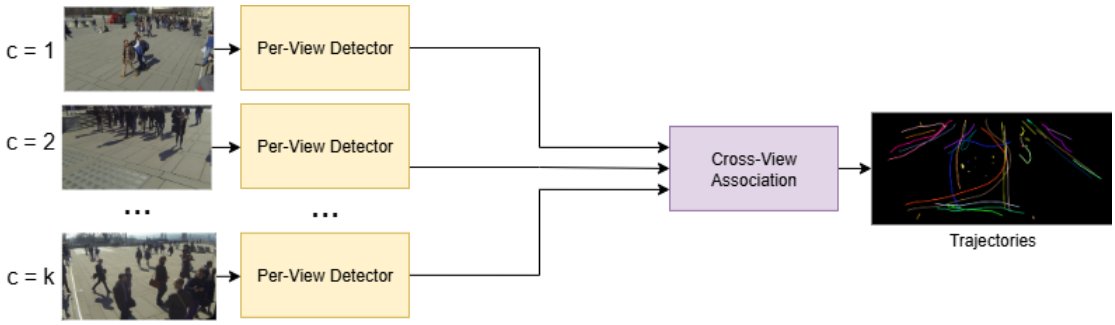
Strengths, Failure Modes, and Representative Systems. Geometry-driven association is attractive for its simplicity (few hyperparameters) and low latency when calibration is accurate and view overlap is substantial; geometric checks act as a strong filter against spurious pairings. Weaknesses emerge under long-term occlusions, heavy crowding (multiple epipolarly consistent candidates), noisy per-view hypotheses, or deviations from the planar model (calibration errors, off-plane bias). Representative operational systems integrate measurement, geometric verification, and lightweight per-view tracking to maintain consistent identities in real time on overlapping setups [35, 36]. The typical output is a set of cross-view correspondences linking per-view hypotheses across cameras and, when estimated, a consolidated ground-plane/3D position for subsequent temporal management.

2.2.3 Global Cross-View Association Methods

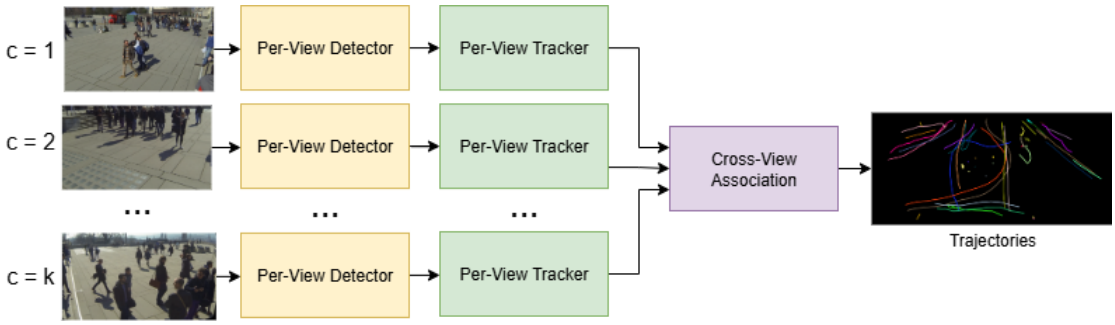
Global identity resolution can be formulated as an optimization over space–time–view, going beyond greedy matching. Depending on the formulation, these methods operate near-online (sliding window) or offline (batch), and consume either frame-level detections or per-view tracklets. The common payoff is increased robustness to misses, false positives, and long occlusions, at the cost of higher complexity and latency. Figure 2.1 contrasts the two late-fusion variants: (a) global cross-view association from per-view detections, and (b) the same association from per-view tracklets.

Network Flow and K-Shortest Paths. Flow-based formulations connect observations across frames (and views) under capacity and flow-conservation constraints, optimizing costs derived from detection confidence and temporal/motion coherence. Source/sink nodes and dummy transitions encode birth/death and missing observations; solving an LP or a k-shortest-paths problem yields globally consistent associations over a window and handles fragmentation effectively [5, 6]. Inputs are typically detections (in some cases occupancy maps), and the regime is near-online or offline depending on the window length and solver budget.

Graph Partitioning (Multicut / Lifted Multicut). Global identity can be cast as graph partitioning where detections are vertices and pairwise compatibilities are edge weights; the lifted variant introduces long-range edges to encode non-local temporal and re-identification constraints [51]. Multi-camera extensions integrate geometric consistency



(a) Late-fusion cross-view association from per-view detections



(b) Late-fusion cross-view association from per-view tracklets

Figure 2.1: Late-fusion cross-view association variants. In calibrated multi-camera setups, per-view hypotheses are first produced independently and then fused across views in a global association stage. (a) Each camera supplies frame-level detections; the global cross-view association links observations across cameras and over time to produce identity-consistent trajectories in the shared BEV frame. (b) Each camera first forms short per-view tracklets; the global association operates on these segments to produce BEV trajectories. Both panels assume a calibrated multi-view setup and can run near-online (sliding window) or offline (batch).

and cross-view edges; for instance, geometry projections to ground/3D are used to build inter-view connections before solving a lifted multicut on the spatio-temporal graph [42]. Recent graph-based models couple spatial and temporal association within a unified optimization while remaining causal at run time via incremental updates [14]. In practice, lifted/multicut consume detections, whereas geometry-augmented variants may start from per-view tracklets.

Higher-Order / Hypergraph on Tracklets. Higher-order formulations operate on tracklets and model compatibilities beyond pairs, capturing joint space–time–view consistency (co-visibility constraints, motion smoothness, appearance agreement) via hyperedges and dense substructure search [59]. The benefit is stronger disambiguation in crowded scenes and during long occlusions; the cost is heavier inference and sensitivity to tracklet quality, since errors can propagate within higher-order cliques.

Long-Window Tracklet Clustering with Re-ID (Offline). Practical pipelines—often in challenge settings—cluster per-view tracklets over long windows using strong re-identification features with spatio-temporal gating, sometimes without explicit calibration [30, 61]. These approaches prioritize identity consistency (e.g., IDF1) and recover fragmented trajectories by linking across extended gaps, accepting increased latency and a reliance on robust appearance models.

Global methods improve robustness to observation gaps, prolonged occlusions, and local noise, but increase computational cost, solver tuning, and delay. "Global" does not necessarily imply offline—graph-based models can be global in formulation yet online in execution [14]. The choice between detection- and tracklet-level inputs affects scalability and error propagation; these methods are preferable when local geometric consistency (Section 2.2.2) is insufficient due to partial overlaps, detector noise, or extended occlusions.

2.3 Early-Fusion Detection and BEV Tracking

This section introduces the early-fusion BEV paradigm for calibrated and synchronized multi-camera networks observing approximately planar scenes. The idea is to fuse multi-view evidence before object detection by projecting per-view features into a shared metric reference frame (BEV), producing outputs that are already cross-view consistent. Compared to late fusion, this reduces duplicate hypotheses across cameras and perspective conflicts, enforces spatial coherence in a single metric reference, and simplifies downstream tracking so that prediction, gating, and association operate in the same space as the measurements. These advantages come with clear assumptions: accurate calibration, temporal alignment, and an approximate ground-plane model.

We distinguish the upstream supplier from the three tracking families that build on its outputs. BEV early-fusion detectors serve as the producer of BEV hypotheses [27, 26]. On top of these, tracking appears in three forms: (i) BEV tracking that reuses detector features with an external linker [53, 2, 17]; (ii) joint BEV detection-and-tracking that co-trains detection and association within a single model [54]; and (iii) detector-agnostic BEV tracking, which constructs identity cues outside the detector and employs an external linker [57]. Under this lens, the detector-feature reuse and the detector-agnostic variants are tracking-by-detection (external linker over BEV detections), whereas the joint model is not strictly so, as detection and linking are co-trained and decided within a single network. In this landscape, our method falls within the detector-agnostic, online BEV tracking family; details are provided in Chapter 3. We proceed accordingly—BEV early-fusion detectors (Section 2.3.1), BEV tracking with detector features (Section 2.3.2), joint BEV detection and tracking (Section 2.3.3), and detector-agnostic BEV tracking (Section 2.3.4).

2.3.1 BEV Early-Fusion Detectors

Early-fusion in bird's-eye view (BEV) decides upstream, in a shared metric space on the ground plane, by fusing multi-view evidence before object detection. Outputs are already cross-view consistent (one hypothesis per subject), which mitigates per-view perspective ambiguities and exploits inter-camera overlap under occlusion. The paradigm assumes known calibration and an approximately planar scene.

Representative detectors illustrate this evolution. MVDet projects per-view features into BEV and performs convolutional aggregation to estimate pedestrian presence/centers in an anchor-free manner [27]. MVDeTr advances this design by replacing purely convolutional fusion with a view/position-aware mechanism based on deformable multi-view attention ("Shadow Transformer"), better handling projection-induced, position-dependent distortions and improving accuracy on multi-camera benchmarks such as WildTrack and MultiViewX [26, 11, 27].

Compared to late fusion, deciding in BEV upstream reduces duplicate hypotheses across cameras and cross-view conflicts, improves the spatial coherence of detections (shared metric frame), and simplifies downstream tracking since gating and association operate in the same metric space as the measurements. Known limitations remain: dependence on calibration quality, potential off-plane bias, and sensitivity to domain shift across scenes.

2.3.2 BEV Tracking with Detector Features

This family builds on BEV early-fusion detectors (see Section 2.3.1) that, beyond producing detections, expose identity cues directly in BEV (e.g., instance-level appearance descriptors, learned temporal affinities, or BEV motion cues). The design separates responsibilities: the detector learns to emit association-friendly signals in the same metric space as the measurements, while the tracking module remains external and replaceable. The linker can follow different paradigms (online assignment, lightweight graph formulations, state-space models). What unifies the family is that the detector provides the similarity evidence, and temporal linking is executed by a separate runtime procedure rather than being trained end-to-end with detection. Figure 2.2 summarizes the high-level pipeline.

Identity Cues Exposed by BEV Detectors. Representative approaches differ mainly in the type of cue they expose rather than in the association backend. One line predicts BEV re-identification embeddings per instance and passes them to the linker [53]. Another learns a temporal affinity in BEV—e.g., via attention across adjacent frames—that the linker uses as a cost [2]. A third line exposes BEV motion cues (e.g., probabilistic human flow on the BEV grid) that an external linker exploits within path-based formulations [17]. In all cases the association logic is modular and causal at runtime, and the detector's role

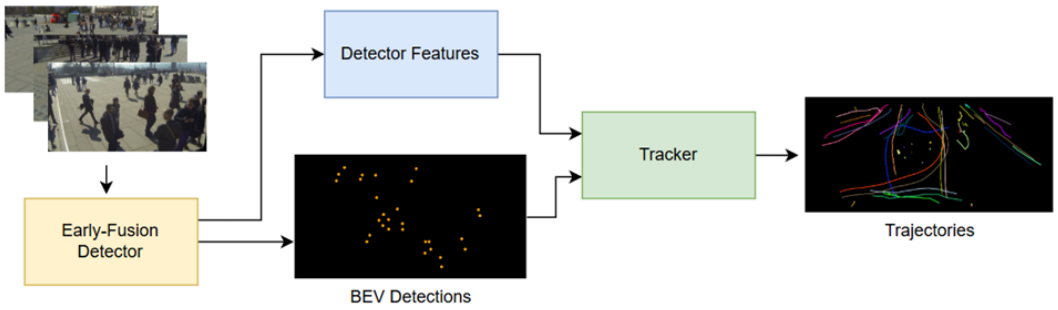


Figure 2.2: BEV tracking with detector features. Multi-view images feed an early-fusion BEV detector that emits two BEV streams: BEV Detections and detector-exposed identity cues (Detector Features). A separate tracker consumes both to link observations over time and manage the track lifecycle, producing identity-consistent trajectories.

is to supply detector-aligned evidence for identity matching.

Strengths, Limitations, and Positioning. The practical payoff is twofold: representation coherence and efficiency (identity cues live in the same BEV space as detections, avoiding per-view cropping or back-projection at runtime and keeping latency and I/O low), and operational modularity (the association backend can evolve—online, near-online, lightly global—without retraining the detector). Trade-offs mirror this coupling: identity cues are tied to the detector's backbone and training domain, so domain shift or calibration drift can affect both detection and appearance; BEV cues may be fragile where projection-induced distortions are pronounced; and prolonged occlusions still challenge purely local/online association. Conceptually, this family sits between detector-agnostic BEV tracking—where identity cues are built outside the detector (see Section 2.3.4)—and joint BEV detection-and-tracking, where cues and temporal linking are learned within a single model (see Section 2.3.3).

2.3.3 Joint BEV Detection and Tracking

This family denotes architectures that learn both detection and temporal association within the same BEV model. A multi-view backbone lifted to BEV produces shared features, and a learned association head decides identity links directly in BEV during training and inference, without relying on an external linker as the primary decision step. Current representatives exemplify the idea by co-training presence/localization and linking in a single pipeline so that identity cues and temporal reasoning are optimized together [54]. Figure 2.3 summarizes the high-level architecture.

Joint models offer native space–time coherence, reduced reliance on hand-tuned association recipes, and improved recovery under occlusions thanks to short-range BEV context.

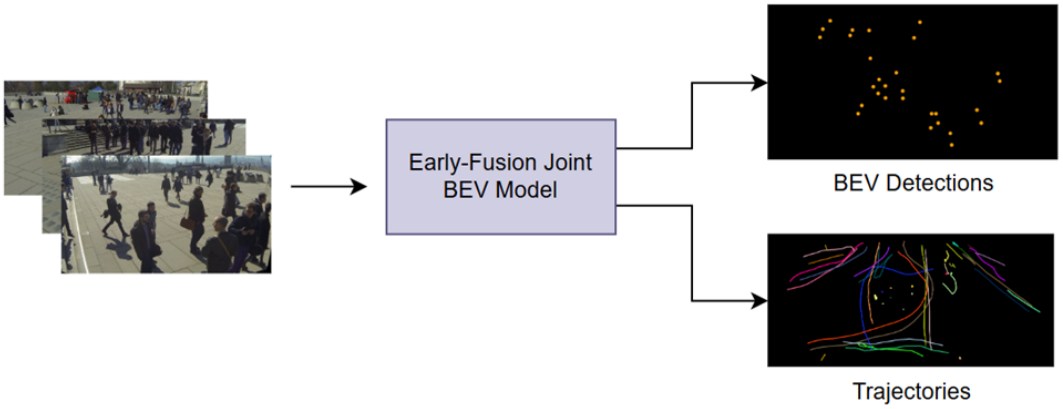


Figure 2.3: Joint BEV detection and tracking. Multi-view images feed a single early-fusion BEV model that builds shared BEV features and, through learned detection and association heads, produces BEV detections and links observations over time into identity-consistent trajectories; both tasks are trained jointly and executed within the same model at inference.

The trade-offs are higher training and memory cost, reduced modularity (detector and associator are not plug-and-play), and potential sensitivity to domain shift when co-training data are limited, with debugging and ablation less transparent than in modular pipelines. Conceptually, joint BEV detection-and-tracking sits at the most integrated end of our continuum—contrasting with detector-agnostic BEV tracking, where identity cues are built outside the detector (see Section 2.3.4), and with BEV tracking that reuses detector features while keeping an external linker (see Section 2.3.2). Outside our setting, joint end-to-end detection–tracking has also been explored in automotive perception—often multi-camera and sometimes multi-sensor—with different sensing and evaluation protocols [29, 16].

2.3.4 Detector-Agnostic BEV Tracking

This family comprises trackers that are independent of the specific BEV detector: they consume BEV detections from any early-fusion model and build identity cues outside the detector, typically via BEV-to-image back-projection to obtain multi-view crops and aggregate per-view appearance embeddings across views. Data association remains an external, replaceable module, implemented either as causal online assignment (short horizons, state-space models) or as a near-online or offline graph optimization. The aim is to maximize modularity and portability with respect to the detector while keeping a single metric output in the BEV frame. Figure 2.4 summarizes the high-level pipeline.

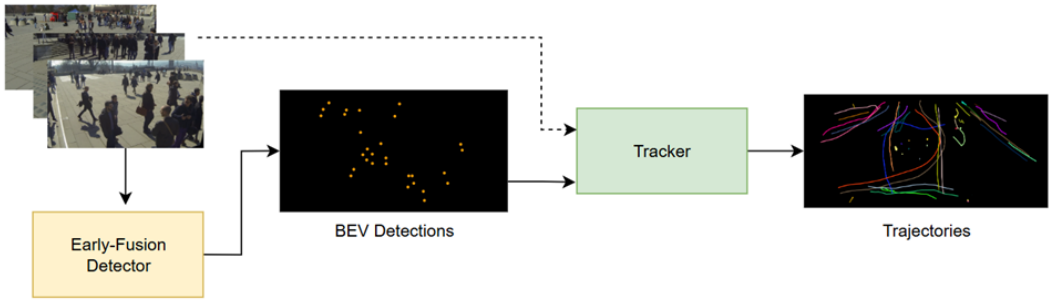


Figure 2.4: Detector-agnostic BEV tracking. Multi-view images feed an early-fusion detector that outputs detections in BEV. The tracker consumes these detections to link observations over time in BEV, maintain persistent identities and target states, and produce trajectories. The dashed arrow indicates that the tracker may access the calibrated multi-view images to derive additional identity/visibility cues.

Detector-Agnostic Identity Cues and Association. Representative systems construct appearance in a detector-agnostic manner: they associate BEV hypotheses to per-view 2D observations, extract features from multi-view crops, aggregate them into cross-view descriptors, and solve association on a spatio-temporal graph with hierarchical GNNs and a global merging layer to handle long occlusions and fragmentation [57]. Some systems additionally perform a 2D–3D detection association to select and verify the per-view crop used for re-identification, improving robustness under projection misalignment [57]. Within this family lies our method: like these systems, we obtain appearance from multi-view crops via BEV-to-image back-projection; unlike graph/global approaches, we replace the graph/GNN with a lightweight online association entirely in BEV—Multiple Object Tracking with Detection and Tracking (MOTDT)–style gating and matching—while keeping appearance decoupled from the detector [13, 53, 2]. See Chapter 3 for details.

Strengths, Limitations, and Positioning. Decoupling identity cues from the detector yields portability across detectors (e.g., MVDet/MVDeTr [27, 26] can be swapped without retraining identity features) and clearer diagnosability/ablation, since motion, appearance, and association can be isolated. Because appearance is built from image-space multi-view crops and then aggregated across views, these trackers are less exposed to BEV projection artifacts; domain adaptation can be handled by fine-tuning the re-ID backbone independently of the detector. The trade-offs are runtime and I/O overhead for multi-view cropping (and any 2D lookup used to select crops), stronger dependence on calibration and per-view visibility for reliable back-projection, and no end-to-end synergy between detection and identity cues; robustness to prolonged occlusions depends on the chosen backend (lightweight online linkers are more fragile, graph/global variants increase complexity and latency). In our continuum, detector-agnostic BEV tracking sits at the most modular end

(cues outside the detector; external linking): compared with hybrid methods that reuse detector features (Section 2.3.2) it trades some latency and tight BEV alignment of cues for greater portability and resilience to projection distortions or backbone changes; compared with joint models (Section 2.3.3) it forgoes co-training gains in exchange for simpler training, easier debugging, and the freedom to swap detector or association without retraining the whole stack.

Chapter 3

Proposed Method for Multi-Camera Multi-Object Tracking

This chapter presents the proposed multi-camera multi-object tracking (MC-MOT) method for calibrated, synchronized networks of fixed cameras observing an approximately planar scene with overlapping fields of view. The formal problem setting, notation, and geometric background (BEV and multi-view calibration) are collected in Appendix A; here we focus on the algorithmic design—how motion prediction, appearance modeling, and online association are combined on top of early-fusion BEV detections to maintain identity consistency under challenging visibility.

We make two contributions. First, we build a detector-agnostic online tracker by combining a BEV Kalman filter with predictive gating for motion [53, 2] and multi-view crop embeddings for appearance via BEV-to-image back-projection [57]. Compared with EarlyBird/attention-aware and BEV-SUSHI, respectively, we retain their strengths—Kalman-based motion and multi-view crop appearance—while replacing the complementary components: detector-head embeddings [53, 2] with crop-based descriptors, and graph neural network (GNN) association [57] with online Kalman-based assignment and geometric gates. To our knowledge, this exact motion–appearance combination has not, to date, appeared in a single online BEV tracker for calibrated multi-camera pedestrian tracking on planar scenes.

Second, we introduce an occlusion-aware visibility score: a purely geometric, per-target measure of visual availability across cameras computed solely from camera poses and BEV predictions (no RGB). We use it to diagnose failure modes (e.g., identity switches and dropouts) and discuss how it can inform visibility-aware weighting within the pipeline.

The chapter is organized as follows. Section 3.1 outlines the high-level system architecture, describing the data flow from multi-view detections to trajectory output and clarifying the roles of individual modules. Section 3.2 details the BEV Kalman filter used for motion prediction, providing spatial priors and gating constraints for association. Section 3.3 describes the extraction and fusion of multi-view appearance descriptors, enabling identity association in visually ambiguous situations. Section 3.4 specifies the online association strategy and track management logic, including spatial gating, appearance-based matching, geometric fallback, and lifecycle policies. Finally, Section 3.5 introduces an occlusion-aware visibility score, a purely geometric measure of visual availability that serves both as a diagnostic tool and as a potential signal for visibility-aware integration in the pipeline.

3.1 Overview of the Proposed System Architecture

This section introduces the main modules of the proposed tracking system and clarifies their roles and interactions. The pipeline operates primarily in the bird's-eye view (BEV) domain, with per-view image crops only for appearance extraction. Input detections come from an upstream early-fusion multi-view detector that projects results into BEV; downstream modules handle motion prediction, appearance representation, online association, and track management. The tracker runs online and causally, updating the state frame by frame without access to future observations.

Figure 3.1 summarizes the architecture and data flow across modules; detailed algorithms are presented later (Sections 3.2, 3.3, 3.4).

Early-fusion detection (upstream) The tracker consumes 2D detections already localized in BEV, produced by an external early-fusion multi-view detector. In this work we adopt *MVDeTr* [26], which aggregates cross-view cues with a transformer architecture and reports competitive results on WildTrack and MultiViewX. Early-fusion methods [27] integrate geometry across views before detection, unlike late-fusion strategies that merge results post hoc. As the detector is upstream, its outputs are directly fed to the tracking pipeline.

Motion model The motion model predicts, in BEV, the likely location of each active track at the next frame, providing spatial priors and gating for association. We use a Kalman filter for its simplicity and efficiency; the state definition and update equations are detailed in Section 3.2.

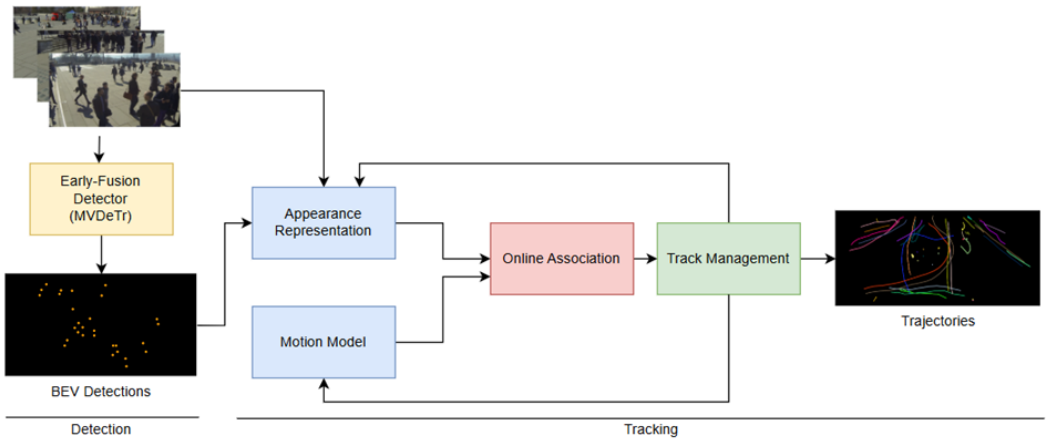


Figure 3.1: Overview of the tracking pipeline. An early-fusion multi-view detector produces BEV detections. A BEV Kalman motion model and multi-view appearance representation provide priors and descriptors for the online association module; track management implements the lifecycle to output identity-consistent trajectories.

Appearance representation Geometry alone is insufficient in crowded scenes. For each BEV detection, we back-project to calibrated camera views, extract fixed-size image patches (ROIs) where the detection is visible, and compute per-view embeddings with a pre-trained network. Per-view descriptors are aggregated into a single representation used during matching. The procedure and design choices are described in Section 3.3.

Online association At each frame, detections are matched to existing tracks using a cost that combines motion priors (gating in BEV) and appearance similarity. The assignment strategy and tie-breaking rules are detailed in Section 3.4.

Track management Based on association outcomes, tracks are initialized, updated, temporarily kept during short miss streaks, or deleted after prolonged absence, following the lifecycle in Appendix A.6. Implementation details (e.g., thresholds and counters) are given in Section 3.4.5.

3.2 Motion Model with Kalman Filtering

The tracking pipeline relies on a motion model to preserve temporal coherence and constrain data association in a strictly online, causal setting. We adopt a linear Kalman filter in the bird's-eye view (BEV) domain, where multi-camera detections are expressed in metric ground-plane coordinates. A constant-velocity prior yields lightweight, closed-form prediction and update, enabling real-time operation while supplying the association module with

spatial priors and statistical gating cues. This choice favors simplicity, interpretability, and clean integration with downstream components, without sequence-level optimization (see Appendix A.2 for all symbol definitions).

The section is organized as follows. We motivate the role of the motion model and its interaction with association and lifecycle logic (Section 3.2.1); specify the BEV Kalman filter—state, dynamics, measurement model, and exported quantities for gating (Section 3.2.2); provide practical initialization and tuning guidelines (Section 3.2.3); summarize the per-frame predict/update loop and missed detections (Section 3.2.4); and finally state assumptions and failure modes (Section 3.2.5).

3.2.1 Role of the Motion Model in Online Tracking

In online MC-MOT, per-frame detections lack temporal memory, making identity maintenance vulnerable to occlusions, missed observations, and clutter. A motion model provides a *predictive prior* over each target's state (position and velocity) in the shared BEV frame. This prior bridges short observation gaps, reduces identity switches, and stabilizes trajectories under noise.

Operationally, the model is invoked *before* association: it forecasts the next state and its uncertainty, which are then used to (i) restrict candidates via spatial gating and thus shrink the assignment problem, and (ii) preserve continuity when no detection arrives by propagating the state forward (see Section 3.2.4). In a causal setting—no access to future frames—such predictive anchoring is essential: decisions at time t rely only on past and present information, so the prior constrains matching in dense scenes and mitigates ambiguity among nearby detections (see Section 3.4.2).

We model motion with a linear Kalman filter in BEV (Section 3.2.2), a standard choice in real-time MOT for its efficiency, interpretability, and closed-form updates [31, 4]. Its role mirrors successful online trackers—e.g., SORT/DeepSORT in monocular settings [9, 60] and recent multi-camera pipelines such as EarlyBird and attention-aware MPT [53, 2]—where a lightweight constant-velocity prior delivers strong robustness without sequence-level optimization.

3.2.2 Kalman Filter in Bird's-Eye View

We model target dynamics with a linear Kalman filter in the Bird's-Eye View (BEV) frame. States are expressed in metric ground-plane coordinates (meters), which removes perspective effects and simplifies motion reasoning. Each track follows a constant-velocity (CV) model with state

$$\mathbf{x}_t \;=\; \left[x_t,\, y_t,\, \dot{x}_t,\, \dot{y}_t\right]^\top,$$

i.e., 2D position and velocity on the ground plane.

State transition. For a sampling interval Δt, the CV transition is

$$\mathbf{F}(\Delta t) \;=\; \begin{bmatrix} 1 & 0 & \Delta t & 0 \\ 0 & 1 & 0 & \Delta t \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}. \tag{3.1}$$

Unmodeled accelerations are captured by process noise $\mathbf{Q}(\Delta t)$. Using the standard white-acceleration model with spectral density σ_a^2 [31, 4],

$$\mathbf{Q}(\Delta t) \;=\; \sigma_a^2 \begin{bmatrix} \frac{\Delta t^4}{4} & 0 & \frac{\Delta t^3}{2} & 0 \\ 0 & \frac{\Delta t^4}{4} & 0 & \frac{\Delta t^3}{2} \\ \frac{\Delta t^3}{2} & 0 & \Delta t^2 & 0 \\ 0 & \frac{\Delta t^3}{2} & 0 & \Delta t^2 \end{bmatrix}. \tag{3.2}$$

When timestamps are not uniform, $\mathbf{F}(\Delta t)$ and $\mathbf{Q}(\Delta t)$ are recomputed per frame using the current inter-frame Δt.

Measurement model. BEV detections provide position only, so the observation model is

$$\mathbf{z}_t \;=\; \mathbf{H}\,\mathbf{x}_t + \boldsymbol{\varepsilon}_t, \qquad \mathbf{H} \;=\; \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix}, \tag{3.3}$$

with $\boldsymbol{\varepsilon}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{R})$. We take $\mathbf{R}$ diagonal (typically anisotropic), e.g. $\mathbf{R} = \mathrm{diag}(\sigma_x^2, \sigma_y^2)$, reflecting BEV localization noise in meters.

Filtering cycle and gating outputs. We denote by $\hat{\mathbf{x}}_{t|t-1}$ and $\mathbf{P}_{t|t-1}$ the predicted mean and covariance (per track), and by $\hat{\mathbf{x}}_{t|t}$ and $\mathbf{P}_{t|t}$ the filtered estimates after a measurement update. At each frame, the filter performs prediction with (3.1)–(3.2) and, when a detection is assigned, a correction with (3.3). The predicted measurement and its covariance,

$$\hat{\mathbf{z}}_{t|t-1} \;=\; \mathbf{H}\,\hat{\mathbf{x}}_{t|t-1}, \qquad \mathbf{S}_t \;=\; \mathbf{H}\,\mathbf{P}_{t|t-1}\,\mathbf{H}^\top + \mathbf{R},$$

are exported to the association module for statistical gating (see Section 3.4.2). Tracks are filtered independently (no cross-covariances).

Initialization, hyperparameters, and numerical details are deferred to Section 3.2.3; the recursive predict/update equations are summarized in Section 3.2.4.

3.2.3 Initialization and Tuning

This subsection details how the BEV Kalman filter is instantiated at track birth and how its noise parameters are chosen. Concrete numeric values are reported in Section 4.3.5.

State and covariance at birth. Given a BEV detection $\mathbf{z}_0 = [x_0\ y_0]^\top$, initialize

$$\hat{\mathbf{x}}_0 = [x_0,\ y_0,\ 0,\ 0]^\top, \qquad \mathbf{P}_0 = \mathrm{diag}(\sigma_{p0}^2, \sigma_{p0}^2, \sigma_{v0}^2, \sigma_{v0}^2),$$

with σ_{p0} on the order of the detector's BEV localization noise and $\sigma_{v0} \gg \sigma_{p0}/\Delta t$ (unknown velocity). If two consecutive detections are available, warm-start $\dot{\mathbf{p}}_0 \approx (\mathbf{z}_1 - \mathbf{z}_0)/\Delta t$ and reduce σ_{v0} accordingly.

Measurement covariance R. Detections provide positions only, so $\mathbf{R} = \mathrm{diag}(\sigma_x^2, \sigma_y^2)$. Two practical recipes: (i) *From BEV resolution:* for grid pitch g [m], use $\sigma_x \approx \sigma_y \approx g/\sqrt{12}$; (ii) *From residuals:* run a short burn-in, collect innovations $\mathbf{r}_t = \mathbf{z}_t - \hat{\mathbf{z}}_{t|t-1}$ on matched pairs, and set $\sigma_x \approx 1.4826\,\mathrm{MAD}\{(r_t)_x\}$, $\sigma_y \approx 1.4826\,\mathrm{MAD}\{(r_t)_y\}$, where MAD is the median absolute deviation (a robust scale estimator). If BEV resolution or calibration accuracy differs along the two axes, use an anisotropic $\mathbf{R}$.

Process noise $\mathbf{Q}(\Delta t)$. We adopt the white-acceleration model (3.2) with spectral density σ_a^2 [31, 4]. Interpret σ_a as typical unmodelled acceleration ($\mathrm{m/s^2}$). A useful heuristic is to choose σ_a so that the expected 3σ position spread over 1 s matches scene dynamics (pedestrians ∼1–2 m), then refine on a validation split. Always recompute $\mathbf{Q}(\Delta t)$ with the current Δt.

Variable frame rate. Compute Δt from timestamps (seconds) and update $\mathbf{F}(\Delta t)$ and $\mathbf{Q}(\Delta t)$ accordingly. For rare long gaps, either cap $\Delta t \le \Delta t_{\max}$ or allow large $\mathbf{Q}$ while capping Δt to avoid excessive covariance blow-up. Ensure units are consistent (meters / $\mathrm{m\,s^{-1}}$).

Gating thresholds. Downstream association uses *Euclidean* gating by a fixed radius τ (Section 3.4.2). To choose τ principledly, match it to the size of the 3σ ellipse implied by the innovation covariance $\mathbf{S}_t = \mathbf{H}\mathbf{P}_{t|t-1}\mathbf{H}^\top + \mathbf{R}$. For reference, a *Mahalanobis* gate accepts pairs with $d^2 \le \chi^2_{2,\alpha}$ (5.99 at 95%, 11.83 at 99.7%) [4]; we do not use it in our implementation but report it for completeness and comparability.

All hyperparameters $(\sigma_{p0}, \sigma_{v0}, \sigma_a, \mathbf{R}, \tau, \Delta t_{\max})$ are fixed per dataset and their values (and selection procedure) are reported in Section 4.3.5.

3.2.4 Predict, Update, and Missed Updates

This subsection summarizes the recursive Kalman cycle executed at every frame and the quantities exposed to the association and lifecycle modules.

Prediction. Given the filtered state and covariance at $t-1$, the one-step prediction with the CV model (Section 3.2.2) is

$$\hat{\mathbf{x}}_{t|t-1} = \mathbf{F}(\Delta t)\,\hat{\mathbf{x}}_{t-1|t-1}, \qquad \mathbf{P}_{t|t-1} = \mathbf{F}(\Delta t)\,\mathbf{P}_{t-1|t-1}\,\mathbf{F}(\Delta t)^\top + \mathbf{Q}(\Delta t).$$

The predicted measurement and its covariance are

$$\hat{\mathbf{z}}_{t|t-1} = \mathbf{H}\,\hat{\mathbf{x}}_{t|t-1}, \qquad \mathbf{S}_t = \mathbf{H}\,\mathbf{P}_{t|t-1}\,\mathbf{H}^\top + \mathbf{R},$$

and are exported to the association module to support gating and threshold selection (Section 3.4.2).

Update (on a matched detection). For an assigned detection $\mathbf{z}_t$, the innovation and Kalman gain are

$$\mathbf{y}_t = \mathbf{z}_t - \hat{\mathbf{z}}_{t|t-1}, \qquad \mathbf{K}_t = \mathbf{P}_{t|t-1}\,\mathbf{H}^\top\,\mathbf{S}_t^{-1},$$

and the corrected estimates are

$$\hat{\mathbf{x}}_{t|t} = \hat{\mathbf{x}}_{t|t-1} + \mathbf{K}_t\,\mathbf{y}_t, \qquad \mathbf{P}_{t|t} = (\mathbf{I} - \mathbf{K}_t\mathbf{H})\,\mathbf{P}_{t|t-1}.$$

The above is the standard covariance update; the Joseph form $\mathbf{P}_{t|t} = (\mathbf{I}-\mathbf{K}_t\mathbf{H})\,\mathbf{P}_{t|t-1}\,(\mathbf{I}-\mathbf{K}_t\mathbf{H})^\top + \mathbf{K}_t\mathbf{R}\mathbf{K}_t^\top$ may be used for additional numerical robustness [4].

Missed Update (no assigned detection). If no detection is associated at time t, the correction is skipped and

$$\hat{\mathbf{x}}_{t|t} = \hat{\mathbf{x}}_{t|t-1}, \qquad \mathbf{P}_{t|t} = \mathbf{P}_{t|t-1},$$

so uncertainty remains elevated and will grow at subsequent predictions. Miss counters and state transitions are handled by the lifecycle logic (Appendix A.6).

Interface to Other Modules. Each frame, the motion model exposes $(\hat{\mathbf{z}}_{t|t-1}, \mathbf{S}_t)$ to support gating and candidate pruning (Section 3.4.2), and the filtered state $(\hat{\mathbf{x}}_{t|t}, \mathbf{P}_{t|t})$ to be used at the next step. The association outcome (hit/miss) determines whether the update is executed and informs track state transitions.

For completeness, we refer to [31, 4] for the classical derivation and properties of the filter.

3.2.5 Assumptions and Limitations

The BEV Kalman filter trades modeling expressiveness for speed, determinism, and ease of integration. We assume targets move on a common ground plane and that cameras are fixed, calibrated, and time-synchronized, so BEV positions are metrically consistent across frames. Observations provide ground-plane positions (x, y) only; velocities are latent. Each target evolves independently under a constant-velocity (CV) prior with per-track filtering and no cross-covariances. Noise covariances are treated as stationary within a sequence, and variable frame rate is handled by recomputing $\mathbf{F}(\Delta t)$ and $\mathbf{Q}(\Delta t)$ per frame (see Sections 3.2.2 and 3.2.3). Downstream gating uses a fixed Euclidean radius (Section 3.4.2), preferred over Mahalanobis for its simplicity and latency stability in our setting.

These assumptions imply clear failure modes. A CV prior struggles with abrupt accelerations, sharp turns, or interaction-driven dynamics (e.g., mutual avoidance in crowds), which inflate prediction error and weaken gating. A strictly planar BEV introduces bias in non-planar environments (ramps, stairs) and increases sensitivity to calibration drift. Position-only measurements limit recovery after long occlusions, since velocity must be inferred from past motion. Finally, fixed (non-adaptive) noise levels may be over- or under-confident under varying local conditions, degrading both updates and gates.

The choice is pragmatic for a real-time, online pipeline: no training, closed-form updates, and clean interfaces with association and lifecycle logic (Sections 3.2.4 and A.6). This mirrors successful online trackers—e.g., SORT/DeepSORT in monocular settings [9, 60] and recent multi-camera early-fusion systems like EarlyBird [53]—where a lightweight CV prior delivers robust performance without sequence-level optimization. By contrast, pipelines with learned long-term association (e.g., hierarchical GNNs in BEV-SUSHI [57]) better handle extended occlusions and complex interactions, at the cost of higher latency, training infrastructure, and tighter module coupling—outside our causal, modular design goal. In practice, we mitigate some of these limits via conservative lifecycle thresholds and a geometric fallback in association (see Section 3.4.5). Visibility cues are discussed in Section 3.5.3.

3.3 Appearance Representation from Multi-Camera Projections

Accurate and robust appearance modeling is essential in multi-camera multi-object tracking (MC-MOT), particularly in crowded or cluttered scenarios where motion prediction alone

cannot reliably preserve target identities. By leveraging calibrated and synchronized camera views, the proposed pipeline extracts, processes, and fuses visual descriptors to enable consistent identity association even under partial occlusion, similar motion patterns, or temporary loss of visual contact. This integration reduces identity switches and trajectory fragmentation, improving overall tracking stability.

This section describes the appearance modeling component, which operates in the image domain while remaining tightly coupled with the BEV-based motion framework. We first motivate the role of visual cues in resolving motion-based ambiguities (Section 3.3.1), then detail the BEV-to-image back-projection process used to localize each tracked individual in all visible camera views (Section 3.3.2). Next, we present the embedding network architecture, operational setup, and descriptor properties (Section 3.3.3), followed by the multi-view aggregation strategy that produces a unified identity representation (Section 3.3.4). We conclude by discussing the main design considerations, current limitations, and potential improvements (Section 3.3.5).

3.3.1 Role of Appearance Features in Association

In multi-object tracking, especially in dense or cluttered scenes, geometric prediction alone is often insufficient to preserve target identity. In the BEV domain, different individuals may follow similar trajectories, remain static for extended periods, or move in close proximity, leading to ambiguous associations. This frequently occurs when pedestrians walk side-by-side, cross paths, or temporarily leave the field of view; upon reappearance, multiple candidates may satisfy the motion-based gate, making identity recovery uncertain.

Appearance cues provide complementary, identity-specific information that is partially invariant to short-term motion ambiguities and moderate viewpoint changes. A compact descriptor—typically obtained from a convolutional re-identification network—encodes stable visual properties such as clothing color, texture, and local patterns. These features remain consistent across frames and have been shown to significantly reduce identity switches in both monocular and multi-camera MOT settings [60, 69, 12, 28, 53, 57].

In the proposed system, each active track is projected back into calibrated camera views where it is expected to be visible (Section 3.3.2). From the resulting Regions of Interest (ROIs), appearance embeddings are extracted and used to compute a similarity cost between tracks and detections during association (Section 3.4.3). This follows a widely adopted paradigm in MOT pipelines [60, 65], where motion-based gating narrows the candidate set and visual similarity refines the final assignment. By combining spatial prediction with visual descriptors, the tracker maintains identity consistency even under occlusion, crowding, or near-identical motion patterns, thereby improving metrics such as IDF1 and reducing trajectory fragmentation in challenging multi-camera scenarios.

3.3.2 BEV-to-Image Back-Projection

To extract appearance cues from image space, each active BEV track must be localized in the calibrated camera views where it is likely visible. This stage bridges the geometric tracking space with the visual embedding pipeline, enabling per-view Region of Interest (ROI) extraction for identity modeling. We leverage the planar ground model and the per-camera calibration defined in Appendix A, Sections A.3 and A.4. For camera c with intrinsics $\mathbf{K}_c$ and extrinsics $[\mathbf{R}_c \mid \mathbf{T}_c]$, the mapping from ground plane to image is given by the plane-to-image homography $\mathbf{H}_c$ induced by z=0 [23, 66].

Let $\tilde{\mathbf{X}}_g = [X\ Y\ 1]^\top$ be the homogeneous ground-plane location of the track; the corresponding image point in camera c is

$$\tilde{\mathbf{x}}^{(c)} \sim \mathbf{H}_c \tilde{\mathbf{X}}_g, \qquad \mathbf{x}^{(c)} = \Pi\big(\tilde{\mathbf{x}}^{(c)}\big),$$

where $\Pi(\cdot)$ denotes dehomogenization. If $\mathbf{x}^{(c)}$ lies outside the image bounds or behind the camera, the view is skipped.

ROI construction. Around the ground-contact point $(X, Y,0)$ we instantiate a vertical 3D box of physical width w and height h (dataset-specific pedestrian priors; see Section 4.3). Let $\mathbf{u} = (u_x, u_y)$ be a unit vector on the ground plane orthogonal to the viewing ray.[1] The two bottom corners are

$$(X, Y,0)^\top \pm \tfrac{w}{2}\,[u_x, u_y,0]^\top,$$

and the two top corners are obtained by adding $(0,0, h)^\top$. The four 3D corners are projected to the image of camera c via $\mathbf{P}_c = \mathbf{K}_c[\mathbf{R}_c \mid \mathbf{T}_c]$. The ROI is the minimal axis-aligned rectangle enclosing the four projected points.

Validity checks and output. ROIs partially outside the image are clipped; ROIs fully outside are discarded. Numerical degeneracies (e.g., vanishing homogeneous scale during dehomogenization) are skipped. For each active track and each visible camera, the output is a valid image-space ROI forwarded to the embedding network (Section 3.3.3). This keeps geometry centralized in BEV while exposing per-view crops for appearance extraction, without rederiving calibration details already covered in Appendix A.4.

[1] Any consistent choice that yields a lateral direction in the ground plane is acceptable; the exact construction is implementation-dependent.

3.3.3 Embedding Network and Descriptor

Following ROI extraction (Section 3.3.2), each image crop is mapped to a fixed-dimensional feature encoding stable, identity-specific visual cues. This descriptor complements motion during association and helps recover from partial occlusions [60, 28].

Architecture choice (design-level). The embedding backbone must (i) produce discriminative features for visually similar individuals under varying viewpoints, (ii) run at low latency, and (iii) integrate seamlessly into multi-camera pipelines without domain-specific retraining. We rely on OSNet [69]: its omni-scale residual blocks with dynamic feature fusion capture both fine- and coarse-grained cues. OSNet has shown strong performance within MOT pipelines (e.g., BoT-SORT [1], ReST [14], 3D-aware trackers [44]). In our experiments we instantiate this component with OSNet; implementation specifics (pretrained weights, input size, descriptor dimensionality) are deferred to Section 4.3.

Output. For each visible camera we obtain a per-view descriptor $\mathbf{f}_c \in \mathbb{R}^d$ (later ℓ_2-normalized in setup), preserving the camera index c. These descriptors are passed unchanged to the multi-view aggregation stage (Section 3.3.4).

3.3.4 Multi-View Aggregation of Appearance Features

As described in Section 3.3.3, each tracked individual observed in a calibrated multi-camera setup produces one appearance descriptor per visible camera view. While these per-view embeddings may contain discriminative cues, they can also be degraded by occlusions, motion blur, truncation at image boundaries, or viewpoint-specific bias. Aggregating descriptors across views mitigates these effects and yields a more stable and complete identity representation, which is particularly advantageous in crowded or cluttered scenes.

Let $\{\mathbf{f}_i\}_{i=1}^{N}$ be the set of ℓ_2-normalized d-dimensional embeddings extracted at the same time step from the N camera views in which the target is visible. The unified multi-view descriptor $\bar{\mathbf{f}}$ is obtained via element-wise average pooling:

$$\bar{\mathbf{f}} \;=\; \frac{1}{N}\sum_{i=1}^{N}\mathbf{f}_i,$$

which is invariant to the number of contributing views and can be efficiently vectorized, making it suitable for real-time operation. This equal-weight fusion avoids the need for per-view confidence scores or scene-specific priors, which may be unavailable or unreliable during online tracking. Similar averaging strategies for cross-view appearance aggregation have been successfully employed in multi-camera tracking frameworks such as [57].

The resulting $\bar{\mathbf{f}}$ becomes the canonical appearance vector for the track at the current frame and is stored in the track state for subsequent use in appearance-based association (Section 3.4.3). More advanced multi-view fusion strategies, including quality- or visibility-weighted averages and learned attention mechanisms, are discussed in Section 3.3.5.

3.3.5 Considerations and Limitations

This subsection outlines the main design choices of the proposed appearance modeling pipeline, followed by its key limitations and possible directions for improvement.

Design considerations. The system adopts a modular pipeline that decouples detection from appearance modeling. Detections are produced in the BEV domain and then back-projected into each visible camera to extract localized crops. This approach, inspired by [57], contrasts with integrated architectures such as [53, 2], where detection and embedding share the same backbone. The modular design enables the use of off-the-shelf re-identification models, allows independent replacement or tuning of components, and preserves a clear separation between spatial tracking logic and appearance-based reasoning.

Limitations. Two main limitations remain:

1. **No ROI content validation.** Unlike [57], which uses an additional image-space detector to verify that projected regions correspond to visible individuals, the proposed system accepts any back-projected ROI falling within image bounds. As a result, projections affected by calibration noise, occlusions, or blind spots may yield crops containing mostly background, fully occluded targets, or partial body fragments. These low-information regions are still processed by the embedding network, potentially degrading descriptor quality and reducing the reliability of appearance-based association (Section 3.4.3).

2. **No domain adaptation.** The embedding network operates in a zero-shot setting, pre-trained on datasets such as Market-1501 and MSMT17. The resulting descriptors may lose discriminative power when the deployment domain differs substantially in illumination, resolution, or appearance statistics—particularly in highly specific indoor environments such as warehouses or hospitals.

Potential improvements. To address these limitations, future work could integrate a lightweight ROI verification step to reject visually unreliable crops and adapt the embedding model to the deployment domain via supervised fine-tuning or self-supervised

adaptation. These enhancements would align the appearance modeling pipeline with state-of-the-art robustness and adaptability standards in multi-camera tracking.

3.4 Online Association and Track Management

The tracking pipeline maintains target identities through a sequential combination of online association and track management. The association module matches incoming detections to active tracks by integrating motion priors and appearance cues, while the management module updates the corresponding states and enforces lifecycle rules. This joint operation ensures temporal consistency in multi-camera multi-object tracking (MC-MOT) even under occlusion, dense interactions, or degraded visual conditions.

The section is organized as follows. Section 3.4.1 defines the scope, inputs, outputs, and dependencies of the two modules. Section 3.4.2 introduces the motion-based candidate filtering that constrains the association search space. Section 3.4.3 describes the computation of appearance-based matching costs. Section 3.4.4 details the two-stage assignment strategy that prioritizes visual consistency while recovering matches through geometric fallback. Finally, Section 3.4.5 specifies the update rules for motion and appearance states and the policies for handling unmatched tracks and detections.

3.4.1 Module Scope and Interfaces

The online association and track management modules jointly maintain consistent target identities over time in a calibrated multi-camera setup. Operating in an online, causal fashion, they integrate motion- and appearance-based cues to match detections with existing tracks at each frame, update the corresponding states, and enforce lifecycle rules. The association module sits downstream of the motion model (Section 3.2) and the appearance representation component (Section 3.3), while the track management module operates immediately after association, before the system's trajectory output stage.

Inputs. At each time step t, the module receives:

- Predicted track states $\{\hat{\mathbf{s}}_j^{(t)}\}$ from the motion model, expressed in the bird's-eye view (BEV) coordinate frame, including predicted position, velocity, and associated uncertainty.
- Candidate detections $\{\mathbf{d}_i^{(t)}\}$ from the upstream multi-view detector, also in BEV coordinates, optionally accompanied by per-camera Regions of Interest (ROIs) and aggregated appearance descriptors from Section 3.3.4.
- Track appearance descriptors $\{\mathbf{a}_j^{(t-1)}\}$ maintained from previous frames, representing each target's historical visual profile.

Outputs. After processing, the module delivers updated track states, correcting BEV positions and velocities for matched tracks while retaining predicted states for unmatched ones. Appearance profiles are refreshed for tracks successfully matched with detections containing valid descriptors. The module also issues lifecycle decisions, such as initializing new tracks, keeping inactive ones alive for short gaps, or removing stale tracks.

Interfaces and dependencies. This module interacts directly with the motion model, which supplies the predicted states and is updated in response to association outcomes; with the appearance extraction and aggregation pipeline, which provides per-frame visual descriptors for detections; and with the track management subsystem, which implements lifecycle policies such as initialization, tentative confirmation, and deletion thresholds [9, 60].

Scope. The association logic is organized into distinct subcomponents: spatial gating, appearance cost computation, matching, and fallback strategies. This modularity allows independent modification or replacement of each stage (e.g., using a stronger candidate filtering strategy as in [65] instead of Euclidean gating), supports fine-grained ablation studies to assess the contribution of individual cues, and facilitates adaptation to different datasets or detection sources without altering the overall control flow.

This module constitutes the operational core of the tracking pipeline's online phase, ensuring identity continuity under real-time constraints while remaining extensible for future enhancements such as learned matching costs [7] or joint detection–tracking architectures [70].

3.4.2 Candidate Filtering via Motion Gating

As introduced in Section 3.4.1, the first step in online association is a motion-based gating stage, acting as a geometric pre-filter. It discards detection–track pairs that are spatially implausible given the predicted states and uncertainties from the motion model (Section 3.2), thereby reducing the size of the assignment cost matrix and lowering the risk of false associations [60, 1]. The gating operates in the bird's-eye view (BEV) domain defined in Appendix A.3, ensuring that spatial constraints are applied consistently across all camera views.

Let $\hat{\mathbf{p}}_t \in \mathbb{R}^2$ denote the predicted BEV position of a track at time t (from the motion model's prediction step, Section 3.2.4) and $\mathbf{d}_i \in \mathbb{R}^2$ the BEV location of a candidate detection. The Euclidean gating condition is:

$$\|\hat{\mathbf{p}}_t - \mathbf{d}_i\|_2 < \tau,$$

where τ is a fixed threshold defining the acceptance region. Detections outside this region are discarded before appearance matching.

A more statistically grounded variant uses the Mahalanobis distance [4] to incorporate the predicted uncertainty:

$$\sqrt{(\hat{\mathbf{p}}_t - \mathbf{d}_i)^\top \mathbf{S}_t^{-1} (\hat{\mathbf{p}}_t - \mathbf{d}_i)} < \tau',$$

where $\mathbf{S}_t$ is the predicted position covariance exported by the motion model. This produces an elliptical acceptance region aligned with the uncertainty ellipse. While this approach adapts to anisotropic uncertainty, it requires matrix inversion and can be less stable when covariance estimates are noisy.

In our implementation, the Euclidean formulation is adopted for its computational efficiency and stability under approximately isotropic error distributions. The threshold τ is a key hyperparameter: larger values improve recall but may introduce ambiguous matches; smaller values improve precision but risk fragmenting trajectories. Only candidate pairs that pass the gating check are forwarded to the appearance-based cost computation (Section 3.4.3), ensuring that the matching stage operates on a geometrically filtered set of candidates.

3.4.3 Appearance-Based Matching Cost

After the spatial gating stage (Section 3.4.2), each admissible detection–track pair is scored using an appearance-based cost derived from the descriptors defined in Section 3.3, and maintained as described in Section 3.4.5. Let $\mathbf{f}_i, \mathbf{a}_j \in \mathbb{R}^d$ denote the ℓ_2-normalized descriptors of detection i and track j, respectively. The cost is computed as the complement of the cosine similarity [60]:

$$\text{cost}_{i,j} = 1 - \mathbf{f}_i^\top \mathbf{a}_j,$$

where $\mathbf{f}_i^\top \mathbf{a}_j$ is the dot product between the normalized vectors, equivalent to the cosine similarity. This measure yields values in the range $[0, 2]$, with lower scores indicating higher visual similarity.

Computation is restricted to detection–track pairs that have passed the motion gating stage, resulting in a sparse cost matrix. This separation of responsibilities preserves the modularity of the tracking pipeline: spatial cues constrain the set of plausible matches, appearance descriptors rank the candidates according to visual compatibility, and acceptance thresholds are applied exclusively in the subsequent assignment stage (Section 3.4.4).

3.4.4 Two-Stage Assignment Strategy

The association between admissible detections and active tracks is formulated as an online one-to-one assignment problem, considering only pairs satisfying the spatial gating constraint (Section 3.4.2). The strategy adopts a two-stage design to balance accuracy and robustness under varying visual conditions, following the paradigm of recent multi-view tracking systems [60, 53, 2].

Stage 1: Appearance-First Matching

In the first stage, a cost matrix $\mathbf{C} \in \mathbb{R}^{M \times N}$ is constructed for all admissible track–detection pairs, where M is the number of active tracks and N is the number of detections passing the spatial gating. Each entry $C_{i,j}$ is the appearance-based cost from Section 3.4.3.

The matrix is solved using the Hungarian algorithm [34], which guarantees an optimal one-to-one assignment and is widely adopted in tracking for its robustness and determinism [60, 53]. While greedy matching can reduce computational cost in dense scenarios, the Hungarian method is generally preferred when accuracy is the priority [60]. After assignment, matches are retained only if their cost satisfies the visual acceptance criterion:

$$C_{i,j} < \tau_{\max},$$

where $\tau_{\max}$ is an empirically tuned threshold rejecting visually ambiguous associations. Pairs failing this check are marked as unmatched and deferred to Stage 2.

Stage 2: Geometric Fallback on Residuals

The second stage sequentially addresses the residuals from Stage 1—tracks and detections left unmatched after appearance-based filtering. For each residual pair, the cost is the Euclidean distance in the BEV plane between the predicted track position and the detection center:

$$\text{cost}_{i,j}^{\text{geo}} = \|\hat{\mathbf{p}}_i - \mathbf{d}_j\|_2.$$

Only pairs within the motion threshold are considered:

$$\text{cost}_{i,j}^{\text{geo}} < \tau_{\text{motion}}.$$

This secondary assignment is solved with the same algorithm used in Stage 1, but previously accepted matches are never reconsidered, preserving the integrity of high-confidence associations.

Threshold roles. τ_{max} controls visual acceptance in Stage 1, ensuring that only visually consistent matches are retained. τ_{motion} constrains the maximum allowed spatial deviation in Stage 2, enforcing geometric plausibility in the absence of reliable appearance cues.

Discussion. This two-stage strategy prioritizes discriminative appearance matching [60, 28], while a motion-based fallback [9, 53, 2] recovers associations in degraded visual conditions such as occlusion, motion blur, or projection errors. By operating sequentially and excluding already validated matches, the fallback minimizes identity switches while improving recall. Although less discriminative than appearance-based matching, it plays a crucial role in maintaining temporal consistency and trajectory continuity when visual evidence alone is insufficient.

3.4.5 Track Management

Once the association stage is completed (Section 3.4.4), each track is updated according to the outcome of the matching process. The underlying state representation, semantic entities, and life-cycle states (TENTATIVE, CONFIRMED, LOST, DELETED) follow the definitions in Appendix A.6. Here we detail the specific update policies adopted in our implementation.

State Update (Motion)

For each matched track–detection pair, the Kalman filter correction step [31, 4] refines the predicted kinematic state using the detection's 2D position in the BEV frame (Appendix A.3) as measurement. This correction reduces drift from the motion model and improves the accuracy of subsequent spatial gating (Section 3.4.2) and motion-based fallback matching (Section 3.4.4). Tracks that remain unmatched are propagated using the prediction step only, preserving a coherent trajectory estimate under temporary occlusion or missed detections.

Feature Maintenance (Appearance)

Matched tracks with valid appearance descriptors are updated according to a temporal aggregation policy. Consistent with prior work [60, 1], two common strategies are:

- **Running average:** incremental update with exponential decay, emphasizing recent observations while retaining long-term memory.
- **Buffer-based:** fixed-size queue storing the most recent descriptors, averaged to form the current appearance profile.

We adopt the buffer-based approach with uniform weighting, which offers robustness to transient noise and a natural forgetting mechanism: outdated or corrupted descriptors are discarded automatically, mitigating *concept drift* [28]. No explicit descriptor quality filtering is applied; while this supports real-time operation, it also increases vulnerability to updates from occluded or low-quality detections. Future extensions could incorporate quality-aware updates based on detection confidence or visibility estimates [14].

Handling Unassigned Tracks and Detections

Unmatched tracks increment their miss counter, triggering state transitions as defined in Appendix A.6. Unassigned detections immediately initialize new TENTATIVE tracks ($N_{\text{init}} = 1$), maximizing responsiveness to new targets. This design choice is well-suited to scenarios with high-quality detections and low false-positive rates. However, it increases sensitivity to transient clutter, potentially generating short-lived tracks. In our implementation, such spurious tracks are quickly pruned when they fail to transition to CONFIRMED.

Discussion. The adopted track management policy balances responsiveness and robustness. Motion updates preserve continuity during brief occlusions, while buffer-based appearance aggregation limits identity drift without unbounded memory. Immediate initialization enables rapid reaction to new targets but trades some robustness in noisy conditions. The absence of descriptor quality control is a deliberate efficiency choice, leaving room for future integration of confidence-based updates.

3.5 Occlusion-Aware Visibility Score

This section introduces a visibility–occlusion-aware score that quantifies the geometric availability of each tracked subject across a calibrated multi-camera setup. The measure is derived exclusively from 3D scene geometry and mutual occlusion reasoning, providing an interpretable and reproducible signal that is independent of dataset-specific visual characteristics. Its motivation stems from the observation that identity preservation in multi-camera multi-object tracking (MC-MOT) often fails under low-visibility conditions, such as partial occlusions or limited cross-view overlap. By isolating the geometric component of visual evidence, the score enables systematic correlation between visibility patterns and association performance (Section 3.4), and offers a principled basis for visibility-aware weighting or gating of appearance features in track management (Section 3.4.5).

The section is organized as follows. Section 3.5.1 motivates the need for an explicit visibility measure and its diagnostic role. Section 3.5.2 details the geometric computation procedure. Section 3.5.3 outlines possible integration points within the pipeline. Section 3.5.4 discusses the main considerations and limitations of the approach.

3.5.1 Motivation and Diagnostic Role

A post-hoc analysis of identity switch (IDSW) events in our experiments (Section 4.6) revealed a recurring failure mode: most mismatches occurred when the target was located near the boundaries of the bird's-eye view (BEV) area, where it was visible from only a small subset of the calibrated cameras, and/or was partially occluded by other subjects. These conditions arise from limited cross-view overlap at the scene periphery and inter-person occlusions in crowded regions. Such geometric constraints are well-known challenges in multi-object tracking [20, 15, 41, 53], yet are often handled implicitly rather than measured explicitly.

To address this, we introduce a visibility–occlusion-aware score that quantifies the visual availability of each tracked subject. For every track at each frame, the score estimates the magnitude of its projected 2D footprint that remains unobstructed in each calibrated camera view, computed as the square root of the number of unoccluded pixels within an idealized bounding box. The computation relies solely on known 3D geometry and mutual occlusions among all tracked subjects (Section 3.5.2). The measure is independent of appearance similarity (Section 3.3) and motion prediction (Section 3.2), isolating the geometric component of visual evidence.

The primary role of this metric is diagnostic: it enables direct correlation of IDSW events with low-visibility conditions, replacing qualitative inspection with a quantitative, frame-level signal. By identifying frames in which the tracker operates on non-informative or ambiguous visual input, it provides actionable insight for (i) interpreting performance degradation in association modules (Section 3.4) and (ii) guiding future integration of visibility-aware update policies in track management (Section 3.4.5).

3.5.2 Geometric Estimation of the Visibility Score

The visibility–occlusion-aware score is computed from first principles of multi-view geometry, relying solely on the calibrated camera parameters and the estimated 3D positions of tracked subjects. The procedure is structured in three stages, reflecting the projection, occlusion reasoning, and score computation steps.

Projection to camera views. Let $\mathbf{p}_t^{(\tau)} \in \mathbb{R}^2$ be the BEV ground-plane position of subject τ at time t, lifted to homogeneous ground-plane coordinates $\tilde{\mathbf{X}}_g^{(\tau)} = [x_t^{(\tau)}, y_t^{(\tau)}, 1]^\top$. For each camera $c \in \mathcal{C}$, the projection into image coordinates is obtained via the homography $\mathbf{H}_c$ (Appendix A.4):

$$\tilde{\mathbf{x}}_t^{(c,\tau)} \sim \mathbf{H}_c \, \tilde{\mathbf{X}}_g^{(\tau)}, \quad \mathbf{x}_t^{(c,\tau)} = \Pi\Big(\tilde{\mathbf{x}}_t^{(c,\tau)}\Big),$$

where $\Pi(\cdot)$ denotes dehomogenization. Around $\mathbf{x}_t^{(c,\tau)}$, we place an *ideal* bounding box $B_t^{(c,\tau)}$ with fixed image-space width w_{box} and height h_{box}, derived from average subject dimensions

in the dataset.

Depth ordering and occlusion masking. For each view c, all subjects are sorted by their Euclidean depth from the camera center:

$$d_t^{(c,\tau)} = \left\| \mathbf{R}_c \mathbf{X}_w^{(\tau)} + \mathbf{T}_c \right\|_2 ,$$

where $\mathbf{X}_w^{(\tau)}$ is the 3D world position of the subject's reference point. If two boxes $B_t^{(c,\tau)}$ and $B_t^{(c,\tau')}$ overlap in the image plane and $d_t^{(c,\tau)} < d_t^{(c,\tau')}$, then τ is considered to occlude τ' in view c. The *occlusion mask* $M_t^{(c,\tau)}$ is defined as the subset of pixels in $B_t^{(c,\tau)}$ covered by any closer subject in the same view.

Visibility score definition. Let $\mathrm{pix}(\cdot)$ denote the number of pixels in a given region. The per-view geometric visibility score is defined as:

$$v_t^{(c,\tau)} = \sqrt{\mathrm{pix}\Big(B_t^{(c,\tau)} \setminus M_t^{(c,\tau)}\Big)}, \tag{3.4}$$

i.e., the square root of the number of unoccluded pixels in the bounding box.

Aggregation across views. The frame-level visibility score for subject τ is obtained by summing over cameras:

$$V_t^{(\tau)} \;=\; \sum_{c \in \mathcal{C}} v_t^{(c,\tau)}, \tag{3.5}$$

where $v_t^{(c,\tau)}$ is the per-view visibility score defined in (3.4).

This purely geometric formulation is independent of appearance features (Section 3.3), motion prediction (Section 3.2), or dataset-specific visual characteristics. It provides an interpretable and reproducible measure of scene geometry and mutual occlusion, directly operationalizing the motivation in Section 3.5.1.

3.5.3 Integration Opportunities within the Tracking Pipeline

The visibility–occlusion-aware score can be incorporated into the tracking pipeline in multiple ways, either as a weighting factor within existing modules or as a gating condition to adapt processing to the reliability of visual evidence. We highlight two primary integration pathways.

Visibility-weighted appearance aggregation. The baseline system aggregates per-view appearance embeddings by uniform averaging (Section 3.3.4), implicitly assuming

equal reliability across camera views. In practice, however, the same subject is often observed from views of uneven quality due to partial occlusion, motion blur, or extreme perspective. This can bias the aggregated descriptor toward distorted or incomplete evidence, degrading downstream matching performance.

Let $\mathbf{f}_t^{(c,\tau)} \in \mathbb{R}^d$ be the ℓ_2-normalized embedding of track τ at time t from camera c, and let $v_t^{(c,\tau)} \geq 0$ be its per-view geometric visibility score from Section 3.5.2, expressed as the square root of the number of visible pixels in the ideal bounding box $B_t^{(c,\tau)}$. The *visibility-weighted* aggregated descriptor is:

$$\bar{\mathbf{f}}_t^{(\tau)} = \frac{\sum_{c \in \mathcal{C}} v_t^{(c,\tau)} \, \mathbf{f}_t^{(c,\tau)}}{\sum_{c \in \mathcal{C}} v_t^{(c,\tau)}}. \tag{3.6}$$

A stricter variant applies a per-view threshold τ_{vis} on $v_t^{(c,\tau)}$, discarding all views where the number of visible pixels (in root scale) falls below this value.

Visibility-based gating in data association. In the online association stage (Section 3.4), matching decisions typically combine geometric proximity with appearance similarity. When a subject has low visibility in the current frame, its per-view embeddings are likely degraded; forcing their use in the appearance-first stage of the two-stage assignment strategy (Section 3.4.4) increases the risk of identity switches or fragmented tracks.

The aggregated visibility score $V_t^{(\tau)}$ is computed as in (3.5). Appearance-based matching is enabled only if $V_t^{(\tau)} \geq \tau_{\text{vis}}$; otherwise, association is performed using geometric cues alone. The same gating rule can be applied in the track management module (Section 3.4.5), updating appearance descriptors only when $V_t^{(\tau)}$ exceeds the threshold.

Integration of the visibility score requires no architectural changes. It can be injected as a weight or gating term in standard modules, enabling incremental adoption in real-time systems with a modest (linear-in-ROIs) computational overhead. Alternatively, it can remain purely diagnostic, supporting post-hoc analysis of occlusion-related failures.

3.5.4 Considerations and Limitations

The visibility–occlusion-aware score $v_t^{(c,\tau)}$ and its aggregated form $V_t^{(\tau)}$ (Section 3.5.2) are defined as purely geometric measures derived from calibrated multi-view projections and mutual occlusion reasoning. While this design yields interpretability and low computational overhead, it imposes intrinsic constraints that limit its descriptive scope.

First, the score models only dynamic occlusions caused by other tracked subjects. Static obstacles such as walls, vehicles, or architectural structures are not represented, so actual line-of-sight interruptions by fixed foreground elements remain unaccounted for. In scenes

with significant static clutter, this can systematically overestimate visibility.

Second, the measure is quality-blind: it encodes geometric availability in terms of the visible pixel count within an idealized bounding box (in root-pixel units) but ignores visual degradations such as motion blur, low illumination, extreme perspective, or truncation at image boundaries. As a result, detections with poor visual evidence may still be scored as nearly fully visible (high $v_t^{(c,\tau)}$), affecting the reliability of appearance-based matching (Section 3.4.3) and descriptor updates in track management (Section 3.4.5).

Third, if a dedicated image-space detector is introduced to validate projected ROIs before appearance extraction, many cases currently flagged as "low visibility" ($v_t^{(c,\tau)} < \tau_{\text{vis}}$) by the geometric score would already be filtered at the detection-validation stage. In such a configuration, the diagnostic value of the visibility score would be reduced, as visually unreliable instances would be excluded upstream, limiting its added benefit for association or feature-weighting.

Finally, in the present system the score is used exclusively for post-hoc diagnostic analysis (Section 3.5.1) and is not integrated into the online association loop (Section 3.4). It is computed from fixed geometric relations without temporal adaptation or learned components, and its operational use would require externally chosen visibility thresholds.

Chapter 4

Experiments

This chapter presents the empirical evaluation of the proposed multi-camera multi-object tracking (MC-MOT) method in calibrated, synchronized environments. A single, reproducible protocol is applied uniformly across datasets. Results are reported and interpreted through controlled comparisons, focusing on identity consistency, localization accuracy, and trajectory persistence.

The chapter is organized as follows. Section 4.1 introduces the benchmarks used, summarizing calibration, scene layout, cross-view coverage, and dataset-specific challenges relevant to reading results. Section 4.2 defines the metrics for identity consistency, localization accuracy, and trajectory persistence, and includes detection-side measures to contextualize tracking scores. Section 4.3 specifies the evaluation protocol and the settings required for reproducibility. Section 4.4 reports comparative tracking results on both benchmarks under the fixed protocol. Section 4.5 isolates the contribution of core modules via an ablation study and considers a visibility-aware variant of the association. Section 4.6 provides a diagnostic error analysis, combining qualitative episodes with a quantitative breakdown that relates error patterns to visibility conditions.

4.1 Datasets

We evaluate on two established multi-camera benchmarks: WildTrack (primary) [11] and MultiViewX (secondary, for generalization) [27]. We chose this pair because both provide calibration and BEV ground-plane evaluation, yet with complementary characteristics: WildTrack offers real-world imagery with strong occlusions and uneven cross-view coverage near BEV borders, whereas MultiViewX provides controlled scenes with denser crowds, higher mean coverage, and stable appearance statistics. This pairing supports fair comparison on standard splits and probes generalization across different coverage regimes. We use the official train/test split and the standard ground-plane evaluation protocol for both

benchmarks [53, 57, 26, 27, 2]; implementation details are deferred to Section 4.3. The following subsections summarize scene layout, camera coverage, calibration availability, and dataset-specific challenges relevant to interpreting our results.

4.1.1 WildTrack (Primary Benchmark)

WildTrack is a real-world multi-camera pedestrian benchmark captured in front of the ETH Zürich main building, with strongly overlapping fields of view. It comprises 7 fixed, synchronized cameras at 1920×1080 px, a ground-plane region of approximately 12×36 m, and an accurate joint calibration (intrinsics, extrinsics, lens distortion) obtained via bundle adjustment on manually curated cross-view correspondences. The ground plane is quantized into a 480×1440 grid (2.5 cm per cell); annotations include ground-plane locations and per-view 2D boxes. Public releases provide synchronized multi-view frames and a densely labeled subset sampled at 2 fps (about 7×400 annotated images), supporting both multi-view detection and ground-plane tracking [11].

Scene and camera layout. The acquisition site is an outdoor plaza with pedestrian traffic, frequent inter-person occlusions, and diverse viewpoints. Cameras are mounted above average head height and arranged to maximize overlap over a predefined area-of-interest (AoI) on the ground plane; the dataset provides per-camera intrinsics/extrinsics and multi-view synchronization, enabling consistent projection/back-projection in the shared BEV frame.

Evaluation region and coverage. All experiments restrict evaluation to the canonical AoI distributed with the dataset and used in prior work [26, 53, 14, 2]. For orientation, we include a BEV panel with camera positions and fields-of-view (FoVs) overlaid on the AoI (Fig. 4.1); this anticipates patterns discussed in the diagnostics (Section 4.6), where identity switches tend to concentrate near BEV borders or in low-overlap zones.

Known challenges. WildTrack is purposely crowded and geometrically challenging: dense interactions, partial occlusions, and frequent truncations at FoV boundaries stress both detection and association. Combined with precise calibration, the evaluator's ground-plane distance thresholds are unforgiving—rewarding metric accuracy and penalizing projection/calibration drift.

4.1.2 MultiViewX (Secondary Benchmark / Generalization)

MultiViewX is a synthetic multi-camera pedestrian dataset introduced alongside MVDet to study multi-view detection under controlled occlusion and crowding [27]. It comprises 6

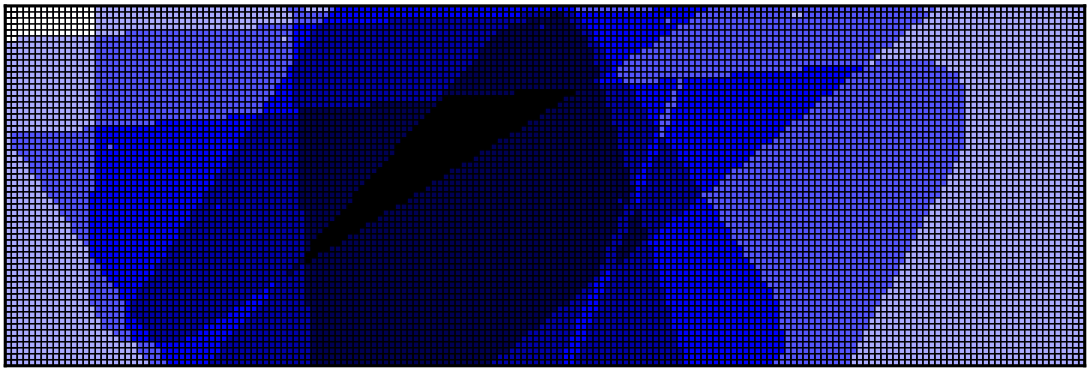


Figure 4.1: WildTrack BEV overview with camera FoVs and evaluation area (AoI). Darker shading indicates higher multi-view coverage. *Reproduced from* [11]. © IEEE.

fixed, synchronized cameras at 1920×1080 px, 400 frames sampled at 2 fps, and a ground-plane region of approximately 16×25 m. The ground plane is quantized into a 640×1000 grid (2.5 cm per cell); annotations include per-frame BEV occupancy maps and per-view 2D boxes. On average the scene contains ≈40 persons per frame, and each BEV location is visible from 4.41 cameras (higher mean coverage than WildTrack).

Scene and camera layout. Rendered in Unity, the scenario exhibits stable illumination and background across frames. Cameras are positioned to produce overlapping fields of view (FoVs) over the evaluation area; the dataset includes full calibration (intrinsics, extrinsics), enabling consistent projection/back-projection in the shared BEV frame.

Evaluation region and coverage. Experiments are restricted to the canonical evaluation area (AoI) distributed with the dataset and used in prior work. Compared to WildTrack, MultiViewX offers fewer cameras but higher average per-location coverage; we revisit this difference in the diagnostics when discussing failure modes in low-overlap zones (Section 4.6).

Characteristics vs. WildTrack. MultiViewX features denser crowds (≈40 persons/frame) and highly consistent calibration and appearance statistics due to its synthetic nature. The higher mean coverage mitigates certain occlusion ambiguities, shifting the balance between motion and appearance cues relative to WildTrack.

4.2 Evaluation Metrics

We report metrics for identity consistency, localization accuracy, and trajectory coverage. IDF1 is used as the primary indicator of identity preservation [47]. CLEAR MOT scores (MOTA, MOTP) and MT/ML are included as complementary context [8]. Because tracking depends on the quality of the detections provided, we also include detection-side metrics (MODA, MODP, Precision, Recall) to contextualize the tracking scores [8]. Protocol choices shared by all metrics—matching in the ground plane, the distance threshold r_{match}, and one-to-one assignment—are fixed in Section 4.3.

The section is organized as follows. Section 4.2.1 defines IDF1 and motivates its role as the main lens for interpreting results. Section 4.2.2 summarizes MOTA, MOTP (meter-valued and normalized forms), and MT/ML. Section 4.2.3 states the detection metrics and their relation to the tracking metrics.

4.2.1 Primary Identity Metric: IDF1

IDF1 is our primary measure of identity preservation. Unlike CLEAR MOT scores, it is an identity-centric metric: it evaluates whether predicted trajectories keep a consistent mapping to ground-truth identities over time (and across cameras), independently of small localization errors as long as matches fall within the evaluation threshold. It is therefore well aligned with the goal of multi-camera MOT, where data association is the principal challenge [47].

Definition (sequence level). Given an entire sequence, the evaluator computes a global one-to-one mapping between predicted IDs and ground-truth IDs that maximizes the number of correctly assigned detections (ID true positives, IDTP). From this mapping:

$$\text{IDP} = \frac{\text{IDTP}}{\text{IDTP} + \text{IDFP}}, \qquad \text{IDR} = \frac{\text{IDTP}}{\text{IDTP} + \text{IDFN}}, \qquad \text{IDF1} = \frac{2\,\text{IDP} \cdot \text{IDR}}{\text{IDP} + \text{IDR}}.$$

Here, IDFP and IDFN count, respectively, assigned predictions that do not belong to the mapped ground-truth identity and ground-truth instances not recovered by their mapped prediction. The global ID mapping is solved once per sequence (optimal assignment), making IDF1 sensitive to identity switches and fragmentation rather than dominated by raw detection recall.

Protocol notes. Matching is performed in the ground plane using the official distance threshold r_{match}, and admissible pairs are assigned one-to-one. The resulting global identity mapping is computed once per sequence to derive IDTP/IDFP/IDFN. Per-dataset settings

(matching space, r_{match}, and assignment algorithm) are reported in Section 4.3. Unless otherwise stated, scores are reported per sequence and aggregated at the dataset level.

Interpretation. Higher IDF1 indicates stable, identity-consistent tracks. Drops in IDF1 typically reflect (i) identity switches at close interactions or partial occlusions, (ii) fragmentation after temporary losses, or (iii) missed detections that break the global ID mapping. In the results (Sections 4.4 and 4.5) we prioritize IDF1 when comparing methods and analyzing variants, using the other metrics as complementary context.

4.2.2 Complementary Metrics: CLEAR MOT and Track Persistence

Beyond IDF1, we report MOTA and MOTP (CLEAR MOT) and the track-level indicators MT/ML to contextualize detection accuracy, localization precision, and trajectory persistence. Definitions follow CLEAR MOT [8] and MOTChallenge reporting conventions [41].

MOTA (accuracy).

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

where GT is the number of ground-truth instances, FN the missed targets, FP the false positives, and IDSW the identity switches (count). MOTA is widely used but is largely driven by FN (recall), hence it is less specific to identity preservation than IDF1 [8]. Note that MOTA can be negative in heavily error-dominated regimes.

MOTP (localization). Following CLEAR MOT [8], we compute the mean ground-plane localization error for accepted matches:

$$\overline{d} \;=\; \frac{\sum_t \sum_i d_i^{(t)}}{\sum_t c_t},$$

where $d_i^{(t)}$ is the BEV distance for the i-th matched pair at time t, and c_t is the number of matches at t. All distances are measured in meters in the BEV plane and are computed only over accepted matches (hence $d_i^{(t)} \le r_{\text{match}}$).

The classical form reports $\text{MOTP}_{\text{m}} = \overline{d}$ in meters (lower is better). To align with common reporting in multi-camera BEV tracking, we also provide a normalized, higher-is-better unitless score:

$$\text{MOTP}_{\text{n}} \;=\; 1 - \frac{\overline{d}}{r_{\text{match}}} \;\in\; [0,1],$$

where r_{match} is the official BEV matching radius (specified in Section 4.3). By definition, $\text{MOTP}_{\text{n}} = 1$ denotes zero mean error; values are clipped at 0 if $\overline{d} > r_{\text{match}}$.

MT/ML (trajectory persistence). Mostly tracked (MT) and mostly lost (ML) quantify how much of each ground-truth trajectory is correctly recovered. Let $r(\tau)$ be the fraction of frames of GT trajectory τ that are correctly matched under the evaluation protocol. A trajectory is counted as MT if $r(\tau)$ exceeds the standard coverage threshold (we adopt the MOTChallenge convention of 80%), and as ML if $r(\tau)$ falls below the complementary threshold (20%); dataset-specific deviations, if any, are stated in Section 4.3 [41]. These indicators capture sequence-level stability and fragmentation.

Reporting conventions. In cross-paper tables we report {IDF1, MOTA, MOTP, MT, ML} , which are consistently available for WildTrack and MultiViewX. IDSW is provided only for our method and ablation variants. Scores are computed over the full sequence for both datasets (each provides a single annotated sequence). For multi-sequence benchmarks, we would report per-sequence scores and their unweighted (macro) average. Operational details that affect all metrics (matching in BEV, distance threshold r_{match}, and one-to-one assignment algorithm) are specified in Section 4.3.

4.2.3 Detection Metrics

For completeness, we summarize the detection metrics used to contextualize tracking performance. Definitions follow the same conventions as for CLEAR MOT; evaluation specifics (matching space and thresholds) are fixed in Section 4.3.

MODA (accuracy).

$$\text{MODA} \;=\; 1 \;-\; \frac{\text{FN} + \text{FP}}{\text{GT}}.$$

It reflects the balance of misses and false positives at the detection stage. In relation to tracking,

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

so the gap (MODA − MOTA) isolates association losses.

MODP (localization). Let $d_i^{(t)}$ be the BEV distance of the i-th accepted match at time t, and let r_{det} denote the detection matching radius specified by the protocol. We report a normalized, higher-is-better score

$$\text{MODP} \;=\; \frac{1}{\sum_t c_t} \sum_t \sum_i \Big(1 - \frac{d_i^{(t)}}{r_{\text{det}}}\Big) \;\in\; [0,1]$$

(expressed in %). The classical meter-valued form corresponds to the mean localization error $\overline{d} = \frac{\sum_t \sum_i d_i^{(t)}}{\sum_t c_t}$, with $\text{MODP} = 1 - \overline{d}/r_{\text{det}}$.

Precision and Recall.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \qquad \text{Recall} = \frac{\text{TP}}{\text{GT}}.$$

The FP/FN balance (Precision/Recall, hence MODA) conditions MT/ML and the attainable identity consistency (IDF1), while localization quality (MODP) relates to the spatial accuracy summarized by MOTP.

4.3 Experimental Setup

This section defines the experimental conditions under which we evaluate the proposed multi-camera multi-object tracking (MC-MOT) system in calibrated environments, with the aim of enabling fair and reproducible comparisons. We fix the evaluation rules on the ground plane (matching space and assignment), specify which data segments enter scoring, and record the operational choices actually consumed by the tracker at runtime. Algorithmic derivations and design rationales are presented in Chapter 3; here we document the concrete settings used in the experiments.

The section is organized as follows. Section 4.3.1 states the evaluation protocol; Section 4.3.2 describes the detector used to generate tracker inputs and the interface between detection and tracking; Section 4.3.3 details the model-selection criterion and search strategy under the compute budget; Section 4.3.4 records per-module runtime configurations; Section 4.3.5 consolidates all numeric settings in a single table; and Section 4.3.6 anchors the software environment for reproducibility.

4.3.1 Evaluation Protocol

We adopt the official ground–plane protocol of the benchmarks to ensure fair and reproducible comparisons [26, 53, 14, 2, 57]. This subsection specifies how predictions and ground truth are matched and which data segments enter the evaluation.

Matching configuration. Associations are computed in the BEV metric space using Euclidean distances on the ground plane. For tracking metrics we enforce a fixed matching radius $r_{\text{match}} = 1.0\,\text{m}$; for detection metrics we adopt a tighter radius $r_{\text{det}} = 0.5\,\text{m}$. Candidate pairs with distance greater than the applicable radius are not eligible. Within each frame, matches are resolved by optimal one-to-one bipartite assignment (Hungarian algorithm) [34] under the corresponding gate.

Evaluation window and aggregation. For both benchmarks we follow the official split [11][27]: the first 360 annotated frames form the training portion and the last 40 annotated

frames form the test portion. Since each dataset comprises a single sequence, all metrics are computed end-to-end on the entire test segment (no resets or windowing), and this protocol is applied consistently across all experiments.

4.3.2 Detector and Preprocessing

We use MVDeTr [26] as an off-the-shelf multi-view detector to produce the inputs consumed by the tracker. This subsection specifies the inference configuration, the per-frame data flow from synchronized multi-view images to BEV predictions, and the absence of additional preprocessing on our side.

Detector identity and configuration. Inference runs with the public implementation released by the authors [25]. For both benchmarks we adopt the weights distributed with the repository and the default inference configuration associated with each checkpoint; no fine-tuning or parameter changes are introduced.

Interface and data flow. At each time instant t, the detector receives the synchronized multi-view images as provided by the dataset, in their released format and ordering. Any resizing or normalization is handled internally by the official pipeline for the target dataset. For each frame, the detector returns a set of predicted locations on the canonical BEV grid (see Appendix A.3), represented as pairs (x, y).

Preprocessing and protocol alignment. No additional preprocessing is applied: no score thresholding, non-maximum suppression, merging, or ROI filtering beyond what is already handled within the detector. The (x, y) predictions are forwarded to the tracker as produced. Detector inference is executed on the same frames and on the same test segment of the official split (360/40) used for scoring, ensuring consistency with the evaluation protocol in Section 4.3.1.

4.3.3 Tuning Protocol

We specify how operating points are selected under a constrained compute budget, ensuring reproducibility and preventing test leakage. Both benchmarks provide a single annotated sequence; selection is performed on the training segment and the chosen configuration is kept fixed when scoring on the held-out test segment.

Selection criterion and data. Hyperparameters are selected by maximizing IDF1 on the official train segment (first 360 frames). In case of ties, we prefer lower localization error (MOTP) and, secondarily, fewer identity switches. The selected configuration is then

frozen and used unchanged on the test segment (last 40 frames). For efficiency, all hyperparameters chosen on WildTrack—both tuned and fixed—are reused as-is on MultiViewX; no information from test frames is used during selection.

Parameters tuned The appearance module (multi-view back-projection, per-camera ROI extraction, and OSNet inference for each visible view of each detection, followed by multi-view fusion and memory updates) dominates wall-clock time on the 360-frame training split: per frame, the cost scales with the number of cameras and detections and entails multiple forward passes through the embedding network. Under our compute budget this makes a full sweep prohibitively slow. We therefore calibrate only the motion-related hyperparameters on a motion-only ablation of the tracker (see Section 4.5), which preserves the same spatial gating and lifecycle logic as the full model while replacing appearance-based matching with purely geometric association. Concretely, we tune the BEV gating radius τ and the miss tolerance N_{miss} on the training segment; the selected values are then reused unchanged in the full model.

Heuristic settings in the full model. In the full two-stage model, the appearance-acceptance threshold τ_{max} is fixed heuristically (value in Table 4.1) to act as a tight visual acceptance at Stage 1, discouraging ambiguous matches and prioritizing precision. The geometric threshold used in Stage 2 is tied to the gate, $\tau_{\text{motion}} = \tau$, enforcing a coherent spatial criterion between candidate filtering and fallback. As future work, it would be possible to include both τ_{max} and τ_{motion} in the search space. For the visibility-enabled variant, we also introduce a visibility switching threshold τ_{vis}, used to enable/disable appearance matching based on the estimated visibility; τ_{vis} is set heuristically (value in Table 4.1) and its diagnostic role and limitations are discussed in Section 4.6.

Search ranges. We explore τ in [30,70] with step 10 (BEV grid units) and N_{miss} in [2,10] with step 2 (frames). Final selected values are reported in Table 4.1. Ablations reuse the full-model settings for applicable parameters; non-applicable ones are disabled without ad-hoc retuning.

4.3.4 Module Configurations (Operational Settings)

This subsection records the runtime choices actually used by the tracker; the algorithmic rationale is provided in Section 3.4. All numeric settings (thresholds, Kalman filter (KF) noises and covariances, buffer length, descriptor dimensionality, lifecycle counters) are consolidated in Table 4.1.

Motion Model (BEV Kalman)

State: constant-velocity kinematics on the ground plane with 2D BEV position measurements from associated detections. The filter executes a predict step every frame and an update step when a measurement is available; unmatched tracks are propagated by prediction while incrementing a miss counter. New tracks are initialized from a single BEV measurement with large initial velocity uncertainty; the initial covariance P_0, process noise Q, measurement noise R, and the frame time step Δt are specified in Table 4.1. Candidate generation for association uses Euclidean motion gating in BEV with radius τ (Table 4.1).

Appearance Embedding and Multi-View Fusion

For each detection, the BEV location is back-projected onto every camera to extract a square patch; views whose projection falls outside the field of view are skipped. Patches are resized and normalized with the standard ImageNet preprocessing [33]. Descriptors are produced by an OSNet-x1.0 [69] backbone via `torchreid`[68] in inference-only mode (pretrained weights, no fine-tuning) and ℓ_2-normalized. Multi-view fusion for a detection is the uniform average of valid per-view descriptors, yielding a single descriptor per detection. If no valid crop is available for a detection, a null descriptor is assigned, which effectively prevents acceptance at the appearance stage while allowing recovery via the geometric fallback. Each track maintains a fixed-length FIFO buffer of recent descriptors with uniform aggregation. Descriptor dimensionality, input size, and buffer length are reported in Table 4.1.

Online Association

Association proceeds in two stages. A spatial gate in BEV with radius τ first filters admissible pairs. Stage 1 solves a one-to-one assignment with the Hungarian algorithm using appearance costs and retains matches that satisfy the acceptance threshold $\tau_{\max}$. Stage 2 addresses residuals with a geometric fallback: the cost is the BEV Euclidean distance between the predicted track position and the detection, and assignment again uses the Hungarian solver; matches accepted in Stage 1 are not reconsidered. In the full model, the motion-fallback threshold is tied to the gate, $\tau_{\text{motion}} = \tau$ (see Section 4.3.3). All thresholds appear in Table 4.1.

Track Management (Lifecycle)

Unassigned detections initialize new TENTATIVE tracks immediately (confirmation threshold N_{init} in Table 4.1). Matched tracks undergo a Kalman correction and update their appearance buffer; unmatched tracks keep the predicted state and increment their miss

counter. Tracks are deleted when the miss counter exceeds N_{miss} (Table 4.1). Track IDs are consistent within each sequence and are not recycled after deletion.

4.3.5 Hyperparameters Summary

This subsection consolidates all numeric operational settings used in our experiments. All values are collected in Table 4.1, which serves as the single source of truth. Values refer to *WildTrack*; unless otherwise noted, the same values are reused on *MultiViewX* without re-tuning.

Table 4.1: Operational hyperparameters actually used at runtime. Values refer to WildTrack and are reused on MultiViewX unless otherwise indicated. Columns list the symbol and units; "grid units" denote the BEV discretization, meters refer to ground–plane distances, "items/frames" indicate buffer lengths or frame counts, and cosine distance is used for $\tau_{\max}$. The table consolidates evaluation, BEV Kalman, association, appearance, lifecycle, and the optional visibility threshold.

Parameter	Symbol	Units	Value
Evaluation			
Matching radius (ground plane)	r_{match}	m	1.0
Motion model (Kalman, BEV)			
Frame time step	Δt	frame	1
Initial covariance (diagonal)	$\text{diag}(P_0)$	state units^2	1000
Process noise (diagonal)	$\text{diag}(Q)$	state units^2	1
Measurement noise (diagonal)	$\text{diag}(R)$	measurement units^2	1
Online association			
BEV gating radius	τ	grid units	40.0
Geometric fallback threshold	τ_{motion}	grid units	40.0
Appearance acceptance (Stage 1)	$\tau_{\max}$	cosine distance	0.2
Appearance			
Descriptor dimensionality	d	dim	512
Embedding input size	$H \times W$	px	256×128
Track descriptor buffer length	K	items	30
Track management (lifecycle)			
Initialization threshold	N_{init}	items	1
Miss tolerance (deletion)	N_{miss}	frames	10
Visibility variant			
Visibility switching threshold	τ_{vis}	—	200.0

4.3.6 Implementation Notes

We fix here the software environment used in all experiments; full configuration files and scripts are available in the repository at the declared commit [39].

Software stack. Python 3.10 (conda). Core libraries: PyTorch 2.2 [43] with `pytorch-cuda` 12.1 (torchvision 0.17), NumPy 1.24 [22], OpenCV, SciPy[56], Pandas [40], Pillow. External packages: torchreid (OSNet-x1.0 backbone) and `motmetrics` for evaluation.

Hardware (GPU). NVIDIA GeForce GTX 1080, driver 550.120; the installed driver (CUDA 12.4) is compatible with the PyTorch runtime `pytorch-cuda` 12.1 used in our environment.

Repository and configs. All configs and scripts used to produce the results are frozen at commit `4224659` of the thesis repository [39]; this commit includes the environment specification and execution scripts referenced throughout the Experimental Setup.

4.3.7 Detection Quality (MVDeTr)

We report the aggregate detection performance of the external detector used throughout the experiments (MVDeTr [26, 25]). Scoring follows the detection protocol in Section 4.3.1. The aggregate values are reported in Table 4.2. These figures are not part of our contribution; they provide context for the tracking results in Section 4.4.

Table 4.2: Detection quality of MVDeTr on WildTrack and MultiViewX. Scores are computed end-to-end on the official test split. These values provide context for the tracking results and are not part of our contribution. Metrics: MODA, MODP, Precision, Recall (percentages); FN and FP are absolute counts.

	MODA↑	MODP↑	Precision↑	Recall↑	FN↓	FP↓
WildTrack	92.02	82.29	96.50	95.48	43	33
MultiViewX	93.64	91.42	99.57	94.04	89	6

4.4 Tracking Results

We present comparative tracking results on WildTrack and MultiViewX under the official ground-plane protocol (Section 4.3.1). Scores are reported in Table 4.3; detection quality used as context is summarized in Table 4.2. Interpretation follows the metrics in Section 4.2: IDF1 is the primary lens; for MOTA we refer to the headroom given by MODA; MOTP and MT/ML provide localization and persistence context. We first discuss WildTrack and then MultiViewX, and conclude with a brief cross-dataset take-home; ablations and diagnostics are deferred to Sections 4.5 and 4.6.

WildTrack On WildTrack (Table 4.3), our method ranks second in IDF1 and MOTA, ties for first in MT, and remains competitive on ML. As the primary identity metric, IDF1 is close to the best method (Attention-Aware [2]); this is consistent with the behavior of our two-stage association (appearance first, geometric fallback), which maintains identities through close interactions and brief dropouts.

For MOTA, reading the results against the detector context (Table 4.2, Section 4.3.7) shows that our score lies only ≈ 0.32 percentage points below MVDeTr's MODA (92.02%), indicating limited association losses beyond FP/FN (cf. Section 4.2.3). This headroom view also clarifies cross-method differences: EarlyBird reports on WildTrack a detector MODA of

Table 4.3: Tracking results on WildTrack and MultiViewX. Scores are computed end-to-end on the official test split. Methods include published baselines and ours. Reading: IDF1 is the primary lens; MOTA, MOTP, and MT/ML provide context. Conventions: best in **bold**, second best underlined.

	IDF1↑	MOTA↑	MOTP↑	MT↑	ML↓
WildTrack					
REST [14]	85.7	81.6	81.8	79.4	**4.7**
EarlyBird [53]	92.3	84.9	86.6	78.0	4.9
BEV-SUSHI [57]	95.6	92.6	**93.7**	80.5	7.3
Attention-Aware [2]	**96.1**	**92.7**	88.8	**87.8**	4.9
Ours	95.7	91.7	89.8	**87.8**	4.9
MultiViewX					
EarlyBird [53]	82.4	88.4	86.2	82.9	1.3
Attention-Aware [2]	85.7	91.3	86.8	89.5	1.3
Ours	**92.0**	**92.2**	**95.0**	**93.4**	**0.0**

91.2% and a tracking MOTA of 89.5% (gap = 1.7 pp) [53], while ReST employs MVDeTr yet trails our scores [14], pointing to tracker-level differences. The role of the detector is further illustrated by BEV-SUSHI, whose tracking improves when fed with stronger detections [57]; as future work, it would be natural to evaluate our tracking pipeline with alternative detectors (e.g., EarlyBird or BEV-SUSHI) within the dataset's canonical BEV.

Regarding MOTP, in this ground-plane setting it is largely detector-limited (computed on accepted matches), so the remaining margin to BEV-SUSHI reflects localization precision rather than association. Finally, MT/ML confirm trajectory persistence (high MT, low ML), attributable to the combination of reliable detections and our miss-tolerant lifecycle with geometric fallback. In summary, with MVDeTr off-the-shelf we consume nearly all available headroom on WildTrack; the principal remaining margin is localization precision, not association.

MultiViewX On MultiViewX (Table 4.3), our method ranks first across all reported metrics. As the primary identity indicator, IDF1 confirms that the two-stage association (appearance followed by geometric fallback) generalizes well from WildTrack to MultiViewX.

For MOTA, reading the scores against the detector context (Table 4.2, Section 4.3.7) shows that our result is only $\approx$ 1.44 percentage points (pp) below MVDeTr's MODA (93.64%), indicating limited association loss beyond FP/FN (cf. Section 4.2.3). In comparison, EarlyBird reports on MultiViewX a detector MODA of 94.2% and a tracking MOTA of 88.4% (gap = 5.8 pp) [53], suggesting a larger association overhead under the same

ground-plane protocol. The detector's influence has already been discussed on WildTrack; here it chiefly serves to quantify headroom, with our pipeline expending little of it.

We also lead in MOTP. In this ground-plane setting, MOTP is largely detector-limited (computed on accepted matches), so the advantage reflects precise localization of the accepted hypotheses rather than differences in association. Finally, MT/ML indicate stable trajectory coverage (very high MT with zero ML), attributable to accurate detections and a miss-tolerant lifecycle with geometric fallback. In summary, on MultiViewX the combined pipeline (MVDeTr plus our tracker) consumes most of the available headroom while maintaining superior identity consistency and persistence relative to EarlyBird [53] and Attention-Aware [2].

Across datasets, association efficiency is high: the MODA–MOTA gap is 0.32 pp on WildTrack and 1.44 pp on MultiViewX, versus 1.7/5.8 pp reported by EarlyBird [53]; this indicates a low overhead due to association under the ground-plane protocol (cf. Table 4.2 and Section 4.2.3). Our strengths are identity and persistence (IDF1, MT/ML), while the remaining margin sits in localization (MOTP), which in this setting is largely detector-limited (computed on accepted matches). The role of the detector remains practical: BEV-SUSHI improves when fed with stronger detections [57]; as a controlled extension, our tracker could be evaluated with alternative detectors within the dataset's canonical BEV to quantify detector sensitivity. Notably, on MultiViewX we reuse hyperparameters tuned on WildTrack and still achieve state of the art, suggesting robust gating/lifecycle and two-stage association. A candid limitation is scope: WildTrack and MultiViewX are single, short sequences (360/40 split); performance on longer or more varied scenes may favor long-range association (e.g., sliding-window/GNN designs) [57]. Ablations and diagnostics are provided in Sections 4.5 and 4.6.

4.5 Ablation Study

This section evaluates how the online association module impacts tracking when the detector and protocol are fixed (MVDeTr; ground–plane rules in Section 4.3). We compare two ablations—motion-only (purely geometric association) and appearance-only (purely visual association)—against the proposed full two-stage pipeline, and we also assess a visibility-aware policy that conditionally activates the visual stage based on a per-candidate visibility estimate. In reading the results, we use IDSW as the direct indicator of association quality; all other metrics follow Section 4.2.

The section is organized as follows. Section 4.5.1 reports the Online Association Ablations (motion-only, appearance-only) against the full two-stage reference. Section 4.5.2 presents the Visibility-Aware Variant and its effect on identity, accuracy, and persistence.

4.5.1 Online Association Ablations

We isolate the effect of the online association policy under fixed detector and protocol: MVDeTr [26] provides detections (Table 4.2) and scoring follows the ground–plane rules in Section 4.3.1. We compare three pipelines that differ only in the association module—*Motion-only*, *Appearance-only*, and the *Full two-stage* reference—while keeping Kalman filter, lifecycle, and all unrelated settings unchanged (Section 4.3). The goal is to verify whether using appearance selectively reduces identity switches and how much a geometric fallback recovers residual pairs without harming persistence. In reading the results (Table 4.4), we use IDSW as the most direct proxy of association quality, with IDF1/MOTA, MOTP, and MT/ML as complementary context (Section 4.2).

Variants and Mechanics

Motion-only (geometric).
Candidates are gated in BEV with radius τ; the cost is the Euclidean distance on the ground plane; assignment is one-to-one (Hungarian algorithm); pairs are accepted if their distance is $\leq \tau$. No appearance features are computed. Operatively, this matches the geometric fallback of the full pipeline. It is efficient and deterministic, but prone to swaps in crowded interactions or local ambiguities.

Appearance-only (visual).
The same BEV gate τ defines admissible pairs; the cost is the cosine distance between descriptors; the Hungarian algorithm solves one-to-one assignment; matches are accepted if $\leq \tau_{\max}$. Residuals are not recovered (no geometric pass). This maximizes visual precision when crops are reliable, but may miss valid associations under partial views or occlusions.

Full two-stage (reference).
Admissible pairs are gated by τ; Stage 1 solves visual assignment with acceptance threshold $\tau_{\max}$; Stage 2 applies a geometric fallback on residuals with threshold $\tau_{\text{motion}}{=}\tau$. The rationale is to exploit appearance where it is informative (reducing ambiguity in crowding) and to recover remaining pairs geometrically. Formal details are given in Section 3.4.

WildTrack: Results & Interpretation

From Table 4.4, the full two-stage achieves the best balance: low IDSW with high IDF1/MOTA and solid MOTP. Appearance-only is almost overlapping, but without geometric recovery a small fraction of pairs remains unassigned or fragments, yielding +1 IDSW and a marginal -0.1 pp drop in MOTA. Motion-only is clearly worse in identity and accuracy (IDSW increases markedly, IDF1/MOTA decrease); MOTP also degrades substantially (-5.8 pp vs.

Table 4.4: Online association ablations on WildTrack and MultiViewX. Scores are computed end-to-end on the official test split. Variants differ only in the association module: Motion-only, Appearance-only, and Full two-stage. Conventions: best in **bold**, second best underlined

	IDF1↑	MOTA↑	MOTP↑	MT↑	ML↓	IDSW↓
WildTrack						
Motion-only	90.0	90.6	84.0	**87.8**	**4.9**	18
Appearance-only	95.6	91.6	**89.8**	**87.8**	**4.9**	8
Full two-stage	**95.7**	**91.7**	**89.8**	**87.8**	**4.9**	**7**
MultiViewX						
Motion-only	89.8	91.0	95.0	**93.4**	**0.0**	40
Appearance-only	91.8	92.0	**95.1**	**93.4**	**0.0**	24
Full two-stage	**92.0**	**92.2**	95.0	**93.4**	**0.0**	**22**

the full), consistent with the definition as a mean over accepted matches—purely geometric association tends to accept pairs that are, on average, farther (though within the 1 m gate).

Causally, appearance stabilizes identities during close interactions (reducing swaps), while the geometric fallback recovers residuals that the visual stage rejects, benefiting MOTA without harming persistence. With the lifecycle fixed, MT/ML remain essentially unchanged across variants; here MOTP reveals the indirect effect of the association policy through the distribution of accepted match distances.

MultiViewX: Results & Interpretation

On MultiViewX the full two-stage again leads across metrics, with the lowest IDSW. Appearance-only is practically aligned; the geometric fallback consolidates a small margin—IDSW −2 and MOTA +0.2 pp. Motion-only worsens IDF1/MOTA and increases IDSW, whereas MOTP stays high—under precise detections, even geometric matches remain close to their ground-truth counterparts within the gate.

In a cleaner scene the visual stage already resolves most associations; the fallback adds limited but consistent robustness. MOTP is largely detector-driven in this setting; policy-induced differences emerge only when match acceptance drifts toward peripheral pairs, which is not pronounced here. For detector context, see Table 4.2.

4.5.2 Visibility-Aware Variant

We introduce a visibility-aware policy that conditionally activates the Stage 1 (appearance) association based on a per-candidate visibility estimate (see Section 3.5.2). The motivation

is straightforward: in crowding and partial occlusions, multi-view crops can be weak or misleading; forcing appearance costs in such regimes may increase identity switches. The two-stage architecture and all thresholds/modules remain unchanged; we intervene only on when appearance is applied.

Mechanics

For each admissible pair, if the visibility score exceeds a threshold τ_{vis}, Stage 1 (appearance) runs with the usual acceptance threshold $\tau_{\max}$; otherwise Stage 1 is skipped and the pair is handled by the geometric fallback (Stage 2) with $\tau_{\text{motion}}{=}\tau$. Invariants: the BEV gate τ, the geometric threshold $\tau_{\text{motion}}{=}\tau$, the appearance threshold $\tau_{\max}$, the lifecycle and Kalman filter, and the scoring protocol (Section 4.3.1) all remain the same. The policy thus changes only the activation condition for Stage 1: use appearance when the visual signal is likely reliable; otherwise defer to geometry.

WildTrack: Results & Interpretation

Relative to the full two-stage (Table 4.5), the variant reduces identity switches and slightly improves accuracy: IDSW 4 vs 7 (-3), MOTA 92.0 vs 91.7 ($+0.3$ pp), IDF1 95.8 vs 95.7 ($+0.1$), with MOTP 89.4 vs 89.8 (-0.4 pp) and MT/ML unchanged. A cautious reading is that, under occlusions/crowding, skipping appearance when crops are unreliable avoids misleading visual costs (fewer swaps), while recovery through geometry preserves coverage; the small MOTP decrease is consistent with a slightly higher reliance on geometric matches (on average less central, but within the 1 m gate).

MultiViewX: Results & Interpretation

On MultiViewX, the same policy turns off appearance where it would help: IDSW 28 vs 22 ($+6$), MOTA 91.8 vs 92.2 (-0.4 pp), IDF1 91.8 vs 92.0 (-0.2), MOTP 94.8 vs 95.0 (-0.2 pp), with MT/ML unchanged. In a cleaner scene, visibility gating can underuse appearance and thus miss associations that the full two-stage resolves; differences in MOTP remain small, being largely detector-driven.

Takeaway. The visibility-aware policy should be read as an activation filter for appearance: use visual features when the estimated visibility suggests they are reliable; otherwise rely on the geometric fallback. Across our two datasets it yields small gains on WildTrack and small losses on MultiViewX, so we avoid general claims and treat it as a scene- and setting-dependent option rather than a new default. The behaviour is sensitive to the threshold τ_{vis}, which here was fixed heuristically and, for WildTrack, informed in part by qualitative inspection that included test frames (see Section 4.6); the gains on WildTrack

Table 4.5: Visibility-aware vs Full two-stage on WildTrack and MultiViewX. Scores are computed end-to-end on the official test split. The visibility-aware policy activates appearance only when the estimated visibility exceeds τ_{vis}; all other components and thresholds remain unchanged. Conventions: best in **bold**.

	IDF1↑	MOTA↑	MOTP↑	MT↑	ML↓	IDSW↓
WildTrack						
Full (two-stage)	95.7	91.7	**89.8**	**87.8**	**4.9**	7
Visibility-aware	**95.8**	**92.0**	89.4	**87.8**	**4.9**	**4**
MultiViewX						
Full (two-stage)	**92.0**	**92.2**	**95.0**	**93.4**	**0.0**	**22**
Visibility-aware	91.8	91.8	94.8	**93.4**	**0.0**	28

may therefore be optimistic. Applying the same value to MultiViewX led to a slight degradation, indicating dataset sensitivity. In practice, τ_{vis} should be tuned on the training split of each dataset; absent such tuning, the plain two-stage policy remains the safer choice. Finally, effectiveness depends on the fidelity of the visibility score itself: the geometric estimator does not model static occluders and can overestimate visibility in scenes with fixed clutter, which can bias activation decisions (see Section 3.5.4).

4.6 Diagnostic Error Analysis

Identity switches are diagnosed through two complementary views. Section 4.6.1 offers a visual intuition ahead of the quantitative analysis, presenting two episodes: (i) a protocol-induced switch in which, between consecutive frames, the evaluator changes the ground-truth–to-prediction pairing while track IDs remain stable; and (ii) a genuine break in identity continuity due to track fragmentation when the subject is occluded or not visible in most cameras. Section 4.6.2 provides the quantitative counterpart, relating error frequencies to the occlusion-aware visibility score (see Section 3.5).

4.6.1 Qualitative Analysis

This subsection provides a visual intuition for how ID switches arise, ahead of the quantitative analysis. We show per-camera crops with a fixed window across the shown instants, centred on the interaction area. Two complementary cases are presented: *Case A*, where the switch is induced by the evaluator's per-frame matching while track IDs remain stable; and *Case B*, where the switch stems from track fragmentation when the subject is occluded or not visible in most cameras.

Case A — ID Switch with Stable Track IDs. Figure 4.2 shows two consecutive instants ($t-1$, t) for five active views (C1, C2, C3, C5, C6); in the remaining cameras the subjects are out of view. Panels display per-camera crops around the interaction area, with a fixed window across the two instants, for readability. Bounding boxes are the tracker's per-view reprojected ROIs; colours indicate the ground-truth identity assigned by the evaluator in each frame (Section 4.3.1). In this window only C2 and C3 provide usable ROIs, while the other views do not contribute; the multi-view coverage is therefore low and asymmetric in both instants (about 2/7 for one subject and 1/7 for the other), with the pedestrians very close in the active views.

Despite the reduced coverage, the online association is stable: the same track IDs remain attached to the same detections from $t-1$ to t, with no re-attach or re-identification. Under the ground-plane evaluation protocol, matches are computed per frame by one-to-one Hungarian assignment within the BEV matching radius. With two ground-truth instances and two predictions at small ground-plane separation, the per-frame assignment at time t selects the crossed pairing as the slightly better solution; the ground-truth–to-prediction mapping thus differs from that at $t-1$ and an ID switch is counted. This is a change in the per-frame assignment, not a failure of the online association.

Case B — ID Switch from Track Fragmentation. Figure 4.3 shows two instants t=391 and t=393 on four active views (C1–C4); in the remaining cameras the subject is out of view. The intermediate frame t=392 is omitted because the subject is missed. Panels display per-camera crops with a fixed window across the two instants. Bounding boxes are the tracker's per-view reprojected ROIs, and colours/labels denote the tracker-assigned track IDs. At t=391 C3 is the only truly informative view; C2 is partial; C4 is distant and fully occluded; C1 is occluded. At t=393 C3 becomes more occluded, C2 remains partial, and C1/C4 stay non-informative. As a result, the effective multi-view coverage drops from one informative view plus one partial view to at most a single partial view; appearance cues are therefore unreliable.

On the tracker side, the online association does not preserve the previous track on the same person: at t=393 the detection of that physical subject is not assigned to the former track i and is instead assigned to another track j. This is a fragmentation: i loses continuity, j continues the same person. Under the ground-plane evaluation protocol (Section 4.3.1), the per-frame BEV assignment then matches ground truth to track IDs and records that the same ground-truth identity is paired with i at t=391 and with $j \neq i$ at t=393; an ID switch is counted. This contrasts with Case A (protocol-driven mapping change with stable track IDs) and exemplifies an association failure when the subject is occluded or not visible in most cameras.

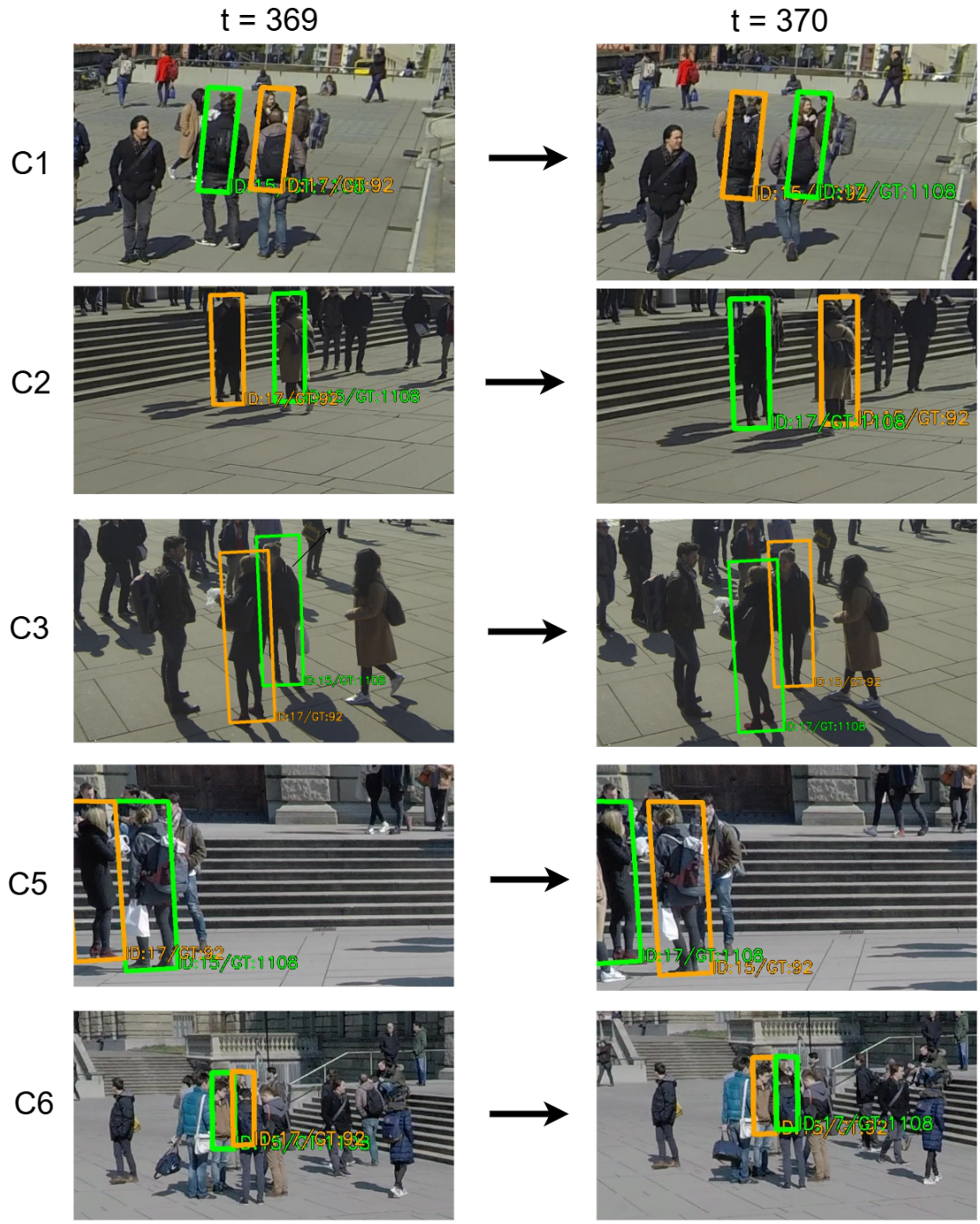


Figure 4.2: Evaluator-induced ID switch with stable track IDs (WildTrack, frames $t = 369$–370). Two consecutive instants are shown for five active views; in the remaining cameras the subjects are out of view. Panels display per-camera crops with a fixed window across $t-1$ and t. Bounding boxes are the tracker's reprojected per-view ROIs; colours denote the ground-truth identity assigned by the evaluator in each frame. Under the ground-plane protocol, per-frame one-to-one Hungarian assignment in BEV selects at t the crossed pairing while track IDs remain unchanged, so the evaluator counts an IDSW.

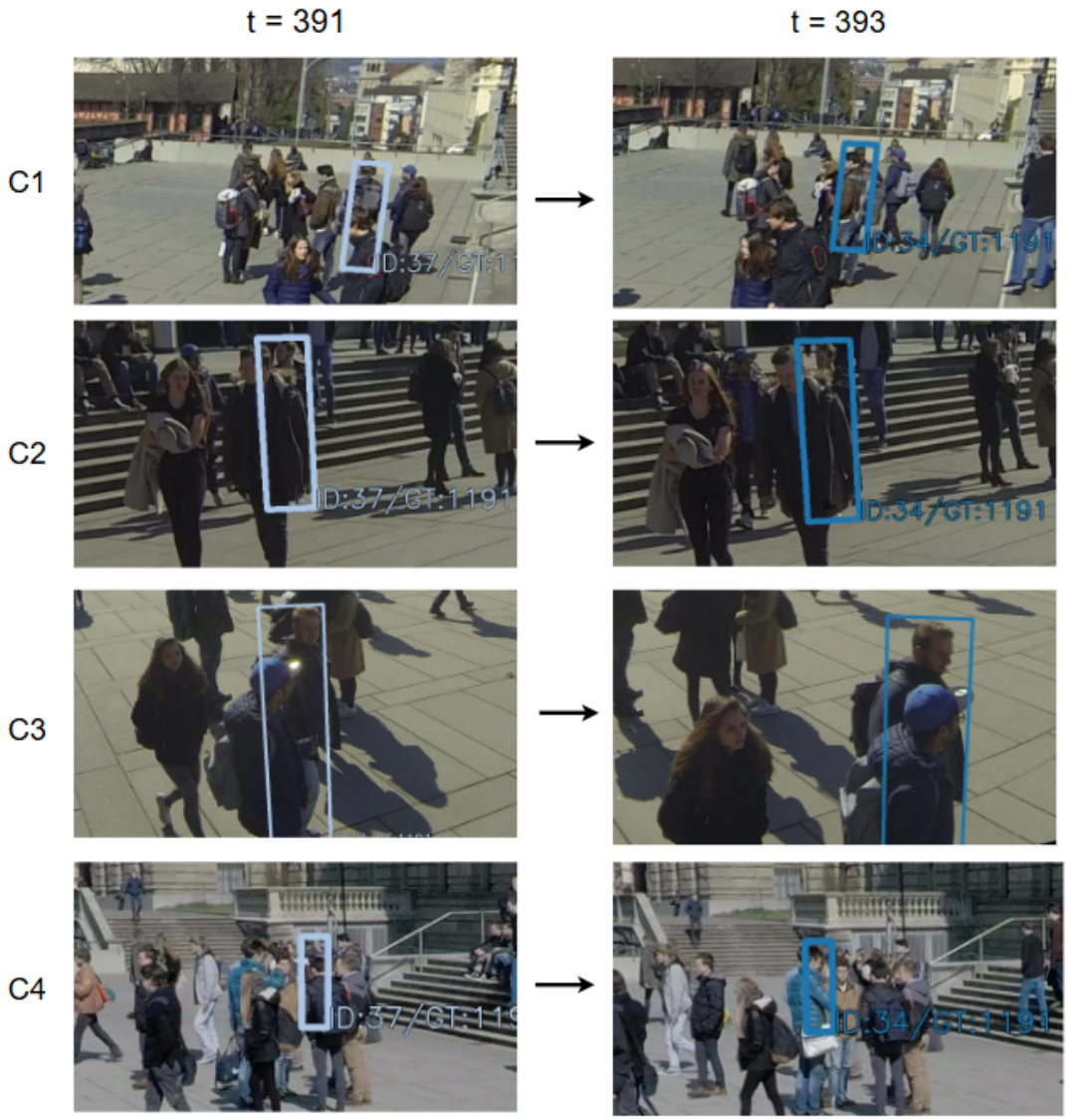


Figure 4.3: ID switch from track fragmentation (WildTrack, frames $t = 391$ and $t = 393$). Two instants are shown for four active views; in the remaining cameras the subject is out of view (the intermediate frame $t = 392$ is omitted due to a miss). Panels display per-camera crops with a fixed window across the two instants. Bounding boxes are the tracker's reprojected per-view ROIs; colours/labels denote the tracker-assigned track IDs. At $t = 393$ the same person is continued by a different track, so the evaluator counts an IDSW.

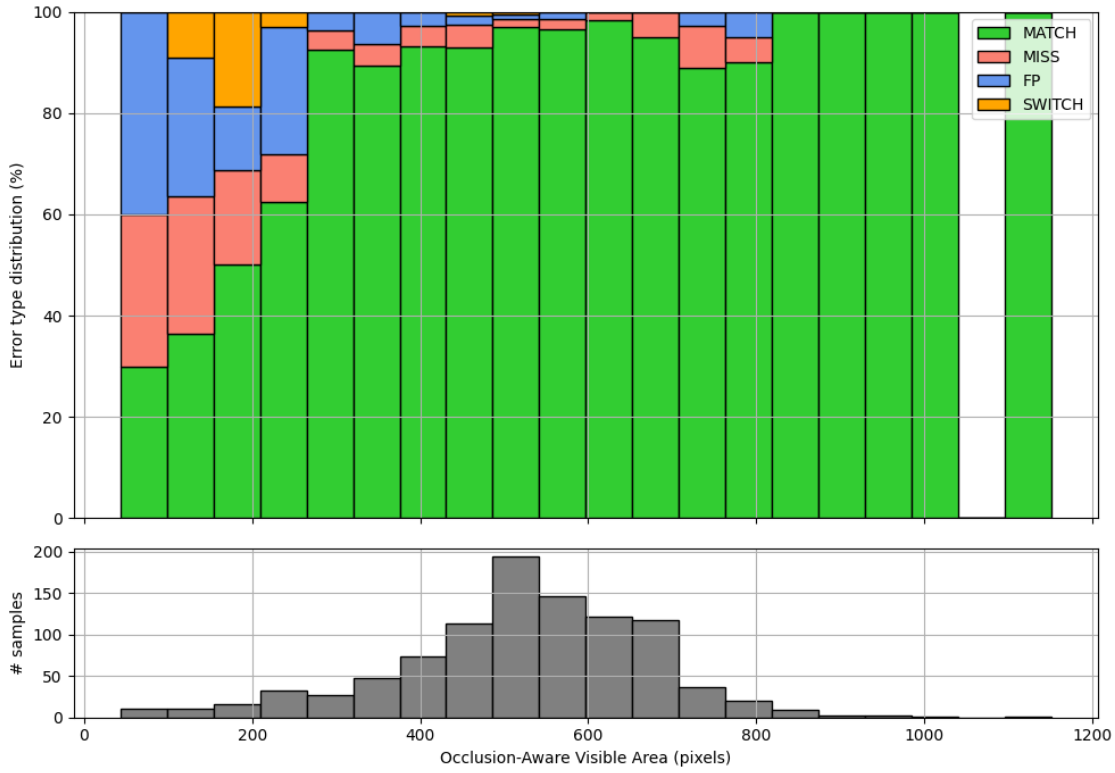
4.6.2 Visibility-Aware Error Analysis

This subsection relates tracking outcomes to the occlusion-aware visibility score (see Section 3.5). The x-axis reports the aggregated visible area (in pixels); the top panel shows, for each bin, the percentages of {MATCH, MISS, FP, IDSW} (summing to 100% within the bin), while the bottom panel reports the number of samples to indicate support. The same binning and legend are used across methods (full and ablations). We focus on IDSW—as the indicator of identity consistency in our pipeline—while MISS/FP are reported for context but are largely detector-driven and thus less informative for association. Bins with low support should be read cautiously and not overinterpreted. We first discuss WildTrack [11](primary dataset, with ablations; Fig. 4.4), then summarize MultiViewX [27] and refer to Appendix B for the corresponding panels. This is a diagnostic (not inferential) analysis, meant to explain when and why IDSW arise, in coherence with the qualitative cases in Section 4.6.1.

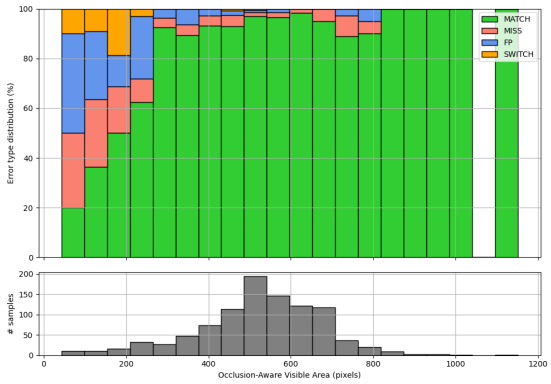
WildTrack On the test split, the Full two-stage exhibits a clear pattern in Fig. 4.4: ID switches appear only in the lowest-visibility bins, while in the modal range—where most samples lie—and beyond they are essentially absent. In that regime the residual errors are mainly MISS/FP, which are largely detector-driven and less informative for association. This reading is consistent with the tracker design: as soon as one–two views yield informative crops, multi-view appearance stabilizes identity; when aggregated visibility is poor, visual cues lose discriminative power and a few switches remain.

The ablations clarify the contribution of each stage. With Appearance-only, the distribution closely matches the full method but there is one extra IDSW in the leftmost bin—precisely the case recovered by the geometric fallback in the two-stage scheme. With Motion-only, switches also occur at medium visibility; in the very lowest bins they are rare because MISS/FP dominate. Without appearance, when multiple candidates fall within the BEV gate the assignment is more prone to crossovers even if visibility is not low (see Section 4.5.1 for aggregated metrics).

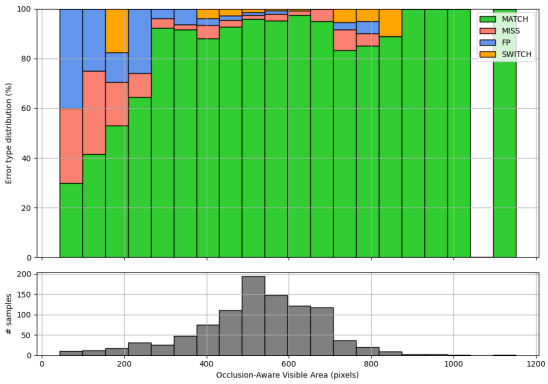
Finally, the visibility-aware diagnostic guided the heuristic choice of the threshold τ_{vis} used in Section 4.5.2: we placed it above the visibility range where the Full two-stage still exhibited IDSW and below the modal regime, so that Stage 1 (appearance) is used only when the aggregated visible area is likely informative, otherwise deferring to the geometric fallback. This choice was made from the test-set distributions on WildTrack—hence it involves test-side peeking and the gains reported there may be mildly optimistic. Applying the same τ_{vis} to MultiViewX did not help, indicating dataset sensitivity. In practice, τ_{vis} should be tuned on the training split of each dataset; absent such tuning, the Full two-stage remains the safer default (see Section 4.5.2).



(a) Full Two-Stage



(b) Appearance-only



(c) Motion-only

Figure 4.4: Visibility-aware error distributions on WildTrack. X-axis: occlusion-aware visible area (pixels). For each bin, the top panel reports the percentage breakdown of {MATCH, MISS, FP, SWITCH} (summing to 100%); the bottom panel shows the number of samples. In (a) ID switches appear only in the lowest-visibility bins; (b) mirrors (a) with one extra switch in the leftmost bin; (c) shows switches also at medium visibility, while the very lowest bins are dominated by MISS/FP.

MultiviewX Using the same protocol, binning, and legend as for WildTrack, the panels for MultiViewX (see Appendix B, Fig. B.1) show a flatter picture. With the Full two-stage method the share of MATCH remains high across almost all bins, while IDSW are rare and do not concentrate at specific visibility levels. The histogram indicates that most samples fall in a mid–visibility range and that the extreme bins have little support, so patterns there should not be over–interpreted. Where they appear, MISS/FP are largely detector–driven and carry limited diagnostic value for association.

The ablations confirm this: Appearance-only and Motion-only produce distributions very similar to the full method; a few switches may occur at intermediate visibility with Motion-only, but counts are small and differences are not robust (see Section 4.5.1). In short, on MultiViewX the two–stage policy is not really stressed by strong occlusions, and the visibility score has limited discriminative power on IDSW compared to WildTrack.

In practice, with a less crowded scene and weaker occlusions, the visibility–aware variant configured with the threshold τ_{vis} chosen on WildTrack does not bring benefits on MultiViewX; τ_{vis} should be tuned per dataset. Absent such tuning, the Full two-stage remains the safer default (see Section 4.5.2). Residual switches on MultiViewX are more plausibly due to factors other than occlusion—e.g., small BEV localization errors near the gate or mild calibration noise—than to insufficient visual evidence.

Chapter 5

Conclusions

This thesis addressed multi-camera multi-object tracking (MC-MOT) for pedestrians in the bird's-eye view (BEV) with calibrated, time-synchronized cameras and overlapping views. MC-MOT serves application domains such as urban mobility, public space safety, sports analytics, and autonomous platforms. These systems require measuring how objects move and recognizing identity stability over time, so that automated counts, flow and occupancy estimates, and behavioral measures can be trusted. However, the multi-view setting poses specific demands: evidence is viewpoint dependent, and overlap creates competing hypotheses that must be resolved across views.

Recent work fuses evidence early in BEV, deciding detections in a shared ground-plane frame to reduce cross-view duplication and stabilize spatial consistency [53, 2, 57]. Within this setting, identity is either kept online in BEV via prediction and geometric gating, yet tied to detector features [53, 2], or made detector-agnostic from image-space crops, with graph-based association using temporal windows and look-ahead rather than frame-by-frame decisions [57]. Despite this progress, a gap remains: a BEV tracker that is both fully online, making per-frame decisions using only past and current frames with no look-ahead, and detector-agnostic under realistic latency budgets is missing.

To address this gap, this thesis contributes two elements. First, it defines a BEV tracker that is fully online and detector-agnostic. We build on the standard online BEV association with prediction and geometric gating [53, 2], replacing appearance features from the detector with detector-agnostic features from back-projected multi-view crops. The separation between motion and appearance keeps the design portable across early-fusion detectors and makes control and diagnosis straightforward. Second, it introduces an occlusion-aware visibility score that quantifies, for each target and time step, the multi-view availability of visual evidence from scene geometry and mutual occlusion. The score is used to analyze where and why identity errors arise in the online regime and supports a visibility-aware association policy that activates the appearance stage only when estimated visibility is

sufficient.

Key Findings The proposed method was evaluated on WildTrack [11] and MultiViewX [27], two calibrated MC-MOT benchmarks. WildTrack captures a real outdoor pedestrian plaza in front of the ETH main building, with people walking, crossing, and forming small groups under overlapping views. MultiViewX is a controlled synthetic multi-view scene for pedestrians. The same ground-plane BEV evaluation protocol as [53, 2, 57, 14, 54] was used for comparison; MVDeTr [26] served as the upstream early-fusion detector.

On WildTrack, the tracker is competitive with these references on identity (IDF1) and overall trajectory quality. On MultiViewX, it ranks first across all reported metrics. In both datasets, the detector's BEV detections are converted into stable identities with minimal loss at the association stage.

Two ablations were studied. Motion-only uses only BEV geometry for matching: candidates are gated and paired by ground-plane distance. Appearance-only uses the same BEV gate to define admissible pairs and then matches by appearance, without the geometric fallback stage. Results show that motion-only increases identity switches, especially in close interactions; by adding an appearance matching stage, the appearance-only variant substantially reduces these switches relative to the motion-only baseline. In the full model, the geometric fallback recovers some residual matches and yields small, consistent gains on both datasets.

The visibility score is used as a diagnostic tool for association. On WildTrack, identity switches tend to occur at low values of the visibility score. On MultiViewX, no clear pattern links the visibility score to switches. Despite the lack of a clear pattern on MultiViewX, the evidence on WildTrack motivated a visibility-aware policy: the appearance stage is applied only when the estimated visibility exceeds a threshold, otherwise the geometric stage is used. The variant behaves as intended on WildTrack, but the gains are modest; on MultiViewX it brings no clear improvement, and outcomes are sensitive to the chosen threshold, so it is left for future development.

Limitations A few limitations set the context for these conclusions. The evaluation is restricted to two outdoor benchmarks of limited duration; indoor scenes, static occluders within the observed area, and long sequences with extended absences and late re-entries are not evaluated. On the method side, appearance is extracted from back-projected image crops and used without additional validation; small calibration mismatches or noise in the BEV hypothesis can introduce background content or partial subjects into the crops, degrading the quality of appearance features. The visibility score is purely geometric and derived from detections; it does not model static occluders or the full scene layout and therefore inherits detector misses and false positives, and in cluttered areas it may overestimate

the available evidence. Finally, although the tracker is designed to be detector-agnostic, it was validated in depth with a single upstream detector; performance sensitivity to the choice of upstream detector and to detection quality has not been measured.

5.1 Future Work

Potential short-term directions include introducing per-view crop validation before extracting appearance as [57] to improve appearance features and increase the reliability of the visibility score. The threshold of the visibility-aware variant can be tuned offline on the training set, in line with the other hyperparameters. The evaluation can be extended to indoor scenes and to longer sequences that stress prolonged occlusions, drift, and late re-entries. Finally, alternative early-fusion upstream detectors can be evaluated to assess sensitivity to detector choice and detection quality.

In the longer term, a potential extension is to move beyond pedestrians and consider road vehicles in fixed, calibrated multi-camera settings. Representative multi-camera vehicle datasets such as CityFlow [52], a real multi-camera traffic dataset, and Synthehicle [24], a synthetic counterpart, can be used to evaluate generalization under different kinematics, appearance regularities, and occlusion patterns.

Appendix A

Problem formulation and geometric background

This appendix collects the formal problem setting, the global notation, and the geometric background (BEV and multi-view calibration) assumed throughout the thesis, with particular reference to the proposed method in Chapter 3. The goal is to provide a single, citable place for symbols, assumptions, and geometric mappings, so that Chapter 3 can focus on the algorithmic design. Readers familiar with calibrated multi-camera tracking may skim this appendix and refer back to it only when notation or geometric details are needed.

A.1 Problem Definition

The goal of multi-target tracking in a multi-camera setting is to estimate, for each physical target τ, a time-indexed trajectory in a common world frame while preserving its identity across frames and across views. We consider fixed, synchronized, and fully calibrated camera networks (intrinsics and extrinsics known), which enable consistent projection and back-projection between image planes and a shared ground-plane reference via planar homographies $\mathbf{H}_c$ [23, 66]. In line with the WildTrack setup, the scene is approximately planar and cameras have overlapping fields of view with precise joint calibration and sequence synchronization [11].

Formally, at each time t the system maintains a set of active targets $\mathcal{T}_t$ (Section A.2). Each target $\tau \in \mathcal{T}_t$ has a kinematic state $\mathbf{x}_t^{(\tau)}$ in the BEV coordinate frame. Each camera $c \in \mathcal{C}$ produces a set of detections $\mathcal{D}_t^{(c)}$ mapped to the ground plane; the union over cameras is $\mathcal{D}_t$. The tracker must (i) associate detections to existing targets or initialize new ones, and (ii) update states over time under measurement noise and intermittent visibility. This yields identity-consistent trajectories $\{\mathbf{x}_t^{(\tau)}\}_t$ in the shared frame.

Compared to single-camera multi-object tracking (MOT), the multi-camera case is challenged by view-dependent appearance, occlusions, and partial coverage, as a target may be visible in only a subset of cameras at any time. Cross-view ambiguity can arise when different targets produce overlapping projections on the ground plane or have similar appearance. These factors make data association under uncertainty the core difficulty and motivate the use of both spatial consistency in the shared frame and appearance cues during matching. The reliance on a ground-plane model via homographies is standard practice in overlapping multi-camera pedestrian tracking and is effective provided calibration is accurate and vertical extent is small relative to the camera height [23, 19].

A.2 Notation

This section collects the symbols and conventions used throughout the thesis; unless stated otherwise, quantities are expressed in the BEV ground-plane frame and time is discrete.

Conventions. Bold lowercase for vectors and bold uppercase for matrices; calligraphic for sets; a tilde denotes homogeneous coordinates; a hat denotes estimates ($\hat{\cdot}$); subscripts $t|t{-}1$ and $t|t$ indicate predicted and filtered quantities. All vectors are column vectors. BEV quantities are in metric units (meters; $\mathrm{m\,s^{-1}}$).

Operators $\top$: transpose; $\|\cdot\|_2$: Euclidean norm; $\mathrm{diag}(\cdot)$: diagonal matrix constructor; $\sim$: equality up to non-zero scale (projective); $\mathrm{norm}(\cdot)$: ℓ_2 re-normalization; $\mathbf{I}$: identity matrix (size by context).

Indices and sets

- t time index; τ track identity; c camera index; k detection index.
- $\mathcal{C}$ set of cameras; $\mathcal{T}_t$ active tracks at time t.
- $\mathcal{D}_t^{(c)}$ detections from camera c at time t; $\mathcal{D}_t = \bigcup_{c\in\mathcal{C}} \mathcal{D}_t^{(c)}$.

State and measurements (BEV)

- $\mathbf{x}_t^{(\tau)} = [x_t, y_t, \dot{x}_t, \dot{y}_t]^\top$: kinematic state of track τ.
- $\mathbf{z}_t^{(c,k)}$ BEV measurement from camera c; shorthand $\mathbf{z}_t$ when unambiguous.
- $\hat{\mathbf{x}}_{t|t-1}$, $\hat{\mathbf{x}}_{t|t}$: predicted and filtered state; $\mathbf{P}_{t|t-1}$, $\mathbf{P}_{t|t}$ covariances.
- $\hat{\mathbf{z}}_{t|t-1}$ predicted measurement; $\mathbf{S}_t$ innovation covariance.
- $\hat{\mathbf{p}}_t \triangleq [\hat{x}_{t|t-1}, \hat{y}_{t|t-1}]^\top$: predicted BEV position.

Kalman filter

- $\mathbf{F}(\Delta t)$ state transition; $\mathbf{Q}(\Delta t)$ process noise (white-acceleration, σ_a).
- $\mathbf{H}$ measurement model; $\mathbf{R} = \mathrm{diag}(\sigma_x^2, \sigma_y^2)$ measurement noise.
- $\mathbf{K}_t$ Kalman gain; Δt inter-frame interval.

Multi-view geometry

- $\mathbf{K}_c$ intrinsics; $\mathbf{R}_c$, $\mathbf{T}_c$ extrinsics; $\mathbf{H}_c$ plane-to-image homography.
- $\mathbf{P}_c = \mathbf{K}_c[\mathbf{R}_c \mid \mathbf{T}_c]$ projection matrix.
- $\mathbf{X}_w$ world point (homog.); $\tilde{\mathbf{X}}_g = [X\, Y\, 1]^\top$ ground-plane (homog.); $\tilde{\mathbf{x}}^{(c)}$ image (homog.); $\Pi(\cdot)$ dehomogenization.
- $\mathbf{u} = (u_x, u_y)$: unit vector orthogonal to the viewing ray on the ground plane.

Appearance descriptors

- $\mathbf{f}_i$ detection descriptor; $\mathbf{a}_j$ track descriptor; both ℓ_2-normalized in $\mathbb{R}^d$.
- $\bar{\mathbf{f}} = \mathrm{norm}\big(\frac{1}{N}\sum_{i=1}^{N} \mathbf{f}_i\big)$ aggregated multi-view descriptor.

Lifecycle and thresholds

- $h_t^{(\tau)}$, $u_t^{(\tau)}$: hit/miss counters.
- $N_{\text{init}}, N_{\text{miss}}, N_{\text{lost}}$: lifecycle thresholds.

Association

- $\mathbf{d}_i$ detection BEV position; $\mathbf{C} \in \mathbb{R}^{M \times N}$ cost matrix, $C_{i,j}$ its entries.
- τ Euclidean gating radius; τ' Mahalanobis gate; $\tau_{\max}$ visual acceptance; τ_{motion} geometric fallback.

Visibility and occlusion

- $B_t^{(c,\tau)}$ ideal image-plane bounding box for subject τ in view c at time t.
- $M_t^{(c,\tau)}$ occlusion mask: pixels of $B_t^{(c,\tau)}$ covered by any closer subject in view c.
- $v_t^{(c,\tau)}$: per-view geometric visibility score (Section 3.5.2).
- $V_t^{(\tau)}$: aggregated visibility score over all cameras for subject τ.
- τ_{vis}: visibility threshold for weighting or gating.

Evaluation protocol

- r_{match}: ground-plane matching radius used by the evaluator (metrics).

Abbreviations. Where unambiguous, we drop superscripts (e.g., $\mathbf{x}_t$ for $\mathbf{x}_t^{(\tau)}$) and use $\mathbf{z}_t$ for a single assigned measurement.

A.3 Shared Reference Space (Bird's-Eye View)

Multi-camera tracking collects observations from different viewpoints, each defined in its own image coordinate frame. To make detections comparable across views, all per-camera observations are mapped to a shared ground-plane reference, the Bird's-Eye View(BEV). In BEV, targets are represented by 2D metric coordinates on an approximately planar surface with a fixed origin and orientation shared across cameras.

Adopting a common BEV frame offers three benefits: (i) view-agnostic spatial reasoning—detections from different cameras can be compared in a single coordinate system; (ii) simplified multi-view fusion and association—spatial proximity tests, gating, and clustering operate in one space; (iii) global motion models defined in metric units, improving temporal consistency.

This abstraction relies on standard assumptions for calibrated multi-camera pedestrian tracking: the scene is approximately planar (pedestrians move on a common ground surface), and each camera is intrinsically and extrinsically calibrated so that a planar homography $\mathbf{H}_c$ exists between the ground plane and the image of camera c [23, 66]. Under these conditions, per-view detections can be projected to BEV with a small, approximately constant bias near the ground plane, aggregated across views, and associated even under partial visibility or limited field-of-view overlap. BEV is therefore the working space for data association and track management; the underlying geometric mappings are detailed in Section A.4.

A.4 Multi-View Geometry and Camera Calibration

We adopt the pinhole camera model with known intrinsics and extrinsics for each camera c. A 3D world point in homogeneous coordinates $\mathbf{X}_w$ projects to homogeneous image coordinates $\tilde{\mathbf{x}}^{(c)}$ as

$$\tilde{\mathbf{x}}^{(c)} \sim \mathbf{K}_c \, [\, \mathbf{R}_c \mid \mathbf{T}_c \,] \, \mathbf{X}_w, \tag{A.1}$$

where "$\sim$" denotes equality up to a non-zero scale factor [23]. Lens distortion is compensated using the calibrated distortion coefficients.

Assuming targets move on a common ground plane (z=0 in the world frame), image points and ground-plane points $\tilde{\mathbf{X}}_g$ are related by a planar homography:

$$\tilde{\mathbf{x}}^{(c)} \sim \mathbf{H}_c \tilde{\mathbf{X}}_g, \tag{A.2}$$

with $\mathbf{H}_c$ computable from $\mathbf{K}_c$, $\mathbf{R}_c$, $\mathbf{T}_c$ when the plane coincides with z=0 [23, 66]. An explicit form (up to scale) is

$$\mathbf{H}_c \sim \mathbf{K}_c \left[\mathbf{r}_{c,1} \; \mathbf{r}_{c,2} \; \mathbf{T}_c \right], \tag{A.3}$$

where $\mathbf{r}_{c,1}$ and $\mathbf{r}_{c,2}$ are the first two columns of $\mathbf{R}_c$.

Practical notes. (i) Calibration yields all required intrinsics, extrinsics, and distortion parameters. (ii) The BEV resolution (meters-per-pixel) is fixed and consistent across the chapter. (iii) Cameras are frame-synchronized to enable temporally consistent multi-view fusion.

Limitations. The planar model is accurate for ground-contact points (e.g., pedestrian footpoints) and introduces deterministic bias for off-plane features, with error magnitude increasing with height and viewing angle. Small calibration errors or deviations from planarity cause systematic BEV misalignments; in our setting these effects are mitigated by spatial gating and robust data association rather than full 3D reconstruction.

A.5 Semantic Entities

The tracking pipeline operates on three levels of semantic entities: *detections*, *tracklets*, and *tracks* [4]. They abstract per-frame evidence, short-term temporal consistency, and long-term identity, respectively.

Detections. At time t, each camera $c \in \mathcal{C}$ (see Section A.2) produces per-view evidence that is mapped to the ground plane (BEV). After optional multi-view aggregation, a detection is a BEV measurement $\mathbf{z}_t^{(c,k)} \in \mathbb{R}^2$ with attributes such as: timestamp t, confidence score, the set of supporting cameras $\subseteq \mathcal{C}$, and (when available) per-view appearance descriptors. Detections are the atomic inputs to the association stage.

Tracklets. A tracklet is a short temporal sequence of detections that is geometrically consistent in BEV and, when available, visually compatible. Tracklets act as a persistence filter: they accumulate evidence across a minimum number of consecutive frames before

a new identity is promoted to a full track. Each tracklet stores a recent history of BEV positions (and optional appearance features) to support subsequent matching.

Tracks. A track represents a persistent target identity maintained over time. Each track carries (i) a kinematic state $\mathbf{x}_t^{(\tau)}$ in BEV (e.g., position and velocity), (ii) a unique identifier τ, and (iii) optional appearance descriptors aggregated over time and views. Tracks are updated frame by frame via prediction and association (defined in Sections 3.2.4 and 3.4.5); they persist through brief occlusions and may be temporarily unmatched without immediate deletion. The life-cycle of a track (tentative, confirmed, lost/deleted) is specified in Section A.6.

A.6 Lifecycle of Tracks

We represent the evolution of a track as a finite-state machine with states Tentative, Confirmed, and Lost, plus the absorbing state Deleted [4, 9, 60]. State transitions are driven by per-frame association outcomes and simple temporal thresholds, independently of the specific matching or motion model used.

Initialization A new Tentative track is spawned when consistent evidence supports the presence of a previously untracked target. In practice, this originates from a short sequence of compatible detections (a tracklet). Let $h_t^{(\tau)}$ and $u_t^{(\tau)}$ be the hit and miss counters of candidate τ at time t. The track remains Tentative until it accumulates at least N_{init} hits over consecutive frames; upon reaching this threshold it becomes Confirmed. Candidates that fail to reach N_{init} due to intervening misses are discarded.

Per-frame update At each time t, a Confirmed or Tentative track receives either a successful association with a detection or a miss. On a hit, $h_t^{(\tau)} \leftarrow h_{t-1}^{(\tau)}+1$ and $u_t^{(\tau)} \leftarrow 0$; on a miss, $u_t^{(\tau)} \leftarrow u_{t-1}^{(\tau)}+1$ and $h_t^{(\tau)}$ is unchanged. The kinematic state is propagated between updates using the motion model defined in Section 3.2.4; association decisions are defined in Section 3.4.

Temporary loss and deletion A Confirmed track transitions to Lost when its miss counter exceeds N_{miss}. While Lost, the track is kept for at most N_{lost} additional frames to allow re-acquisition; if a valid association occurs within this window, the track returns to Confirmed (with $u_t^{(\tau)} \leftarrow 0$). If no observation is obtained for N_{lost} frames, the track becomes Deleted and is permanently removed.

Summary of transitions

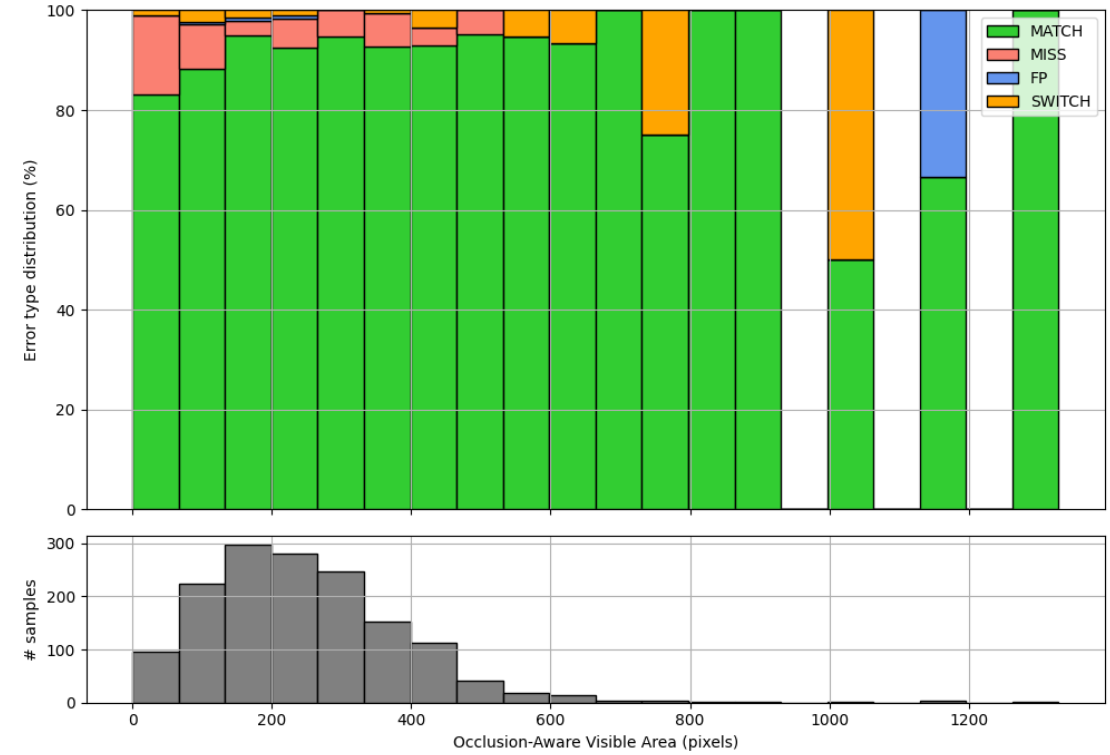
$$\textsc{Tentative} \xrightarrow{h \geq N_{\text{init}}} \textsc{Confirmed}$$
$$\textsc{Confirmed} \xrightarrow{u > N_{\text{miss}}} \textsc{Lost}$$
$$\textsc{Lost} \xrightarrow{\text{hit within } N_{\text{lost}}} \textsc{Confirmed}$$
$$\textsc{Lost} \xrightarrow{\text{no hit for } N_{\text{lost}}} \textsc{Deleted}$$

Rationale This staged lifecycle filters out spurious detections at initialization, preserves identity across brief occlusions or coverage gaps by tolerating short miss streaks, and bounds drift by removing unsupported hypotheses. The thresholds $N_{\text{init}}, N_{\text{miss}}, N_{\text{lost}}$ are set once and reported in Section 4.3.

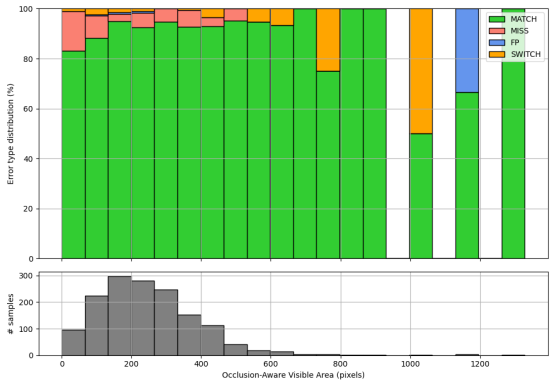
Appendix B

Additional Diagnostic Figures

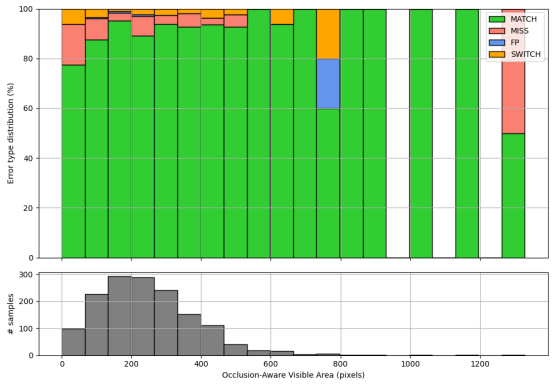
For completeness we report the visibility–aware error distributions on *MultiViewX*. The reading convention is the same used in Section 4.6.2.



(a) Full two-stage



(b) Appearance-only



(c) Motion-only

Figure B.1: Visibility-aware error distributions on MultiViewX. X-axis: occlusion-aware visible area (pixels). For each bin, the top panel reports the percentage breakdown of {MATCH, MISS, FP, SWITCH} (summing to 100%); the bottom panel shows the number of samples. Compared to WildTrack, the profile is flatter: MATCH remains high across bins, SWITCH are rare and not concentrated, and the extreme bins have little support. Both ablations broadly resemble the full method.

Bibliography

[1] N Aharon, R Orfaig, and BZ Bobrovsky. "BoT-SORT: Robust associations multi-pedestrian tracking. arXiv 2022". In: *arXiv preprint arXiv:2206.14651* ().

[2] Reef Alturki, Adrian Hilton, and Jean-Yves Guillemaut. "Attention-Aware Multi-View Pedestrian Tracking". In: *arXiv preprint arXiv:2504.03047* (2025).

[3] Temitope Ibrahim Amosa et al. "Multi-camera multi-object tracking: A review of current trends and future advances". In: *Neurocomputing* 552 (2023), p. 126558.

[4] Yaakov Bar-Shalom, X.-R. Li, and T. Kirubarajan. *Estimation with Applications to Tracking and Navigation*. Wiley, 2001.

[5] Jerome Berclaz, Francois Fleuret, and Pascal Fua. "Multiple object tracking using flow linear programming". In: *2009 Twelfth IEEE international workshop on performance evaluation of tracking and surveillance*. IEEE. 2009, pp. 1–8.

[6] Jerome Berclaz et al. "Multiple object tracking using k-shortest paths optimization". In: *IEEE transactions on pattern analysis and machine intelligence* 33.9 (2011), pp. 1806–1819.

[7] Philipp Bergmann, Tim Meinhardt, and Laura Leal-Taixe. "Tracking without bells and whistles". In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2019, pp. 941–951.

[8] Keni Bernardin and Rainer Stiefelhagen. "Evaluating multiple object tracking performance: the clear mot metrics". In: *EURASIP Journal on Image and Video Processing* 2008.1 (2008), p. 246309.

[9] Alex Bewley et al. "Simple Online and Realtime Tracking". In: *ICIP*. 2016, pp. 3464–3468.

[10] Navaneeth Bodla et al. "Soft-NMS–improving object detection with one line of code". In: *Proceedings of the IEEE international conference on computer vision*. 2017, pp. 5561–5569.

[11] Tatjana Chavdarova et al. “WILDTRACK: A Multi-Camera HD Dataset for Dense Unscripted Pedestrian Detection”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2018, pp. 5030–5039.

[12] Lin Chen et al. “Multi-camera multi-target tracking with space-time-view hypergraph”. In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2020.

[13] Long Chen et al. “Real-time multiple people tracking with deeply learned candidate selection and person re-identification”. In: *2018 IEEE international conference on multimedia and expo (ICME)*. IEEE. 2018, pp. 1–6.

[14] Cheng-Che Cheng et al. “ReST: A Reconfigurable Spatial-Temporal Graph Model for Multi-Camera Multi-Object Tracking”. In: *arXiv preprint arXiv:2308.13229* (2023).

[15] Wongun Choi. “Near-online multi-target tracking with aggregated local flow descriptor”. In: *Proceedings of the IEEE international conference on computer vision*. 2015, pp. 3029–3037.

[16] Shuxiao Ding et al. “Ada-track: End-to-end multi-camera 3d multi-object tracking with alternating detection and association”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2024, pp. 15184–15194.

[17] Martin Engilberge, Weizhe Liu, and Pascal Fua. “Multi-view tracking using weakly supervised human motion prediction”. In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 2023, pp. 1582–1592.

[18] Ran Eshel and Yael Moses. “Homography based multiple camera detection and tracking of people in a dense crowd”. In: *2008 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE. 2008, pp. 1–8.

[19] Francois Fleuret et al. “Multicamera people tracking with a probabilistic occupancy map”. In: *IEEE transactions on pattern analysis and machine intelligence* 30.2 (2008), pp. 267–282.

[20] Dariu M Gavrila and Stefan Munder. “Multi-cue pedestrian detection and tracking from a moving vehicle”. In: *International journal of computer vision* 73.1 (2007), pp. 41–59.

[21] Zheng Ge et al. “Yolox: Exceeding yolo series in 2021”. In: *arXiv preprint arXiv:2107.08430* (2021).

[22] Charles R Harris et al. “Array programming with NumPy”. In: *nature* 585.7825 (2020), pp. 357–362.

[23] Richard Hartley and Andrew Zisserman. *Multiple view geometry in computer vision*. Cambridge university press, 2003.

[24] Fabian Herzog et al. “Synthehicle: Multi-vehicle multi-camera tracking in virtual cities”. In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 2023, pp. 1–11.

[25] Yizhou Hou. *Multiview Detection with Shadow Transformer (and View-Coherent Data Augmentation)*. GitHub; exact version pinned at commit `66cae15`. 2021. URL: https://github.com/hou-yz/MVDeTr (visited on 08/21/2025).

[26] Yunzhong Hou and Liang Zheng. “Multiview detection with shadow transformer (and view-coherent data augmentation)”. In: *Proceedings of the 29th ACM International Conference on Multimedia*. 2021, pp. 1673–1682.

[27] Yunzhong Hou, Liang Zheng, and Stephen Gould. “Multiview Detection with Feature Perspective Transformation”. In: *European Conference on Computer Vision (ECCV)*. 2020. eprint: arXiv:2007.07247.

[28] Yunzhong Hou et al. “IA-MOT: Instance-Aware Multi-Object Tracking with Motion and Appearance”. In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2020.

[29] Yihan Hu et al. “Planning-oriented autonomous driving”. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2023, pp. 17853–17862.

[30] Hsiang-Wei Huang et al. “Enhancing multi-camera people tracking with anchor-guided clustering and spatio-temporal consistency id re-assignment”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023, pp. 5239–5249.

[31] Rudolph E. Kalman. “A New Approach to Linear Filtering and Prediction Problems”. In: *Journal of Basic Engineering* 82.1 (1960), pp. 35–45.

[32] Sohaib Khan and Mubarak Shah. “Consistent labeling of tracked objects in multiple cameras with overlapping fields of view”. In: *IEEE transactions on pattern analysis and machine intelligence* 25.10 (2003), pp. 1355–1360.

[33] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. “ImageNet classification with deep convolutional neural networks”. In: *Communications of the ACM* 60.6 (2017), pp. 84–90.

[34] Harold W Kuhn. “The Hungarian method for the assignment problem”. In: *Naval research logistics quarterly* 2.1-2 (1955), pp. 83–97.

[35] Martijn Liem, Arnoud Visser, and Frans Groen. “A hybrid algorithm for tracking and following people using a robotic dog”. In: *Proceedings of the 3rd ACM/IEEE international conference on Human robot interaction*. 2008, pp. 185–192.

[36] Martijn C Liem and Dariu M Gavrila. "Joint multi-person detection and tracking from overlapping cameras". In: *Computer Vision and Image Understanding* 128 (2014), pp. 36–50.

[37] Tsung-Yi Lin et al. "Focal loss for dense object detection". In: *Proceedings of the IEEE international conference on computer vision*. 2017, pp. 2980–2988.

[38] Wei Liu et al. "Ssd: Single shot multibox detector". In: *European conference on computer vision*. Springer. 2016, pp. 21–37.

[39] Julius Maliwat. *msc-thesis: code and configurations for the experiments*. Commit 4224659. 2025. URL: https://github.com/JuliusMaliwat/msc-thesis (visited on 08/23/2025).

[40] Wes McKinney et al. "pandas: a foundational Python library for data analysis and statistics". In: *Python for high performance and scientific computing* 14.9 (2011), pp. 1–9.

[41] Anton Milan et al. "MOT16: A Benchmark for Multi-Object Tracking". In: *arXiv preprint arXiv:1603.00831*. 2016. URL: https://arxiv.org/abs/1603.00831.

[42] Duy MH Nguyen et al. "Lmgp: Lifted multicut meets geometry projections for multi-camera multi-object tracking". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2022, pp. 8866–8875.

[43] Adam Paszke et al. "Pytorch: An imperative style, high-performance deep learning library". In: *Advances in neural information processing systems* 32 (2019).

[44] Jathushan Rajasegaran et al. "Tracking people with 3D representations". In: *arXiv preprint arXiv:2111.07868* (2021).

[45] Joseph Redmon and Ali Farhadi. "Yolov3: An incremental improvement". In: *arXiv preprint arXiv:1804.02767* (2018).

[46] Shaoqing Ren et al. "Faster r-cnn: Towards real-time object detection with region proposal networks". In: *Advances in neural information processing systems* 28 (2015).

[47] Ergys Ristani et al. "Performance measures and a data set for multi-target, multi-camera tracking". In: *European conference on computer vision*. Springer. 2016, pp. 17–35.

[48] Gemma Roig et al. "Conditional random fields for multi-camera object detection". In: *2011 International Conference on Computer Vision*. IEEE. 2011, pp. 563–570.

[49] Thiago T Santos and Carlos H Morimoto. "Multiple camera people detection and tracking using support integration". In: *Pattern Recognition Letters* 32.1 (2011), pp. 47–55.

[50] Shuai Shao et al. “Crowdhuman: A benchmark for detecting human in a crowd”. In: *arXiv preprint arXiv:1805.00123* (2018).

[51] Siyu Tang et al. “Multiple people tracking by lifted multicut and person re-identification”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 3539–3548.

[52] Zheng Tang et al. “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*. 2019, pp. 8797–8806.

[53] Torben Teepe et al. “EarlyBird: early-fusion for multi-view tracking in the bird’s eye View”. In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 2024, pp. 102–111.

[54] Torben Teepe et al. “Lifting multi-view detection and tracking to the bird’s eye view”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2024, pp. 667–676.

[55] Zhi Tian et al. “Fcos: Fully convolutional one-stage object detection”. In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2019, pp. 9627–9636.

[56] Pauli Virtanen et al. “SciPy 1.0: fundamental algorithms for scientific computing in Python”. In: *Nature methods* 17.3 (2020), pp. 261–272.

[57] Yizhou Wang et al. “BEV-SUSHI: Multi-Target Multi-Camera 3D Detection and Tracking in Bird’s-Eye View”. In: *arXiv e-prints* (2024), arXiv–2412.

[58] Longhui Wei et al. “Person transfer gan to bridge domain gap for person re-identification”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, pp. 79–88.

[59] Longyin Wen et al. “Multi-camera multi-target tracking with space-time-view hypergraph”. In: *International Journal of Computer Vision* 122.2 (2017), pp. 313–333.

[60] Nicolai Wojke, Alex Bewley, and Dietrich Paulus. “Simple Online and Realtime Tracking with a Deep Association Metric”. In: *WACV*. 2018, pp. 1974–1980.

[61] Ryuto Yoshida et al. “Overlap suppression clustering for offline multi-camera people tracking”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2024, pp. 7153–7162.

[62] Yixiao Yun, Irene Yu-Hua Gu, and Hamid Aghajan. “Maximum-likelihood object tracking from multi-view video by combining homography and epipolar constraints”. In: *2012 Sixth International Conference on Distributed Smart Cameras (ICDSC)*. IEEE. 2012, pp. 1–6.

[63] Lei Zhang et al. “Branch-cooperative osnet for person re-identification”. In: *arXiv preprint arXiv:2006.07206* (2020).

[64] Shanshan Zhang, Rodrigo Benenson, and Bernt Schiele. “Citypersons: A diverse dataset for pedestrian detection”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 3213–3221.

[65] Yifu Zhang et al. “ByteTrack: Multi-Object Tracking by Associating Every Detection Box”. In: *European Conference on Computer Vision (ECCV)*. 2022.

[66] Zhengyou Zhang. “A flexible new technique for camera calibration”. In: *IEEE Transactions on pattern analysis and machine intelligence* 22.11 (2002), pp. 1330–1334.

[67] Liang Zheng et al. “Scalable person re-identification: A benchmark”. In: *Proceedings of the IEEE international conference on computer vision*. 2015, pp. 1116–1124.

[68] Kaiyang Zhou and Tao Xiang. “Torchreid: A Library for Deep Learning Person Re-Identification in Pytorch”. In: *arXiv preprint arXiv:1910.10093* (2019).

[69] Kaiyang Zhou et al. “Omni-scale feature learning for person re-identification”. In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2019, pp. 3702–3712.

[70] Xingyi Zhou, Vladlen Koltun, and Philipp Krähenbühl. “Tracking Objects as Points”. In: *ECCV*. 2020.