Review

Object Tracking: A Comprehensive Survey From Classical Approaches to Large Vision-Language and Foundation Models

Rahul Raja *, Arpita Vats, Omkar Thawakar, Tajamul Ashraf

Posted Date: 6 October 2025

doi: 10.20944/preprints202509.2051.v2

Keywords: object tracking; Single Object Tracking (SOT); Multi-Object Tracking (MOT); Long-Term Tracking (LTT); vision-language models; multimodal fusion; foundation models

Article

Object Tracking: A Comprehensive Survey From Classical Approaches to Large Vision-Language and Foundation Models

Rahul Raja [1,*], Arpita Vats [2], Omkar Thawakar [3] and Tajamul Ashraf [4]

[1] LinkedIn, USA and Carnegie Mellon University, USA
[2] LinkedIn, USA and Boston University, USA
[3] Mohamed bin Zayed University of Artificial Intelligence, UAE
[4] Mohamed bin Zayed University of Artificial Intelligence, UAE
[*] Correspondence: rahul.110392@gmail.com

Abstract

Object tracking remains a central problem in computer vision with broad applications in surveillance, autonomous driving, augmented reality, and human–computer interaction. This paper presents a comprehensive survey that unifies the progression of tracking methodologies, from handcrafted and probabilistic models to deep learning paradigms and recent advances with large vision–language and foundation models. We categorize tracking into Single Object Tracking (SOT), Multi-Object Tracking (MOT), and Long-Term Tracking (LTT), systematically reviewing CNN, Siamese, transformer, and hybrid-based approaches alongside detection-guided, detection-integrated, and re-identification–aware pipelines. Special emphasis is placed on emerging trends, including open-vocabulary tracking, promptable models, and multimodal fusion across RGB, depth, thermal, LiDAR, and event-based inputs. We highlight benchmark datasets, evaluation protocols, and taxonomy refinements that reveal convergence toward unified and generalizable tracking systems. Finally, we discuss open challenges—such as occlusion, scalability, identity consistency, and cross-dataset transferability—and outline future directions in self-supervised learning, adapter tuning, and efficient foundation model adaptation. This survey aims to serve as a reference for both academic researchers and practitioners, bridging classical paradigms with the rapidly evolving landscape of foundation- and vision-language–driven tracking.

Keywords: object tracking; single object tracking (SOT); multi-object tracking (MOT); long-term tracking (LTT); vision–language models; multimodal fusion; foundation models

1. Introduction

Object tracking is a core problem in computer vision that aims to localize one or more objects over time in a sequence of video frames. It plays a vital role in a wide array of real-world applications, including video surveillance, autonomous driving, robotics, augmented reality, human-computer interaction, and sports analytics. Tracking involves not only detecting objects in each frame but also maintaining consistent identities despite challenges such as occlusion, motion blur, illumination variation, and object deformation.

To structure this evolving field, this survey categorizes tracking into three major settings: *Single Object Tracking (SOT)*, where a single target is tracked through a video; *Multi-Object Tracking (MOT)*, which involves identifying and maintaining multiple object identities; and *Long-Term Tracking (LTT)*, where the tracker must re-detect the object after occlusion or disappearance. Each setting poses unique challenges and inspires different methodological choices.

Over the years, object tracking methods have progressed from traditional algorithms based on handcrafted features and probabilistic models to modern deep learning-based frameworks. Recent advancements include Siamese-based trackers, transformer architectures, and end-to-end joint detection and tracking models. Additionally, the rise of vision-language models and open-vocabulary trackers marks a new direction toward more flexible and generalizable tracking systems.

To provide a comprehensive overview, this survey is organized into several key sections, each focused on a different axis of the object tracking landscape: foundational problem settings and taxonomy, traditional methods, deep learning approaches, multi-object tracking architectures, long-term tracking, benchmarks and metrics, and finally, emerging trends such as foundation models and multimodal tracking.

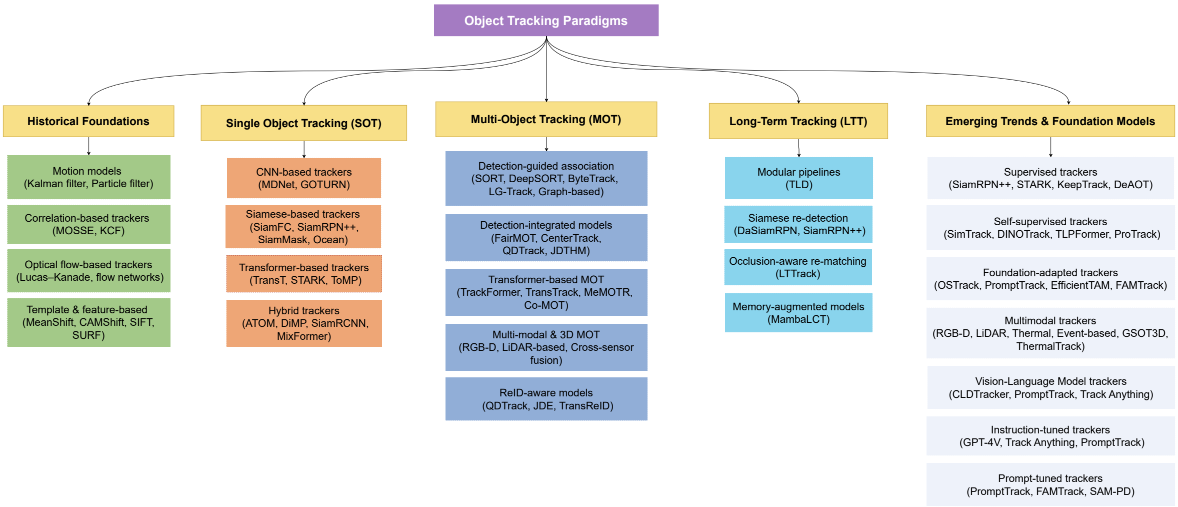
Object Tracking Paradigms
Historical Foundations
Motion models (Kalman filter, Particle filter)
Correlation-based trackers (MOSSE, KCF)
Optical flow-based trackers (Lucas–Kanade, flow networks)
Template & feature-based (MeanShift, CAMShift, SIFT, SURF)
Single Object Tracking (SOT)
CNN-based trackers (MDNet, GOTURN)
Siamese-based trackers (SiamFC, SiamRPN++, SiamMask, Ocean)
Transformer-based trackers (TransT, STARK, ToMP)
Hybrid trackers (ATOM, DiMP, SiamRCNN, MixFormer)
Multi-Object Tracking (MOT)
Detection-guided association (SORT, DeepSORT, ByteTrack, LG-Track, Graph-based)
Detection-integrated models (FairMOT, CenterTrack, QDTrack, JDTHM)
Transformer-based MOT (TrackFormer, TransTrack, MeMOTR, Co-MOT)
Multi-modal & 3D MOT (RGB-D, LiDAR-based, Cross-sensor fusion)
ReID-aware models (QDTrack, JDE, TransReID)
Long-Term Tracking (LTT)
Modular pipelines (TLD)
Siamese re-detection (DaSiamRPN, SiamRPN++)
Occlusion-aware re-matching (LTTrack)
Memory-augmented models (MambaLCT)
Emerging Trends & Foundation Models
Supervised trackers (SiamRPN++, STARK, KeepTrack, DeAOT)
Self-supervised trackers (SimTrack, DINOTrack, TLPFormer, ProTrack)
Foundation-adapted trackers (OSTrack, PromptTrack, EfficientTAM, FAMTrack)
Multimodal trackers (RGB-D, LiDAR, Thermal, Event-based, GSOT3D, ThermalTrack)
Vision-Language Model trackers (CLDTracker, PromptTrack, Track Anything)
Instruction-tuned trackers (GPT-4V, Track Anything, PromptTrack)
Prompt-tuned trackers (PromptTrack, FAMTrack, SAM-PD)

Figure 1. Taxonomy of object tracking paradigms, spanning historical foundations, single-object tracking (SOT), multi-object tracking (MOT), long-term tracking (LTT), and emerging trends leveraging foundation and vision-language models. Each branch highlights representative methods and architectures across the evolution of tracking research

This survey aims to serve as a unified reference for researchers and practitioners in the object tracking community by:

- Presenting a structured taxonomy of object tracking paradigms.
- Reviewing traditional tracking methods such as correlation filters, optical flow, and probabilistic filters.
- Detailing deep learning-based methods, including CNN-based, Siamese-based, and transformer-based trackers.
- Highlighting key advances in multi-object tracking, with a focus on data association, identity preservation, and joint detection-tracking architectures.
- Discussing recent developments in open-vocabulary and multimodal tracking using foundation models.
- Summarizing widely-used datasets and

2. Problem Formulation and Taxonomy

Object tracking is a fundamental task in computer vision that involves estimating the state (e.g., position, scale, and shape) of a target object as it moves through a video sequence. The primary objective is to maintain a continuous and accurate trajectory of the object over time, despite challenges such as occlusions, abrupt motion, background clutter, and varying illumination or viewpoint conditions. Tracking algorithms must be robust to appearance variations, efficient enough for real-time performance, and capable of distinguishing the target from distractors in complex environments.

Let $\{I_t\}_{t=1}^{T}$ denote a sequence of T video frames. The goal is to predict a set of spatial coordinates or bounding boxes B_t in each frame I_t that correspond to the same physical object across time. More formally, the tracker estimates the evolving object states S_t using the current and past observations, possibly incorporating prior knowledge such as motion patterns, category information, or contextual cues.

Contemporary tracking systems typically involve several key components:

- **Feature Extraction:** Transforming raw pixels into semantically meaningful embeddings using deep neural networks, such as convolutional or transformer-based backbones.
- **Motion Modeling:** Capturing temporal dynamics using techniques like optical flow, Kalman filters, or learned motion predictors to estimate object displacement across frames.
- **Data Association:** Linking estimated object states across time by matching detections or predictions using appearance similarity, spatial proximity, or temporal consistency.
- **Model Update:** Adapting the tracking model online to accommodate changes in object appearance, scale, pose, and environmental context.

While the core objective remains consistent, tracking formulations differ based on the number of objects tracked, the temporal length of tracking, and the constraints on initialization or supervision. These scenarios are explored in detail in the

3. Problem Formulation and Taxonomy

Object tracking is a fundamental task in computer vision that involves estimating the state (e.g., position, scale, and shape) of a target object as it moves through a video sequence. The primary objective is to maintain a continuous and accurate trajectory of the object over time, despite challenges such as occlusions, abrupt motion, background clutter, and varying illumination or viewpoint conditions. Tracking algorithms must be robust to appearance variations, efficient enough for real-time performance, and capable of distinguishing the target from distractors in complex environments.

Let $\{I_t\}_{t=1}^{T}$ denote a sequence of T video frames. The goal is to predict a set of spatial coordinates or bounding boxes B_t in each frame I_t that correspond to the same physical object across time. More formally, the tracker estimates the evolving object states S_t using the current and past observations, possibly incorporating prior knowledge such as motion patterns, category information, or contextual cues.

Contemporary tracking systems typically involve several key components:

- **Feature Extraction:** Transforming raw pixels into semantically meaningful embeddings using deep neural networks, such as convolutional or transformer-based backbones.
- **Motion Modeling:** Capturing temporal dynamics using techniques like optical flow, Kalman filters, or learned motion predictors to estimate object displacement across frames.
- **Data Association:** Linking estimated object states across time by matching detections or predictions using appearance similarity, spatial proximity, or temporal consistency.
- **Model Update:** Adapting the tracking model online to accommodate changes in object appearance, scale, pose, and environmental context.

While the core objective remains consistent, tracking formulations differ based on the number of objects tracked, the temporal length of tracking, and the constraints on initialization or supervision. These scenarios are explored in detail in the

4. Single Object Tracking (SOT)

Single Object Tracking (SOT) involves estimating the trajectory of a target object across a video sequence, given its initial bounding box in the first frame [1]. Unlike multi-object tracking, SOT concentrates on localizing a single instance without requiring class labels or re-identification. The task is challenging because the target may undergo occlusion, scale variation, rotation, deformation, or illumination change. While traditional methods relied on handcrafted features, correlation filters, and template matching [2,3], modern deep learning-based trackers represent the task as a learnable similarity function between the initial target template $z = I_1(B_1)$ and candidate regions $x_t(r)$ in the search frame I_t:

$$B_t = \arg\max_{r \in R_t} f_\theta(z, x_t(r)),$$

where R_t denotes the set of candidate regions and f_θ is a learned similarity function parameterized by neural networks.

4.1. CNN-Based Trackers

The first wave of deep trackers used convolutional networks to replace handcrafted features with learned representations. MDNet cast tracking as a binary classification task, using shared convolutional layers trained across multiple domains and domain-specific fully connected layers updated online [4]. This enabled robust discrimination under occlusion but suffered from slow inference due to repeated online fine-tuning. GOTURN instead approached tracking as direct bounding box regression, predicting target coordinates in a feed-forward manner without online updates [5]. Although highly efficient, it lacked adaptation to appearance change. Hybrid CNN–detector trackers, inspired by the TLD framework, incorporated global re-detection modules to recover after failures, though they remained sensitive to drift when noisy updates occurred [6].

4.2. Siamese-Based Trackers

Siamese networks marked a major shift by learning template–search similarity through shared feature embeddings. SiamFC introduced the fully convolutional Siamese design, where cross-correlation between template and search features produced a response map [7]. While simple and real-time, it lacked scale adaptability. SiamRPN integrated a region proposal network to predict both classification scores and bounding box offsets, enabling more precise localization [8]. Its successor, SiamRPN++, extended this with deeper backbones and spatial-aware sampling, which improved robustness at the cost of increased computation [9]. Later extensions such as SiamMask and Ocean expanded the paradigm to include segmentation and attention-based matching for more fine-grained target modeling [10,11].

4.3. Transformer-Based Trackers

Transformers brought global attention and long-range reasoning to SOT. TransT removed reliance on anchors and region proposals by employing cross-attention between template and search regions, learning discriminative correlations directly [12]. STARK extended this approach by treating tracking as sequence prediction, using spatio-temporal transformers to model historical and current features jointly [13]. Other transformer-based designs such as ToMP adapted generic vision transformer backbones to tracking, tokenizing frames into patches and applying dense attention for fine-grained matching [14]. These methods demonstrated the strength of attention-based models in handling clutter, occlusion, and appearance variability, though often at higher computational cost.

4.4. Hybrid Tracking Architectures

Hybrid trackers combine the strengths of convolutional, Siamese, and transformer paradigms. ATOM separated feature extraction from bounding box estimation, using a ResNet backbone and a target-specific regression head optimized online [15]. DiMP improved upon this by introducing a meta-learned optimization module for faster and more robust online adaptation [16]. SiamRCNN bridged Siamese matching and region-based detection, handling scale and aspect ratio variation more flexibly [17]. More recent designs such as SiamBAN and MixFormer unified classification and regression into joint prediction heads, with MixFormer coupling transformers and convolutional backbones to achieve strong performance with fewer parameters [18,19]. These approaches illustrate the ongoing convergence of architectural ideas to achieve a balance between accuracy, adaptability, and efficiency.

Single Object Tracking has progressed from CNN classifiers and regressors to efficient Siamese similarity learning, transformer-driven architectures, and hybrid systems. The central theme across these methods is balancing real-time performance with robustness to appearance change and occlusion, with unified frameworks offering a promising path forward. Table 1 summarizes prominent Single Object Tracking (SOT) models, highlighting their architectural categories, backbone designs, key strengths, limitations, and benchmark performance.

Table 1. Single Object Tracking (SOT) models categorization.

Method	**Category**	**Backbone**	**Key Strength**	**Key Weakness**	**Performance (Dataset/Metric)**
MDNet [4]	CNN	Conv + FC (multi-domain)	Learns domain-invariant representations through multi-domain training; robust under challenging conditions such as heavy occlusion and background clutter; demonstrates strong generalization to unseen objects.	Computationally expensive due to frequent online updates; suffers from low inference speed, limiting real-time usability; performance degrades in long sequences with rapid appearance variation.	AUC: 0.678 (OTB100)
GOTURN [5]	CNN	Dual-input CNN regressor	Extremely fast (>100 FPS), enabling real-time deployment; simple fully offline pipeline; no online fine-tuning needed; efficient feed-forward regression.	Lacks adaptation to target appearance changes; brittle under occlusion, deformation, and scale variation; struggles with long-term robustness in cluttered environments.	AUC: 0.46 (OTB100)
TLD-CNN [6]	CNN	CNN + online learner	Combines detection and tracking, allowing recovery from failures; online learning enables adaptation to dynamic targets; can re-detect objects after drift or loss.	Online updates prone to noise accumulation, leading to drift; high complexity compared to simpler Siamese models; unstable in highly cluttered or fast-moving scenarios.	Precision: ∼0.56 (OTB100)
SiamFC [7]	Siamese	Shared CNN encoder	Lightweight and efficient design; achieves real-time operation (86 FPS); end-to-end similarity learning via cross-correlation; robust against moderate distractors.	Relies on fixed template without update, limiting adaptability; weak handling of scale and aspect ratio changes; fails under long occlusion or drastic appearance variation.	AUC: 0.582 (OTB100)
SiamRPN [8]	Siamese	Siamese CNN + RPN	Incorporates region proposal network (RPN) for accurate localization; handles scale variation better than SiamFC; improved robustness for short- to mid-term tracking.	Strong dependency on anchor design introduces rigidity; limited adaptability to unseen object classes; inference cost increases compared to SiamFC.	EAO: 0.41 (VOT2018)
SiamRPN++ [9]	Siamese	ResNet-50 Siamese	Leverages deep residual features for strong representation; improved receptive fields and robustness; achieves state-of-the-art accuracy on multiple benchmarks.	Computationally heavier than earlier Siamese trackers; constrained by anchor-based formulation; reduced efficiency in embedded or resource-limited systems.	AUC: 0.696 (OTB100)
TransT [12]	Transformer	Cross/self-attention modules	Exploits global context through self- and cross-attention; anchor-free design avoids hand-crafted priors; generalizes well to diverse object categories.	High computational overhead during inference; requires large-scale pretraining for stability; slower than Siamese models on resource-limited devices.	AUC: 0.649 (LaSOT)
STARK [13]	Transformer	Encoder–decoder Transformer	Models temporal dependencies explicitly with spatio-temporal attention; stable under jitter, occlusion, and appearance changes; strong bounding box regression accuracy.	Memory- and compute-intensive; performance sensitive to sequence design; not suitable for lightweight or mobile scenarios.	AUC: 0.678 (LaSOT)
ToMP [14]	Transformer	Transformer + predictor	Performs dense feature matching with strong robustness to appearance changes; eliminates need for frequent online updates; flexible prediction capability.	Dense patch-token representation increases latency; higher complexity hinders real-time performance; resource-demanding for large-scale deployment.	AUC: 0.70 (TrackingNet)
TLD [20]	Hybrid	Optical flow + detector	First to propose tracking–learning–detection loop; global re-detection enables recovery from failures; adaptive appearance models for long-term use.	Online updates prone to drift; handcrafted features limit robustness; unstable in rapidly changing or dynamic scenes.	AUC: 0.53 (OTB100)
ATOM [15]	Hybrid	ResNet + regressor head	Accurate bounding box estimation via dedicated regression head; robust under challenging appearance changes; strong baseline for hybrid trackers.	Requires per-sequence optimization online; prevents real-time deployment; increased latency compared to lightweight models.	AUC: 0.669 (LaSOT)
DiMP [16]	Hybrid	ResNet + meta-learner	Discriminative classification combined with meta-learned updates; adapts quickly to new targets; competitive accuracy on long sequences.	Still requires online optimization; additional training complexity; slower compared to pure Siamese architectures.	AUC: 0.678 (LaSOT)
SiamRCNN [17]	Hybrid	Siamese backbone + R-CNN	Achieves high accuracy under challenging conditions; integrates detection for robust target estimation; flexible handling of aspect ratio and scale.	Computationally heavy two-stage design; significantly slower inference; complex pipeline compared to single-stage trackers.	AUC: 0.64 (LaSOT)
SiamBAN [18]	Hybrid	Siamese + unified head	Anchor-free classification and regression head improves flexibility; balanced trade-off between accuracy and efficiency; robust across diverse conditions.	Still limited by fixed template; lacks strong recovery under long occlusion; less competitive for long-term tracking.	AUC: 0.63 (LaSOT)
MixFormer [19]	Hybrid	Transformer encoder + joint head	Unified CNN+Transformer architecture with fewer parameters; competitive performance across datasets; strong robustness to appearance variation.	Requires extensive pretraining to achieve best performance; heavier than lightweight Siamese designs; reduced efficiency in embedded scenarios.	AUC: 0.704 (LaSOT)

5. Multi-Object Tracking (MOT)

Multi-Object Tracking (MOT) focuses on the task of simultaneously localizing and maintaining consistent identities for multiple objects across video frames [21]. Formally, given a set of object detections at each time step $\mathcal{D}_t = \{d_t^1, d_t^2, \ldots, d_t^{n_t}\}$, the goal is to estimate a set of trajectories $\mathcal{T} = \{\tau_1, \tau_2, \ldots, \tau_M\}$ such that each trajectory $\tau_i = \{(d_t^i, t)\}_{t=t_s}^{t_e}$ corresponds to the same real-world object.

A key challenge in MOT is the **data association problem** - correctly linking detections across frames despite occlusions, abrupt motion, camera shifts, and the presence of visually similar objects. Modern MOT systems often adopt a tracking-by-detection paradigm [22], which separates object detection and temporal association into modular stages. Although this pipeline simplifies learning and improves flexibility, it may suffer from detector errors and delayed identity recovery after occlusion.

Over the past decade, MOT research has shifted from classical filtering and matching algorithms to deep learning-based approaches that integrate appearance models, motion prediction, and re-identification features. Public benchmarks such as MOT17 [23] and DanceTrack [24] have driven progress, enabling fair comparison across identity preservation, detection recall, and trajectory fragmentation.

Despite progress, MOT remains a fundamentally ambiguous problem in crowded scenes and long-term tracking due to ID switches, partial occlusions, and class-agnostic interactions. This has motivated ongoing research into joint detection-tracking models, temporal attention mechanisms, and multi-modal inputs such as LiDAR and radar in autonomous driving.

5.1. Detection-Guided Association Models

Detection-guided tracking frameworks follow a decoupled pipeline wherein objects are first detected in each frame and then associated across time to form trajectories. This paradigm remains dominant due to its modularity and ability to leverage advances in object detection. Formally, let $\mathcal{D}_t = \{d_t^1, d_t^2, \ldots, d_t^{N_t}\}$ be the set of detections at frame t, and $\mathcal{T}_{t-1}$ be the set of active tracks. The goal is to associate $\mathcal{D}_t$ to $\mathcal{T}_{t-1}$ via an assignment matrix A minimizing total cost:

$$A^* = \arg\min_A \sum_{(i,j)\in A} \mathrm{Cost}(d_t^i, \tau_{t-1}^j),$$

where the cost integrates appearance, motion (e.g., Kalman prediction), and geometric similarity. The design choice of how to compute this cost has led to several families of approaches.

5.1.1. SORT Extensions: From Motion-Only to ReID-Aware

Early MOT systems such as SORT relied purely on motion cues and Kalman filtering, yielding efficiency but poor identity consistency. DeepSORT [25] extended this baseline by adding deep appearance embeddings trained for person re-identification (ReID), significantly reducing ID switches. StrongSORT [26] further incorporated Kalman updates and outlier suppression, showing how stabilizing identity propagation improves robustness in noisy scenes. Introducing ReID embeddings transformed MOT from motion-driven matching into a vision-guided task, improving occlusion handling at the cost of higher compute.

5.1.2. Detector-Driven Propagation

Instead of explicit association, some methods reuse the detector itself to propagate trajectories. Tracktor++ [27] leverages the regression head of the detector to move bounding boxes across frames, using classification scores to terminate occluded tracks. Detector-driven propagation simplifies the pipeline but is limited by the detector's recall and struggles under long occlusion or multi-class tracking.

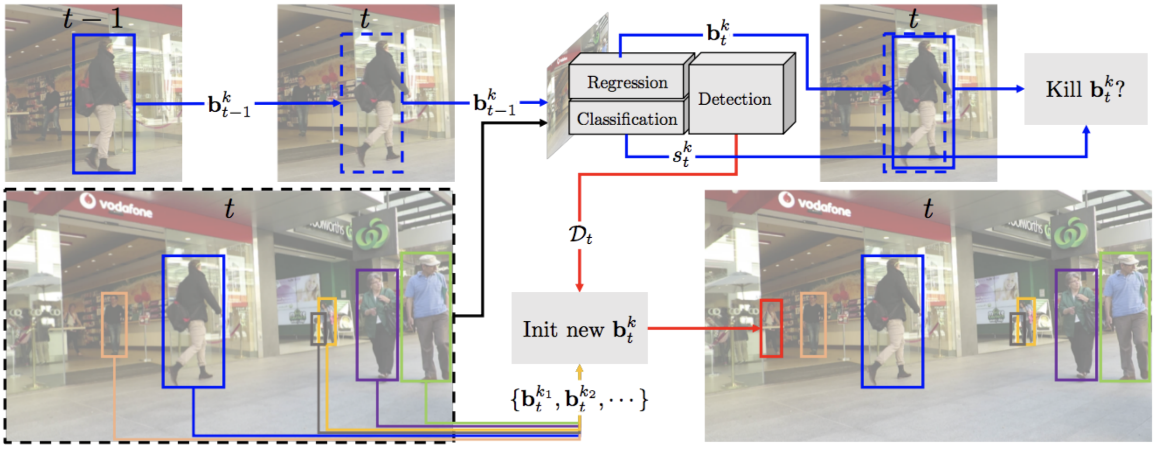

Figure 2. Overview of the Tracktor pipeline for multi-object tracking [27]. Regression aligns bounding boxes from frame $t-1$ to t, while classification scores determine termination. New detections are introduced when overlaps are low.

5.1.3. Recall-Boosting Trackers

Another line of work focuses on reducing false negatives by leveraging low-confidence detections. ByteTrack [28] retains these candidates and applies a two-stage association, improving robustness in crowded scenes. MR2-ByteTrack [29] adapts this principle for embedded platforms using resolution-aware matching to preserve accuracy under resource constraints.Recall-boosting methods highlight the precision–recall trade-off, demonstrating that retaining noisy detections can improve identity stability if handled with careful association.

5.1.4. Confidence-Aware Association

LG-Track [30] distinguishes between classification and localization confidence, improving association by retaining well-localized but low-score detections. Deep LG-Track [31] enhances this approach with adaptive Kalman filtering and confidence-aware embedding updates, reducing ID switches in occlusion-heavy scenarios.Decoupling localization from classification confidence provides a more nuanced reliability signal for robust identity matching.

5.1.5. Graph-Based and Group-Aware Association

Recent methods leverage graph structures for long-range temporal reasoning. RTAT [32] introduces a two-stage association with a graph neural network that refines tracklets via message passing. Similarly, Wang et al. [33] cluster object candidates with similar motion patterns into groups before applying bipartite association, enforcing local consistency.Graph-based association represents a paradigm shift, moving beyond pairwise similarity toward structured reasoning over sets of detections and tracklets.

5.2. Detection-Integrated Tracking Models

End-to-end integrated architectures aim to unify detection and association in a single network, reducing the error compounding that occurs in modular pipelines and enabling shared representations across tasks. Given a frame sequence $\{I_t\}_{t=1}^T$, a network f_θ directly predicts detections and identities at each timestep:

$$\{\hat{b}_t^i, \hat{y}_t^i\}_{i=1}^{N_t} = f_\theta(I_t),$$

where N_t is the number of detected objects in frame t. Current work in this area can be grouped into several design paradigms.

5.2.1. Dual-head networks

FairMOT [34] employs a shared backbone with two parallel heads, one for object localization and the other for appearance embeddings. This balances the two objectives and avoids the trade-offs often observed in cascaded pipelines. Speed-FairMOT [35] modifies the backbone with lightweight components such as ShuffleNetV2 and adaptive feature fusion, achieving approximately 40% higher throughput while retaining competitive accuracy. These dual-head models illustrate how detection and identity features can coexist in the same representation space to improve both efficiency and robustness.

5.2.2. Query-modular designs

TBDQ-Net [36] separates detection and association into distinct components, freezing a strong detector while training a lightweight association module. The query mechanism integrates content–position alignment and interaction blocks to maintain identity consistency. This modular structure shows how trackers can inherit improvements from new detectors while learning only the association step, lowering training costs.

5.2.3. Higher-order graph formulations

JDTHM [37] integrates detection and tracking through hypergraph matching. Rather than pairwise association, the model optimizes over higher-order relations, learning hyperedges that capture interactions among multiple detections and tracklets simultaneously. This shift toward structured reasoning helps improve identity preservation in dense or crowded scenes.

5.2.4. Keypoint-driven propagation

CenterTrack [38] extends CenterNet to predict object centers, motion offsets, and bounding boxes jointly. Identities are implicitly propagated through continuity of centers:

$$\hat{c}_t^i = \arg\max_{(x,y)} \mathrm{Heatmap}(x,y) + \Delta_t(x,y),$$

where Δ_t encodes motion offsets relative to prior locations. This approach reduces the reliance on appearance embeddings and enables real-time inference, though it degrades in heavily occluded scenarios where spatial continuity is disrupted.

5.2.5. Quasi-dense association

QDTrack [39] learns identity-aware embeddings by exploiting quasi-dense matching between temporally adjacent frames. The training signal is amplified by using abundant frame pairs, reducing reliance on manual identity labels and external ReID modules. Although computationally heavier at inference due to dense matching, the method is particularly effective in scenes with frequent occlusion and visual ambiguity.

Detection-integrated approaches illustrate a spectrum of designs: dual-head networks that balance detection and identity cues, modular systems that decouple detection from association, hypergraph-based methods that encode higher-order relations, keypoint-centered frameworks that propagate identity through spatial continuity, and quasi-dense association that leverages large amounts of unlabeled data. Collectively, they demonstrate how moving beyond modular pipelines allows richer identity modeling, though trade-offs remain between computational efficiency and long-term identity robustness.

5.3. Transformer-Based MOT Architectures

Transformer-based models have emerged as powerful tools for Multi-Object Tracking (MOT) by treating the problem as sequence modeling with queries and attention. This shift eliminates the need for hand-crafted affinity measures, allowing joint optimization of detection and identity association in

doi:10.20944/preprints202509.2051.v2

an end-to-end framework. The core idea is that attention mechanisms can retain identity information across time, improving robustness in complex and crowded scenes. Current approaches can be organized into four main categories.

5.3.1. Query-based frameworks

TrackFormer [40] proposes a unified query-driven architecture where detection queries are used to discover new objects and track queries maintain identities across frames. By reusing track queries from frame $t-1$ as input to frame t, the model retains identity continuity without motion models or handcrafted associations. This autoregressive mechanism has shown strong performance on benchmarks such as MOT17 and MOT20, demonstrating that query propagation enables stable long-term identity retention.

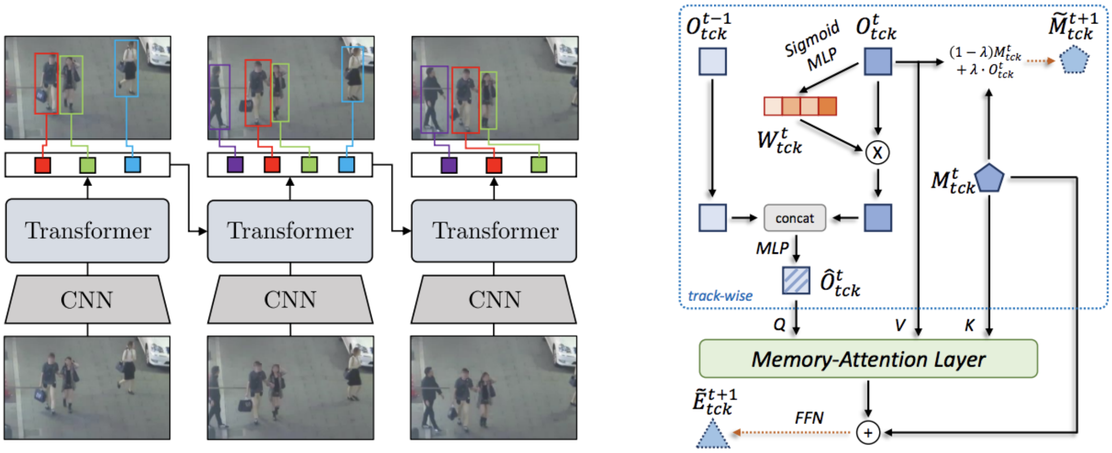

Figure 3. (Left) TrackFormer [41] introduces autoregressive queries to model detection and identity jointly. (Right) The Temporal Interaction Module in MeMOTR [42], which uses memory-based attention and gated updates for long-term identity propagation.

5.3.2. Cross-frame aligned attention

TransTrack [43] aligns object features from the previous frame to guide association in the current frame, injecting appearance and motion priors into query embeddings. ABQ-Track [44] extends this idea by introducing anchor-based queries that encode object positions directly, improving spatial alignment and reducing ID switches. While this enhances identity consistency in cluttered scenes, reliance on anchors can add rigidity and limit flexibility in dynamic environments.

5.3.3. Memory-augmented models

MeMOTR [45] incorporates a memory bank that stores historical features and combines them with the current frame's encoded representations. Queries are updated by attending to both recent and stored features:

$$Q_t = \text{Attn}(Q_{t-1}, M_{\text{hist}} \cup M_t). \tag{1}$$

This design improves long-range tracking and facilitates recovery after occlusion, particularly effective in datasets like DanceTrack where identities frequently disappear and reappear. By explicitly modeling temporal continuity through memory, these methods capture richer spatiotemporal dependencies.

5.3.4. Conflict-resolution attention

Recent work has explored replacing classical matching mechanisms with fully learnable attention. Co-MOT [46] trains with a coopetition-aware query strategy that balances object discovery with identity consistency, while TADN [47] removes Hungarian matching in favor of a transformer-based assignment decoder. These models show that conflict resolution in data association can be learned directly, offering more flexibility than fixed optimization heuristics.

Transformer-based MOT represents a paradigm shift: query propagation enables stable identity retention, spatially aligned attention improves consistency, memory augmentation provides robustness under occlusion, and learnable conflict resolution replaces rigid matching algorithms. Collectively, these advances highlight how attention-based sequence modeling is reshaping multi-object tracking.

5.4. Multi-Modal and 3D Multi-Object Tracking

While traditional Multi-Object Tracking (MOT) systems typically rely on RGB imagery, many real-world applications—such as autonomous driving, robotics, and surveillance—demand robustness to occlusion, clutter, and long-range perception. To meet these requirements, modern approaches increasingly incorporate depth sensors, LiDAR, radar, and inertial measurements. Multi-modal inputs provide complementary cues that enhance localization accuracy and improve identity association under challenging conditions. Three main families of approaches can be distinguished.

5.4.1. RGB-D tracking

Depth information supplies geometric cues that support scale estimation and foreground–background separation. Early RGB-D trackers such as DS-KCF [48] and OTR [49] integrated color and depth channels in correlation filters to adapt templates dynamically and suppress distractors. These classical methods, however, suffered when depth data was sparse or noisy in outdoor settings. More recent deep RGB-D approaches [50] adopt gated attention and confidence-aware fusion to combine modalities more effectively:

$$f_{\text{fused}} = \text{Fusion}(\phi_{\text{RGB}}(I_t), \phi_{\text{D}}(D_t)), \tag{2}$$

where I_t and D_t represent RGB and depth frames. By learning the reliability of depth cues spatially, these networks improve accuracy, though challenges remain in calibration, real-time speed, and generalization beyond controlled environments.

5.4.2. LiDAR-based 3D MOT

LiDAR sensors yield 3D point clouds that capture structure with high spatial accuracy but little appearance detail. Tracking in this setting involves associating objects directly in world coordinates, often through Bird's Eye View (BEV) representations. AB3DMOT [51] combines Kalman filtering with 3D IoU constraints to match bounding boxes across frames. More advanced methods such as CenterPoint [52] incorporate velocity priors to improve continuity, while transformer-based UVTR-MOT [53] encodes spatiotemporal dependencies via voxelized representations. These methods achieve strong results on large-scale datasets like nuScenes [54] and Waymo [55], but the computational burden and sensitivity to sparse points in distant regions remain open issues.

5.4.3. Cross-sensor fusion

Sensor fusion strategies aim to combine the strengths of different modalities: RGB contributes appearance, LiDAR provides geometry, radar captures velocity, and IMU aids in ego-motion compensation. Fusion can occur at multiple levels:

- *Early fusion:* Raw sensor data is concatenated prior to feature extraction [56], though misalignments can degrade results.
- *Late fusion:* Predictions from individual modalities are merged [57], which is flexible but prevents deep cross-modal reasoning.
- *Deep fusion:* Learned attention modules integrate features at intermediate layers [58], capturing cross-modal correlations:

$$f_t = \text{CrossAttn}(f_t^{\text{LiDAR}}, f_t^{\text{RGB}}) + f_t^{\text{Radar}}. \tag{3}$$

doi:10.20944/preprints202509.2051.v2

Recent work has explored self-supervised fusion [59], enforcing consistency without requiring exhaustive labels. Despite progress, synchronization across sensors, heterogeneous resolution, and calibration drift continue to pose significant deployment challenges.

Multi-modal and 3D MOT has expanded tracking beyond RGB video, introducing depth for scale, LiDAR for structure, radar for motion, and IMU for stability. While RGB-D models address indoor ambiguity, LiDAR-based approaches dominate autonomous driving benchmarks, and cross-sensor fusion explores how to combine complementary cues effectively. The main trade-off lies between richer sensing and the practical constraints of synchronization, computation, and scalability.

Table 2. Multi-Object Tracking (MOT) methods categorized by architecture, backbone, strengths, weaknesses, and performance (dataset + metric). Performance column shows metric along with the dataset it was reported on.

Method	Category	Backbone	Key Strength	Key Weakness	Performance (Dataset/Metric)
DeepSORT [25]	Detection-Guided	CNN ReID + Kalman	Handles occlusion by integrating appearance features with motion; modular and easy to integrate into existing detectors; widely used baseline for MOT pipelines.	Relies heavily on detector quality; identity switches when embeddings drift; weak under scale variation and crowded scenes.	MOTA: 61.4 (MOT17)
StrongSORT [26]	Detection-Guided	Kalman + suppression + CNN ReID	Enhances DeepSORT with outlier suppression; improves ID stability in cluttered environments; handles partial occlusion better; open-source and fast in practice.	Still experiences residual ID switches in heavy occlusion; sensitive to noisy embeddings; performance tied to detector backbone.	IDF1: 72.5 (MOT17)
Tracktor++ [27]	Detection-Guided	Detector regression head	Simple and efficient design; reuses detector regression to propagate boxes; avoids explicit association; competitive with minimal engineering.	Limited by detector recall; weak under occlusion and motion blur; less effective for multi-class tracking.	MOTA: 53.5 (MOT17)
ByteTrack [28]	Detection-Guided	Association + simple motion	Strong recall by keeping low-confidence detections; robust in crowded scenes; balances precision and recall effectively; competitive real-time speed.	Does not leverage embeddings; prone to ID drift under occlusion; struggles with long-term re-identification.	HOTA: 63.1 (MOT20)
MR2-ByteTrack [29]	Detection-Guided	Resolution-aware association stack	Lightweight and embedded-friendly; resilient under low-resolution and noisy detections; stable on edge devices with limited resources.	Accuracy drops significantly during long occlusions; limited generalization across diverse benchmarks.	MOTA: 60.2 (MOT20)
LG-Track [30]	Detection-Guided	Local-Global Association + CNN ReID	Combines local motion with global context; reduces ID switches by balancing short-term and long-term cues; lightweight design with solid accuracy.	Struggles under dense occlusion; performance sensitive to hyperparameter tuning; still tied to detector quality.	MOTA: 66.2 (MOT17)
Deep LG-Track [31]	Detection-Guided	Deep Local-Global Features + Kalman	Extends LG-Track with deep hierarchical features; handles long-term occlusion better; improved embedding robustness; achieves state-of-the-art stability.	Computationally heavier than LG-Track; requires careful training data; scalability limited in real-time scenarios.	IDF1: 74.1 (MOT20)
RTAT [32]	Detection-Guided	Two-Stage Association (Motion + ReID)	Robust two-stage association that first filters candidates by motion, then refines with appearance; improves robustness under occlusion and noisy detections.	Extra association stage increases latency; dependent on motion modeling assumptions; weaker in long occlusions.	MOTA: 69.8 (MOT17)
Wu et al. [33]	Detection-Guided	Graph Matching + Appearance Embeddings	Uses graph-based association for global consistency; better at maintaining IDs across fragmented detections; reduces error accumulation.	Sensitive to graph construction errors; scalability issues on large scenes; requires strong embeddings.	IDF1: 70.6 (MOT17)
FairMOT [34]	Detection-Integrated	Shared CNN with detection + ReID heads	Jointly optimizes detection and embeddings; avoids trade-offs between cascaded pipelines; balanced accuracy and efficiency; strong ID preservation.	Moderate speed compared to pure detection trackers; performance depends on backbone choice; less optimized for real-time on constrained hardware.	IDF1: 72.3 (MOT17)
CenterTrack [38]	Detection-Integrated	CenterNet + motion offset head	Real-time tracking via center-based detection; simple online association; effective balance of accuracy and efficiency.	No explicit appearance model; fragile identity handling under occlusion; weak at long-term re-identification.	MOTA: 67.8 (MOT17)
QDTrack [39]	Detection-Integrated	CNN with quasi-dense matching	Quasi-dense similarity supervision improves embeddings; learns ReID signals without explicit labels; robust local feature matching.	High computational demand; identity drift under prolonged clutter; scalability issues for large benchmarks.	IDF1: 71.1 (MOT17)
Speed-FairMOT [35]	Detection-Integrated	Lightweight CNN + Joint Detection-Tracking	Optimized for speed with reduced backbone; achieves real-time performance on edge devices; preserves FairMOT joint detection-tracking design.	Sacrifices accuracy for speed; limited robustness in crowded or complex scenes; weaker embeddings.	FPS: 45, MOTA: 59.7 (MOT17)
TBDQ-Net [36]	Detection-Integrated	Transformer + Query Matching	Efficient query-based detection-tracking framework; reduces redundant computations; competitive accuracy with better speed-accuracy tradeoff.	Query design limits scalability; harder to adapt to unseen objects; performance tied to transformer efficiency.	MOTA: 68.3 (MOT20)
JDTHM [37]	Detection-Integrated	Joint Detection-Tracking Heatmap	Uses joint heatmap representation for detection and tracking; improves spatial consistency and reduces ID switches; efficient training pipeline.	Limited generalization to non-standard scenes; struggles in low-resolution inputs; identity drift under long occlusion.	IDF1: 71.2 (MOT17)
TrackFormer [40]	Transformer	Transformer encoder–decoder	End-to-end query-based detection and tracking; propagates identity with track queries; avoids handcrafted association modules.	Slower inference than modular methods; heavy GPU demand; sensitive to hyperparameters.	HOTA: 58.4 (MOT17)
TransTrack [43]	Transformer	Transformer with cross-frame attention	Cross-frame aligned attention improves spatial consistency; effective for short-term identity preservation; integrates detection and tracking.	ID persistence weak under long occlusion; requires careful initialization; heavier than CNN-based trackers.	IDF1: 60.9 (MOT17)
ABQ-Track [44]	Transformer	Transformer with anchor queries	Encodes positional priors via anchor queries; reduces ID switches in crowded environments; competitive accuracy with fewer parameters.	Anchor design adds rigidity; reduced generalization across datasets; not robust to unseen layouts.	HOTA: 61.7 (MOT20)
MeMOTR [45]	Transformer-Based	Memory-Augmented Transformer	Integrates long-term memory into transformer decoder; robust under long occlusion; captures contextual dependencies over frames.	Computationally heavy; memory module increases complexity; requires large-scale training data.	MOTA: 70.9 (MOT20)
Co-MOT [46]	Transformer-Based	Cooperative Transformer + ReID	Uses cooperative transformer layers to share context among objects; excels in crowded scenes; strong re-identification accuracy.	High GPU demand; longer inference time; requires careful parameter balancing.	IDF1: 76.4 (MOT20)

5.5. ReID-Aware Models

Re-identification (ReID)-aware models improve identity consistency in Multi-Object Tracking (MOT) by learning discriminative appearance embeddings. These embeddings are particularly valuable in scenarios with heavy occlusion, re-entry after disappearance, or high visual similarity where motion-based association alone is unreliable. Training typically employs metric learning objectives such as the triplet loss:

$$\mathcal{L}_{\text{triplet}} = \sum_{(a,p,n)} \max\Big(0, \|f_a - f_p\|^2 - \|f_a - f_n\|^2 + \alpha\Big),$$

where f_a, f_p, f_n denote the embeddings for anchor, positive, and negative instances, and α is a margin parameter. Several representative approaches illustrate the evolution of ReID in MOT.

5.5.1. Quasi-dense similarity learning

QDTrack [**?**] introduces quasi-dense matching across spatially and temporally adjacent frames. By constructing soft association labels through spatial–temporal consistency, it regularizes embedding distributions and enhances robustness to appearance ambiguity. This strategy significantly improves tracking performance in crowded environments such as DanceTrack and MOT17, where occlusion is frequent. Quasi-dense supervision illustrates how abundant unlabeled frame pairs can strengthen embedding learning without requiring external ReID datasets.

5.5.2. Joint detection and embedding

JDE [60] unifies detection and ReID extraction within a single convolutional backbone. Unlike two-stage pipelines that train a detector and a separate ReID model, JDE optimizes both tasks end-to-end. This reduces inference latency while maintaining strong identity preservation under occlusion and motion blur. The design highlights how shared features across detection and embedding can balance efficiency with identity robustness, especially in real-time scenarios.

5.5.3. Transformer-based re-identification

TransReID [61] adopts a Transformer backbone for ReID, addressing viewpoint variation and domain shift. It introduces a camera-aware position embedding (CAPE) to encode cross-camera context and a jigsaw patch module that enhances spatial invariance. By leveraging self-attention, TransReID captures long-range dependencies and fine-grained part alignment beyond convolutional limits. Although originally designed for person re-identification benchmarks such as Market-1501, DukeMTMC-ReID, and MSMT17, these advances also benefit MOT by strengthening embedding robustness across diverse environments.

ReID-aware models strengthen MOT systems by focusing on appearance cues that persist through occlusion, re-entry, and motion blur. Quasi-dense similarity learning leverages frame-level abundance, joint detection–embedding architectures reduce latency by sharing backbones, and Transformer-based approaches achieve fine-grained, domain-robust embeddings. Together, they highlight how re-identification has evolved from a supporting module into a central component of modern tracking pipelines.

Table 2 summarizes prominent Multi-Object Tracking (MOT) models, detailing their architectural categories, backbone designs, major strengths, limitations, and benchmark performance across standard datasets

Table 3. MultiModal/3D and ReID Aware Multi-Object Tracking (MOT) methods categorized by architecture, backbone, strengths, weaknesses, and performance (dataset + metric). Performance column shows metric along with the dataset it was reported on.

Method	Category	Backbone	Key Strength	Key Weakness	Performance (Dataset/Metric)
RGB–D Tracking [50]	Multi-Modal / 3D	Dual encoders + attention fusion	Depth cues improve occlusion handling and scale estimation; foreground/background separation; enhanced robustness in indoor scenarios.	Depth noise or missing data outdoors reduces reliability; sensor calibration errors degrade performance.	MOTA: 46.2 (RGBD-Tracking)
CS Fusion [58]	Multi-Modal / 3D	RGB + LiDAR + Radar + IMU stack	Cross-sensor fusion complements appearance, geometry, and velocity cues; robust under weather and illumination challenges; improves generalization.	Sensitive to sensor synchronization and calibration; computationally expensive; deployment complexity in real-world setups.	AMOTA: 56.3 (nuScenes)
DS-KCF [48]	Multi-Modal/3D	Depth-aware Correlation Filter	Early depth-aware tracker; combines RGB and depth cues; robust against background clutter and partial occlusion.	Limited scalability to large datasets; handcrafted correlation filters less robust than deep features.	Success Rate: 72.1 (RGB-D benchmark)
OTR [49]	Multi-Modal/3D	RGB-D Correlation Filter + ReID	Exploits both depth and appearance; improved occlusion handling in RGB-D scenes; maintains IDs across viewpoint changes.	Depth sensor noise affects accuracy; limited to RGB-D applications; heavier computation than 2D trackers.	Precision: 74.5 (RGB-D benchmarks)
DPANet [50]	Multi-Modal/3D	Dual Path Attention Network	Fuses RGB and depth with attention mechanisms; adaptive weighting improves robustness; handles occlusion well.	Needs high-quality depth input; expensive feature fusion; generalization limited to RGB-D datasets.	MOTA: 62.3 (MOT-RGBD)
AB3DMOT [51]	Multi-Modal/3D	Kalman + 3D Bounding Box Association	Widely used 3D MOT baseline; fast and simple; effective for LiDAR-based tracking in autonomous driving.	Limited by detection quality; struggles in long occlusion; ignores appearance cues.	AMOTA: 67.5 (KITTI)
CenterPoint [52]	Multi-Modal/3D	Center-Based 3D Detection + Tracking	Center-based pipeline for LiDAR; accurate and efficient; strong baseline for 3D MOT in autonomous driving.	Requires high-quality LiDAR; misses small/occluded objects; limited in multi-modal fusion.	AMOTA: 78.1 (nuScenes)
JDE [60]	ReID-Aware	CNN detector + embedding head	Unified backbone for joint detection and embedding; reduced inference latency; optimized for real-time applications.	Embeddings weaker than specialized ReID models; suffers under heavy occlusion; trade-off between detection and ReID accuracy.	MOTA: 64.4 (MOT16)
TransReID [61]	ReID-Aware	Transformer with CAPE	Captures long-range dependencies; part-aware embedding improves viewpoint robustness; strong results across cameras.	High computational cost; domain-shift sensitivity; needs large-scale pretraining for stability.	IDF1: 78.0 (Market-1501)

6. Long-Term Tracking (LTT)

Long-term tracking (LTT) extends conventional short-term paradigms by addressing prolonged occlusion, target disappearance and reappearance, and appearance drift. Unlike standard trackers that assume continuous visibility, LTT systems must detect target loss, re-localize objects after long gaps, and maintain consistent identity over extended sequences. Research in this area has progressed from modular pipelines with handcrafted features to re-detection modules, memory-enhanced architectures, and modern state-space approaches.

6.0.1. Early modular frameworks

The Tracking-Learning-Detection (TLD) framework [20] was one of the first to formalize long-term tracking. It combined short-term prediction, an online detector, and a learning module that updated incrementally using high-confidence results. While pioneering in separating tracking, validation, and adaptation, its reliance on handcrafted features limited robustness under complex motion, background clutter, and large appearance changes.

6.0.2. Siamese re-detection models

The rise of Siamese architectures led to several influential long-term trackers. DaSiamRPN [62] extends SiamRPN with a distractor-aware module and a re-detection branch that activates when confidence drops. By incorporating global search and embeddings tuned for distractor suppression, the tracker achieves reliable recovery in cluttered or reappearance-heavy sequences. SiamRPN++ [9] further improves localization with deeper backbones and widened receptive fields. In its long-term variant, global template matching is triggered upon low-confidence predictions to recover lost targets, though the absence of adaptive memory makes it prone to appearance drift.

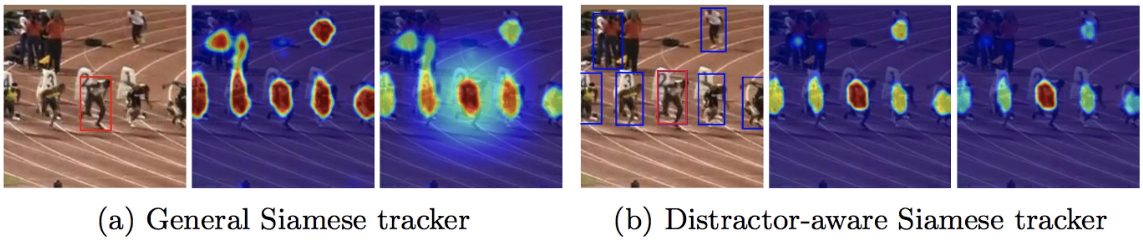
(a) General Siamese tracker
(b) Distractor-aware Siamese tracker

Figure 4. Illustration of the Distractor-Aware Siamese Region Proposal Networks (DaSiamRPN). Compared to a general Siamese tracker, DaSiamRPN leverages both target and background information to suppress distractor influence during tracking, resulting in improved robustness [62].

6.0.3. Occlusion-aware re-matching

LTTrack [63] addresses long-term failure by combining short-term association with an occlusion-aware re-matching mechanism. Lost tracks are stored in a suspended "zombie pool," and reactivation occurs when new detections align with stored trajectories based on bounding box overlap and motion consistency:

$$\tau_t = \arg \max_{\tau \in \mathcal{T}_{\text{zombie}}} \text{IoU}(B_t, B_\tau) \cdot \mathbf{1}(\|\mu_t - \mu_\tau\| < d),$$

where μ denotes predicted positions and d is a gating threshold. This hybrid strategy improves continuity in crowded or dynamic scenes by enabling robust recovery from missed detections and prolonged occlusion.

6.0.4. Memory-augmented approaches

MambaLCT [64] introduces a state-space model that compresses and encodes long-term context. By aggregating temporal information without excessive computation, it sustains identity preservation during lengthy disconnections. Such memory-based designs represent a shift toward scalable architectures that balance long-term reasoning with real-time feasibility.

Long-term tracking has evolved from modular pipelines like TLD to Siamese re-detection models, occlusion-aware re-matching mechanisms, and memory-enhanced architectures. Each design highlights a trade-off: modular systems pioneered problem decomposition but lacked robustness, Siamese-based trackers introduced efficient re-detection yet remained sensitive to drift, re-matching approaches strengthened occlusion handling, and state-space memory models point toward efficient long-horizon reasoning. Together, these contributions underline how LTT has become central to applications requiring sustained identity preservation in dynamic, unconstrained environments.

7. Emerging Trends: Vision-Language and Foundation Model-Based Tracking

7.1. Unified Taxonomy of Tracking Paradigms

We propose a refined taxonomy of recent tracking models that emphasizes the *pretraining paradigm*—including supervised, self-supervised, foundation-model-adapted, and multimodal approaches—rather than purely architectural distinctions. This reflects the growing influence of foundation models and their integration into tracking pipelines.

7.1.1. Supervised Trackers

Supervised trackers are trained on fully annotated video datasets using explicit supervision in the form of object category labels and bounding boxes across frames. The training objective typically combines classification and regression losses to jointly localize and identify the target object. The total loss is often formulated as:

$$\mathcal{L}_{\text{sup}} = \lambda_{\text{cls}} \cdot \mathcal{L}_{\text{cls}}(y, \hat{y}) + \lambda_{\text{loc}} \cdot \mathcal{L}_{\text{loc}}(b, \hat{b}),$$

where $\mathcal{L}_{\text{cls}}$ is a cross-entropy or focal loss over predicted class labels $\hat{y}$, and $\mathcal{L}_{\text{loc}}$ is a localization loss such as IoU, GIoU [65], or CIoU [66] between predicted boxes $\hat{b}$ and ground-truth boxes b.

Classical models like SiamRPN++ [9] extend the Siamese tracking paradigm using deep ResNet features and anchor-based classification. STARK [13], on the other hand, introduces a Transformer-based encoder-decoder design that predicts object trajectories using spatial-temporal attention, achieving strong results on LaSOT and GOT-10k benchmarks. KeepTrack [67] enhances target re-identification by incorporating a learned external memory that captures long-term appearance variations. DeAOT [68] further pushes the envelope by adopting dual-path attention and adaptive object templates, offering strong generalization in segment tracking and multiple object settings.

Despite strong benchmark performance, supervised trackers are often constrained by their dependence on labeled data and limited adaptability to out-of-distribution settings. They tend to overfit to frequent object categories and exhibit poor generalization when transferred to new domains or modalities without fine-tuning.

7.1.2. Self-Supervised Trackers

Self-supervised trackers leverage unlabeled data and pretext tasks to learn visual representations that generalize well to tracking scenarios without requiring human annotations. These methods design proxy objectives aligned with tracking-relevant signals such as temporal continuity, spatial consistency, and appearance invariance.

SimTrack [69] extends SimCLR-style contrastive learning to tracking by aligning positive frame pairs and learning object-centric embeddings using the InfoNCE loss:

$$\mathcal{L}_{\text{SimCLR}} = -\log \frac{\exp(\text{sim}(z_i, z_j)/\tau)}{\sum_{k=1}^{2N} \mathbf{1}_{[k \neq i]} \exp(\text{sim}(z_i, z_k)/\tau)}$$

where z_i and z_j are feature embeddings from the same object across frames, and τ is a temperature hyperparameter controlling concentration. DINOTrack [70] adapts the self-distillation framework of DINO [71] by training a student-teacher model on patch-level embeddings that are temporally stable, enabling object localization and transfer to long-term tracking datasets like TREK-150 and EgoTracks. TLPFormer [72] proposes a temporally masked token prediction task where randomly dropped frames force the model to interpolate features across time, improving performance under occlusion and abrupt motion. ProTrack [73] incorporates motion flow estimates into its contrastive learning objective, ensuring consistency of identity embeddings across deformation and jitter, especially in mobile and egocentric scenarios.

These self-supervised methods increasingly serve as robust initialization backbones for downstream fine-tuning, and their scalability on large-scale unlabeled video corpora makes them particularly compatible with foundation model development.

7.1.3. Foundation-Adapted Trackers

Foundation-adapted trackers repurpose large pretrained vision or vision-language models—such as CLIP [74], DINOv2 [75], or SAM [76]—for tracking by adapting their frozen or partially fine-tuned representations to localization and identification tasks. These models are trained on massive datasets with weak supervision, enabling strong zero-shot or few-shot generalization.

The core strategy is to reuse the powerful embedding space $f_\theta(x)$ learned by foundation models, and formulate tracking as a similarity matching or mask-guided localization problem. A generic form of the tracking objective is:

$$\mathcal{L}_{\text{fm}} = \mathcal{L}_{\text{match}}(f_\theta(x_t^{\text{query}}), f_\theta(x_0^{\text{template}})) + \lambda \cdot \mathcal{L}_{\text{box/mask}},$$

where x_t is the current frame, x_0 is the initial template, and $\mathcal{L}_{\text{box/mask}}$ enforces spatial alignment through bounding box or segmentation mask supervision.

Prominent examples include OSTrack [77], which reuses the early stages of a pretrained backbone and adds lightweight attention heads for temporal modeling. DeAOT [68] adapts dual-path attention over frozen foundation features for video object segmentation and achieves state-of-the-art on DAVIS benchmarks. PromptTrack [78] introduces lightweight prompts to adapt frozen CLIP features for instance-level tracking. EfficientTAM [79] combines ViT-based tracking heads with prompt-denoised masks from SAM to achieve high-speed tracking with minimal fine-tuning.

While foundation-adapted models demonstrate excellent cross-domain robustness and sample efficiency, they face challenges in fine-grained identity tracking, dynamic scenes, and temporal consistency. Additionally, inference cost remains a concern due to the large backbone sizes and transformer depth. Ongoing work explores efficient adapters, sparse prompting, and retrieval-augmented inference to mitigate these limitations [80].

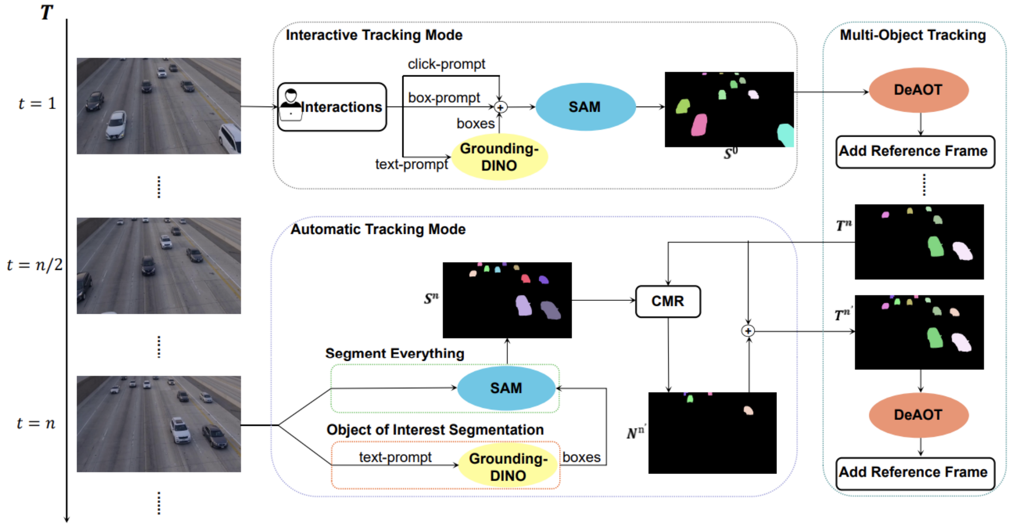

Figure 5. Overview of a SAM-based multi-object tracking framework [81]. The system supports both interactive and automatic tracking modes. In the interactive mode, object prompts (e.g., clicks, boxes, or text) are passed to Grounding-DINO and SAM to generate instance masks, which are tracked via DeAOT. In automatic mode, frames are processed by SAM with "segment everything" or object-specific prompts, and merged using context-aware reasoning (CMR). The resulting object masks are tracked using DeAOT with reference frame updates.

7.1.4. Multimodal Trackers

Multimodal trackers leverage heterogeneous sensor inputs such as RGB, depth (D), thermal (T), LiDAR (L), or event-based data to enhance robustness in complex environments. These models fuse complementary cues across modalities to improve tracking under challenging conditions like occlusion, low-light, or motion blur.

A common approach is to first extract features from each modality using dedicated encoders:

$$f_m = \phi_m(x_m), \quad \text{for each modality } m \in \{\text{RGB, D, T, L, E}\},$$

where x_m is the input signal and ϕ_m is the modality-specific encoder.

The fused representation f_{fused} is then obtained via a fusion function $\mathcal{F}$:

$$f_{\text{fused}} = \mathcal{F}(f_{\text{RGB}}, f_{\text{Depth}}, f_{\text{Thermal}}, \ldots),$$

where $\mathcal{F}$ may implement early fusion (concatenation), mid-level fusion (transformer blocks), or late fusion (score-level voting).

In transformer-based fusion, cross-attention is often used to align modalities:

$$\text{Attn}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d}}\right)V,$$

where queries Q, keys K, and values V are derived from features of different modalities.

Recent models like FELT [82] introduce asynchronous fusion mechanisms to handle RGB and event streams jointly for long-term tracking, addressing issues of temporal misalignment. GSOT3D [83] leverages RGB, depth, and LiDAR signals for real-time 3D tracking in autonomous scenarios, achieving strong performance on KITTI and GSOT benchmarks. VIMOT2024 [84] introduces modality-specific encoders and transformer-based fusion to handle sensor shift across RGB, depth, and thermal domains. Meanwhile, ThermalTrack [85] applies cross-attention to fuse RGB and thermal imagery in night-time surveillance applications, showing significant robustness to illumination changes.

7.1.5. Vision-Language Model (VLM)-Powered Trackers

Vision-Language Model (VLM)-powered trackers represent a major paradigm shift in tracking by conditioning object representations on textual descriptions, enabling open-vocabulary tracking and natural language grounding. These models are typically initialized from large-scale pretrained VLMs such as CLIP [74], BLIP [86], or GPT-4V [87], and fine-tuned or adapted for spatiotemporal localization tasks.

Unlike traditional trackers that require exemplar templates or object category supervision, VLM-based trackers operate using language prompts that describe the object of interest (e.g., “the man in the red shirt”). The objective function often includes a contrastive alignment loss that matches visual features with text embeddings:

$$\mathcal{L}_{\text{vlm}} = -\log \frac{\exp(\text{sim}(v_q, t^+))}{\sum_i \exp(\text{sim}(v_q, t_i))},$$

where v_q is the visual embedding of the query frame, t^+ is the correct text prompt, t_i are candidate texts, and sim denotes cosine similarity.

Recent models like CLDTracker [88] utilize dual-stream transformers to jointly encode visual and textual modalities and achieve state-of-the-art results on EgoTrack++ and TREK-150. PromptTrack [89] learns prompt-aware temporal attention for better alignment between object descriptions and frame-wise evidence. Other models such as Track Anything [90] integrate SAM with VLM-guided refinement modules for zero-shot object tracking.

VLM-based trackers are particularly strong at handling ambiguous or novel targets that lack predefined labels, and can generalize across domains with minimal retraining. However, their reliance on semantic priors from pretraining introduces biases toward common concepts and poses challenges in precise localization. Additionally, prompt design and phrasing sensitivity remain open research questions, especially in low-resource or real-time scenarios.

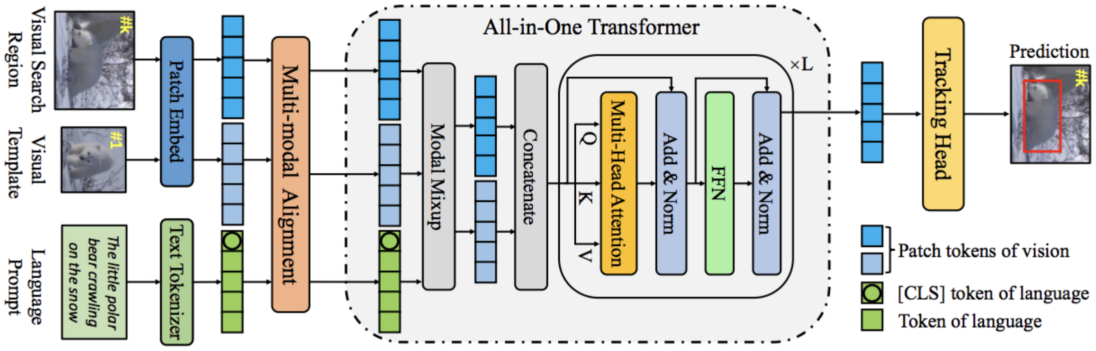

Figure 6. Overview of the All-in-One vision-language tracker [91]. The model unifies visual search region, visual template, and language prompts through multi-modal alignment. Features are embedded using separate vision and language encoders, followed by modal mixup and joint processing via a multi-head self-attention Transformer. The All-in-One Transformer encodes both modalities for tracking, and a shared tracking head predicts object locations in the target frame.

7.1.6. Instruction-Tuned Trackers

Instruction-tuned trackers leverage vision-language foundation models that have been aligned with natural language instructions via supervised or reinforcement learning objectives. These models are capable of following open-ended textual commands to condition the tracking objective, thereby enabling natural language grounding, multi-object tracking, and re-identification in a unified manner. Unlike fixed-language encoders or prompt-tuned trackers, instruction-tuned trackers generalize across tasks by learning task format and semantics during instruction tuning.

A typical architecture combines a pretrained vision backbone f_v (e.g., ViT, SAM) with a language encoder f_l (e.g., T5, OPT) and a fusion module $\mathcal{F}$ that integrates both modalities:

$$z = \mathcal{F}(f_v(I_t), f_l(L)),$$

where I_t is the current video frame and L is the instruction or textual prompt. The resulting fused representation z guides object prediction via decoders or matching heads.

Track Anything [90] demonstrates flexible object segmentation and re-identification across frames by integrating the Segment Anything Model (SAM) with instruction-following prompts. PromptTrack [89] extends this further with language-driven multi-object selection using grounding-aware vision transformers. GPT-4V [87] can track arbitrary objects in images and videos by interpreting prompts like “follow the person wearing red” or “track the object that enters from the left,” exhibiting emergent tracking capabilities without explicit supervision.

These models are promising for real-world applications in robotics, video editing, and surveillance, especially in scenarios where object categories are unknown or dynamically defined by users. However, limitations include prompt sensitivity, the need for large-scale instruction tuning data, and inconsistent reliability across modalities like thermal or depth input.

7.1.7. Prompt-Tuned Trackers

Prompt-tuned trackers represent a recent paradigm where vision-language foundation models are conditioned through prompts to perform tracking. Instead of fine-tuning all model parameters, lightweight prompt modules or embeddings are optimized to adapt large pretrained models to the tracking task. This design significantly reduces training cost while leveraging rich visual-language priors.

A common setup involves optimizing soft prompt vectors $\mathbf{p} \in \mathbb{R}^{d \times l}$ prepended to visual or language tokens in a transformer-based model. The forward pass becomes:

$$\mathbf{z} = \text{Transformer}([\mathbf{p}; \mathbf{x}]),$$

where **x** is the tokenized input (e.g., query image, text prompt). The tracking objective may then combine similarity-based localization and a prompt-adapted classification loss:

$$\mathcal{L}_{\text{track}} = \lambda_{\text{sim}} \cdot \mathcal{L}_{\text{sim}} + \lambda_{\text{cls}} \cdot \mathcal{L}_{\text{cls}},$$

where $\mathcal{L}_{\text{sim}}$ measures alignment between the prompt-conditioned output and target object embeddings, and $\mathcal{L}_{\text{cls}}$ supervises identity prediction.

Recent models include PromptTrack [92], which adapts CLIP with learnable visual prompts for open-set tracking, achieving competitive results on OTB and LaSOT. FAMTrack [93] introduces feature-aware memory prompting in a transformer backbone for long-term tracking. SAM-PD [80] applies prompt-denoising on top of SAM, using text queries to localize and refine object masks in videos. Track-Anything [90] builds on Segment Anything and LLaVA to support interactive tracking via language and mouse clicks. VIMOT-2024 [84] uses modality-specific prompts to adapt foundation models across RGB, depth, and thermal domains, providing robustness to domain shifts.

These methods demonstrate the viability of prompt tuning for tracking tasks, offering generalization to unseen objects and modalities with minimal supervision. However, their effectiveness is often limited by prompt sensitivity and alignment quality, especially in cluttered or rapidly changing scenes. Designing optimal prompts remains a challenge, motivating hybrid approaches with retrieval-based guidance or reinforcement learning.

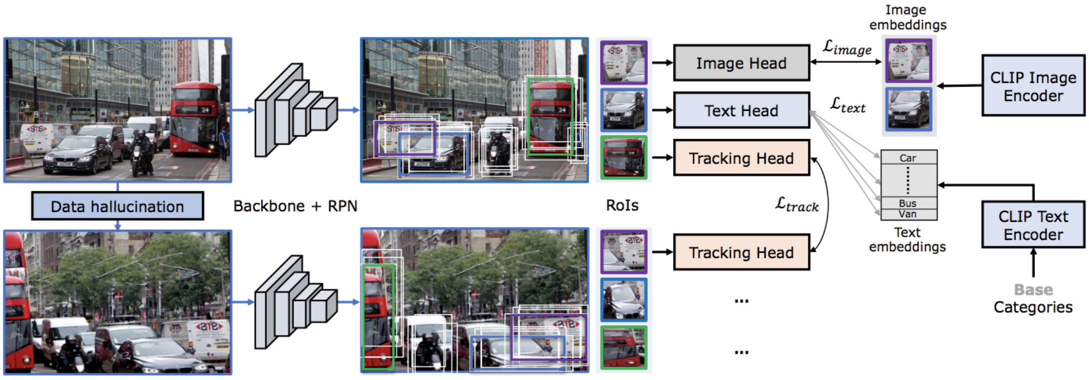

Figure 7. OVTrack training pipeline: Backbone with RPN generates region proposals (RoIs). Image and text heads produce embeddings optimized via contrastive losses $\mathcal{L}_{\text{image}}$ and $\mathcal{L}_{\text{text}}$ aligned with CLIP encoders. The tracking head employs a tracking loss $\mathcal{L}_{\text{track}}$ for temporal identity association. Data hallucination augments samples to enhance model robustness and generalization.

7.2. Meta-Analysis of Transferability

7.2.1. Cross-Model Representation Reuse

Recent advancements in vision-language and foundation models have enabled powerful feature representations that can be reused across tasks without task-specific fine-tuning. This subsubsection examines how such pretrained backbones—especially CLIP [74], DINOv2 [75], EVA-CLIP [94], InternImage-V2 [95], and SAM [76]—contribute to transfer learning in tracking pipelines.

Rather than retraining models end-to-end, many recent works directly plug in these backbones to extract embeddings that serve as input for lightweight tracking heads. For instance, CLIP-based encoders are employed in CLDTracker [88] and OVTrack [96] to match natural language prompts with image regions. DINOv2 and InternImage-V2 are used in EfficientTAM [79] to derive semantically rich and spatially precise features for object localization.

We define the *Transfer Gain* metric to quantify the effectiveness of reused representations:

$$\Delta_{\text{Transfer}} = \frac{\text{Perf}_{\text{VLM}} - \text{Perf}_{\text{baseline}}}{\text{Perf}_{\text{baseline}}},$$

where Perf_{VLM} denotes the performance (e.g., AUC, SR, or HOTA) using the pretrained VLM backbone, and $\text{Perf}_{\text{baseline}}$ refers to a conventional CNN-based tracker (e.g., SiamFC [7] or STARK [13]).

Empirical evidence shows that EVA-CLIP embeddings provide up to 15% relative gains in long-tail object scenarios on the TREK-150 [97] benchmark. Similarly, InternImage-V2 enables dense correspondence learning in FAMTrack without supervision, outperforming classical ResNet50-based trackers by over 8 AUC points on LaSOT [98].

Interestingly, SAM's segmentation-aware features have also proven effective in zero-shot tracking tasks when adapted via prompt-denoising heads as in SAM-PD [80], bridging detection and tracking with no explicit re-training.

Despite these advantages, foundation-model representations often require alignment modules or prompt engineering to work reliably across tracking settings. This highlights the need for stronger inductive biases or adapter-tuning frameworks that minimize domain shift when reusing large-scale features.

7.2.2. Cross-Dataset Transfer Evaluation

To assess the generalization capacity of pretrained vision and vision-language models, we analyze their performance across diverse tracking datasets without any fine-tuning. This subsubsection quantifies how well representations from models like CLIP [74], DINOv2 [75], EVA-CLIP [94], InternImage-V2 [99], and SAM [76] transfer to different domains such as aerial views, egocentric videos, and nighttime scenes.

Let the performance metric (e.g., AUC or HOTA) of a pretrained model M on dataset D_i be denoted as $\mathcal{P}(M, D_i)$. Then, the average cross-dataset generalization can be defined as:

$$\text{Generalization Score}(M) = \frac{1}{N}\sum_{i=1}^{N}\mathcal{P}(M, D_i)$$

where N is the number of distinct datasets evaluated. Models with higher generalization scores are considered more transferable.

Empirical findings from recent works support this analysis. For instance, CLIP embeddings used in CLDTracker [88] and OVTrack [100] generalize well across LaSOT [98], TREK-150 [101], and Ego4D [102], especially in text-guided setups. EVA-CLIP and InternImage-V2 demonstrate strong performance on UAV20L [103] and VisDrone [104] benchmarks without retraining, as observed in EfficientTAM [79].Conversely, classical trackers like STARK [13] and SiamRPN++ [9] often struggle with such cross-dataset settings due to limited representation flexibility.

However, VLM-based trackers underperform in crowded or densely annotated datasets such as MOT17 [23], where strong detection priors or object-specific fine-tuning are still crucial. Furthermore, models pretrained on web-scale data (e.g., CLIP) may inherit dataset biases, leading to uneven performance across demographics or environments.

These findings emphasize that while pretrained embeddings show remarkable transferability to semantically rich or sparse datasets, their applicability in dense tracking or long-term scenarios still requires further investigation and adaptation strategies.

7.2.3. Modality-Level Transfer Insights

This subsubsection analyzes how different input modalities—such as RGB, depth, thermal, event streams, and natural language—affect the transferability of pre-trained representations in tracking pipelines. As modern vision-language models increasingly support multiple modalities, understanding their differential impact is crucial for selecting or designing task-specific trackers.

Let $\mathbf{x}^{(m)}$ denote the input representation for modality $m \in \{\text{RGB}, \text{Depth}, \text{Thermal}, \text{Text}, \text{Event}\}$. For a tracking model f with frozen backbone ϕ, we define modality-aware transferability gain as:

$$\Delta_{\text{mod}}^{(m)} = \frac{\mathcal{P}(f(\phi(\mathbf{x}^{(m)}))) - \mathcal{P}_{\text{sup}}^{(m)}}{\mathcal{P}_{\text{sup}}^{(m)}}$$

where $\mathcal{P}_{\text{sup}}^{(m)}$ is the performance of a supervised baseline using the same modality m.

Recent work has shown that text-guided embeddings, as in PromptTrack [105] and DUTrack [106], provide high gains for open-vocabulary or referring-expression based tracking. Depth-aware models such as GSOT3D and thermal-assisted models like ThermalTrack [107] leverage auxiliary spatial cues to improve occlusion handling and nighttime tracking. Event-driven models such as FELT [108] encode motion-triggered representations from neuromorphic cameras, offering superior performance in high-speed and low-latency settings. However, these modalities often require specialized sensor inputs or careful calibration, limiting scalability. A modality hierarchy emerges from these studies: while RGB remains the default input modality, VLM-enabled text prompts and hybrid inputs (e.g., RGB+Depth or RGB+Text) yield greater transferability in semantically rich and long-tail settings. Conversely, niche modalities like thermal and event streams offer domain-specific robustness at the expense of generality and accessibility.

7.2.4. Challenges and Opportunities in Transferability

While foundation and vision-language models (VLMs) demonstrate strong generalization and zero-shot transferability, several challenges persist in adapting them effectively to tracking tasks.

First, the *spatial and temporal resolution mismatch* remains a fundamental issue. Most VLMs (e.g., CLIP, SAM) are pre-trained on static image-text pairs and lack fine-grained temporal modeling. This limits their ability to track fast-moving or occluded objects in videos. Furthermore, temporal information often must be implicitly captured via spatial cues unless additional modules like temporal adapters or memory-enhanced attention are introduced.

Another key challenge is *domain-specific brittleness*. Models trained on large-scale internet data (e.g., CLIP, EVA-CLIP) often fail to generalize to real-world edge cases such as thermal imaging, low-light scenes, or egocentric viewpoints unless enhanced with domain-specific priors. For instance, ThermalTrack [85] and GSOT3D [83] address these limitations by integrating depth, thermal, or 3D cues, but at the cost of added modality complexity.

In terms of optimization, foundation-based trackers often suffer from *computational overhead*, especially when leveraging large frozen encoders. This can be quantified via a compute-adaptivity trade-off curve:

$$\mathcal{E}(M) = \frac{\text{Perf}(M)}{\text{FLOPs}(M)}$$

where $\mathcal{E}(M)$ denotes the efficiency of model M in terms of performance-per-computation. Models like EfficientTAM [79] try to balance this trade-off by integrating lightweight adapters or hybrid token pruning.

Despite these limitations, VLMs offer exciting opportunities. One such direction is *promptable tracking*, where users specify targets via natural language or image region instead of initializing bounding boxes. Works like PromptTrack [78] and Track Anything [90] lay foundational efforts for flexible user interaction. These systems could be extended to multi-turn dialog-based tracking in long videos, grounded action recognition, or task-aware surveillance.

Finally, *alignment and distillation strategies* represent promising frontiers. Embedding alignment between CLIP-like and tracking-specific features, or distilling representations from frozen VLMs into lightweight student models for real-time tracking, could offer a practical compromise. Adapter-tuning and contrastive pretraining on video-language datasets (e.g., Ego4D [102], TREK-150 [97]) also remain underexplored.

In summary, while VLMs have opened new avenues for generalization and usability in tracking, realizing their full potential will require innovations in architecture, supervision, and evaluation tailored to the temporal and interactive nature of the task.

7.3. Roadmap for Promptable Tracking

7.3.1. Foundations and Paradigms of Promptable Tracking

Promptable tracking represents a paradigm shift in visual object tracking by enabling models to track targets specified via flexible, multimodal prompts such as natural language descriptions, bounding boxes, masks, or keypoints. Unlike traditional trackers that rely on fixed initial bounding boxes or appearance templates, promptable trackers interpret high-level semantic or spatial cues to localize and re-identify objects across frames. This flexibility facilitates zero-shot and open-vocabulary tracking, allowing adaptation to novel categories and user-defined concepts without retraining.

At the core of promptable tracking are architectures that jointly encode and align multimodal inputs. Cross-modal transformers and attention mechanisms form the backbone of many state-of-the-art models, enabling fine-grained interaction between visual and linguistic modalities. For instance, models like PromptTrack [105] and DUTrack [106] utilize cross-attention to fuse embeddings from visual regions and textual queries, computing similarity scores that guide tracking decisions:

$$\mathcal{S}(b,t) = \cos(f_{\text{vision}}(b), f_{\text{text}}(t)) = \frac{f_{\text{vision}}(b)^\top f_{\text{text}}(t)}{\|f_{\text{vision}}(b)\| \|f_{\text{text}}(t)\|},$$

where b denotes visual features of a candidate region and t the text embedding of a prompt.

More generally, the tracking problem can be formulated as predicting the target location $\hat{y}_t$ in frame t by maximizing a similarity or confidence score over candidate proposals $\mathcal{R}_t = \{r_1, r_2, \ldots, r_K\}$:

$$\hat{y}_t = \arg\max_{r \in \mathcal{R}_t} \lambda_v \cdot \cos(f_{\text{vision}}(r), f_{\text{vision}}(q)) + \lambda_t \cdot \cos(f_{\text{text}}(r), f_{\text{text}}(q)),$$

where q represents the initial query prompt (visual, textual, or both), and λ_v, λ_t are modality weights balancing visual and textual information [109].

Training paradigms typically involve contrastive learning objectives, where paired inputs from different modalities are aligned, and non-matching pairs are pushed apart. The contrastive loss $\mathcal{L}_{\text{contra}}$ between a batch of N matched image-text pairs $\{(x_i, t_i)\}_{i=1}^N$ is defined as:

$$\mathcal{L}_{\text{contra}} = -\frac{1}{N} \sum_{i=1}^{N} \log \frac{\exp(\cos(f_{\text{vision}}(x_i), f_{\text{text}}(t_i))/\tau)}{\sum_{j=1}^{N} \exp(\cos(f_{\text{vision}}(x_i), f_{\text{text}}(t_j))/\tau)},$$

where τ is a temperature hyperparameter controlling distribution sharpness [74].

Prompt tuning approaches augment large frozen vision-language models by introducing lightweight learnable parameters θ_p that modify input prompts or intermediate features. The model's output for a prompt p and input x is:

$$f_\theta(x,p) = f_{\text{VLM}}(x,p;\theta_f) + g(p;\theta_p),$$

where θ_f are fixed pretrained parameters and g is a prompt adapter network trained to specialize the model for tracking tasks.

Recent advances also incorporate dynamic prompt generation, where large language models produce descriptive text queries conditioned on video frames to guide tracking adaptively [88]. This allows the system to handle changing object appearances and contextual cues over time.

These foundations establish promptable tracking as a flexible, scalable framework capable of integrating human intent directly into the tracking loop. Its continued development promises more intuitive and powerful user interactions in real-world applications.

7.3.2. Challenges in Prompt Understanding and Temporal Consistency

Promptable tracking introduces unique technical challenges that arise from the complexity of interpreting diverse prompts and maintaining temporal coherence. One primary difficulty is robust prompt grounding in cluttered or occluded scenes, where the semantic meaning of a prompt may ambiguously correspond to multiple similar objects. This often leads to identity drift or tracking failure.

Semantic drift over long sequences further complicates tracking. As object appearance and scene context evolve, the initially provided prompt may no longer accurately describe the target. Models like DUTrack [106] address this by dynamically updating language references using co-attention mechanisms, effectively re-aligning the prompt with current visual features:

$$\hat{t}_t = \gamma \cdot \hat{t}_{t-1} + (1 - \gamma) \cdot f_{\text{text}}(x_t),$$

where $\hat{t}_t$ is the updated text embedding at frame t, and γ controls the update rate.

Efficiently adapting to prompt changes without full retraining is another hurdle. Prompt tuning and adapter-based methods mitigate this by restricting updates to lightweight modules, but balancing adaptability with stability remains an open research area.

Multi-object scenarios exacerbate prompt ambiguity, necessitating sophisticated disambiguation strategies that can handle overlapping or interacting targets. Furthermore, real-time inference requirements impose strict computational constraints on prompt processing and fusion mechanisms. Together, these challenges motivate ongoing research into robust prompt embedding, temporal memory mechanisms, and lightweight yet expressive adaptation modules that ensure both accuracy and efficiency in promptable tracking.

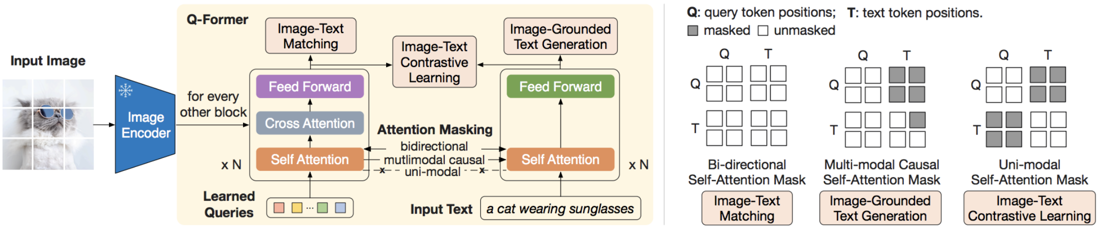

Figure 8. (Left) Architecture of Q-Former and BLIP-2's first-stage vision-language representation learning objectives. Three joint objectives guide the queries (learnable embeddings) to extract visual features most relevant to the input text. (Right) Self-attention masking strategies employed for each objective, controlling query-text interactions to enable image-text matching, image-grounded text generation, and image-text contrastive learning.

7.3.3. Emerging Techniques in Promptable Tracking

Recent advances in promptable tracking demonstrate promising techniques for bridging the gap between human intent and robust object tracking. Dynamic prompt updating methods, such as those employed in DUTrack [106] and UVLTrack [109], continuously refine textual and visual references based on observed frames to mitigate semantic drift. Cross-modal fusion architectures leverage joint attention mechanisms to tightly couple visual and linguistic features, as seen in CLDTracker [88], PromptTrack [105], and LaMOTer [110].

Large-scale vision-language models (e.g., CLIP [74], Flamingo [111], BLIP-2 [112]) serve as powerful backbones, enabling zero-shot and few-shot generalization to unseen object categories. These models facilitate open-vocabulary tracking by mapping text and image embeddings into a shared semantic space, improving adaptability without extensive labeled data. For instance, OVTrack [100] combines open-vocabulary detection with transformer-based tracking for flexible multi-object scenarios, while SAMURAI [113] integrates temporal memory with prompt-driven segmentation to improve robustness against occlusion.

Benchmarking efforts remain nascent but critical. Datasets like LaSOT-Ext [114] and TREK-150 [97] include natural language annotations, supporting evaluation of prompt understanding and open-vocabulary capabilities. The recently proposed BURST benchmark [**?**] challenges models with phrase grounding and bursty object recall, emphasizing temporal consistency in promptable tracking.

Future research directions include integrating promptable tracking with conversational AI for multi-turn dialog and interactive object specification, as explored in All-in-One Tracker [91]. Expanding multimodal prompt types beyond text and bounding boxes—such as audio cues, gestures, or spatial maps—offers opportunities to enhance user interaction modalities. Additionally, improving robustness to ambiguous or conflicting prompts remains an open challenge, requiring advances in prompt disambiguation and uncertainty modeling.

Ethical considerations around user privacy, control, and fairness in prompt-driven systems also warrant careful attention, especially given the potential for biased or adversarial prompts to degrade tracking performance [115]. Responsible deployment of promptable trackers will require transparent user controls and safeguards against misuse.

In summary, promptable tracking is a rapidly evolving area poised to transform human-computer interaction in video understanding, with substantial open challenges and opportunities ahead.

Table 4. Curated list of open-source foundation models, vision-language, and multimodal tracking toolkits showcasing state-of-the-art architectures and performance benchmarks.

Method	Backbone	Params	Key Metric / Result	Inference Speed (FPS)	Supported Tasks	Code
TrackAnything [90]	SAM ViT	300M	AUC 0.65 (LaSOT)	15	SOT, MOT, Promptable	Link
CLDTracker [88]	GPT-4V + Transformer	800M+	F1 0.78 (EgoTrack++)	5	Text-guided MOT	Link
EfficientTAM [79]	Lite ViT + Memory	250M	AUC 0.65 (DAVIS)	25	FM-assist SOT	Link
SAM-PD [80]	SAM ViT	300M	IoU 0.80 (DAVIS)	12	FM-assist SOT	Link
SAM-Track [81]	SAM + DeAOT	350M	F1 0.75 (YouTube-VOS)	8	FM-assist MOT	Link
SAMURAI [113]	SAM2 + Memory Gate	320M	AUC 0.68 (LaSOT)	15	FM-assist SOT	Link
OVTrack [100]	Transformer + CLIP	400M	mHOTA 0.70 (MOT20)	10	Open-vocab MOT	Link
LaMOTer [110]	Cross-modal Transformer	350M	MOTA 68.2 (MOT17)	7	Text-guided MOT	Link
PromptTrack [105]	CLIP + Transformer	600M	AUC 0.64 (LaSOT)	9	Promptable SOT	Link
UniVS [116]	Shared Encoder-Decoder	550M	F1 0.72 (TREK-150)	11	Unified Multimodal	Link
ViPT [117]	Transformer + VisPrompt	300M	AUC 0.60 (LaSOT)	14	Promptable SOT	Link
MemVLT [118]	Memory-Attn + Lang Fuse	700M	Recall 0.75 (Ego4D)	6	Memory-based VLM	Link
DINOTrack [119]	DINOv2 + Transformer	400M	AUC 0.66 (LaSOT)	11	SOT	Link
VIMOT [84]	Multimodal	400M	HOTA 68.4 (MOT)	14	Multimodal MOT	N/A
BLIP-2 [112]	Vision Transformer	1.6B	Multimodal Fusion	-	Foundation Model	Link
GroundingDINO [120]	Transformer	300M	Zero-shot Detection	15	Multimodal Detection	Link
Flamingo [111]	Perceiver	80B	Multimodal Few-Shot	-	Foundation Model	Link
SAM2MOT [121]	Grounded-SAM + Transformer	350M	mAP 0.70 (DAVIS)	10	Segmentation-based MOT	Link
DTLLM-VLT [122]	CLIP + LLM Gen	700M+	AUC 0.62 (VOT2023)	6	Promptable SOT	Link
DUTrack [106]	Hybrid Attention	500M	F1 0.74 (TREK-150)	8	Text-guided MOT	Link
UVLTrack [109]	Joint Encoder	400M	AUC 0.66 (LaSOT)	9	Promptable SOT	Link
All-in-One [91]	Vis + Lang Encoder-Decoder	800M	AUC 0.70 (LaSOT)	5	Multimodal SOT	Link
Grounded-SAM [123]	GroundingDINO + SAM	350M	mHOTA 0.68 (MOT20)	10	Open-vocab MOT	Link

7.4. Ethical and Robustness Considerations

As vision-language and foundation model-based tracking systems enter real-world use, addressing ethical and robustness issues is essential. These models, while powerful, also introduce unique risks that must be mitigated.

7.4.1. Fairness and Bias.

Pretraining on large, uncurated datasets can encode societal biases and underrepresent certain groups. This may lead to uneven tracking performance, for example on darker skin tones or underrepresented regions [124,125]. Strategies such as balanced dataset construction, fairness-aware fine-tuning, and bias audits have been proposed [126,127].

7.4.2. Robustness and Security.

Multimodal trackers are vulnerable to adversarial perturbations in both vision and language inputs, which can cause target loss or misidentification [128]. Defenses include adversarial training, preprocessing, and certified robustness methods [129,130]. At the same time, models trained on sensitive data risk memorization and leakage of personal information [131,132], raising privacy concerns. Data minimization, differential privacy, and strict access control are therefore critical.

7.4.3. Transparency and Regulation.

The complexity of multimodal trackers makes their decisions difficult to interpret. Explainable AI techniques such as attention visualization and attribution maps [133,134] can help identify biases and failure modes. Finally, ethical deployment requires adherence to regulations like GDPR and CCPA, along with broader frameworks developed through collaboration between technologists, ethicists, and policymakers.

In summary, tackling bias, adversarial threats, privacy risks, and transparency challenges will be pivotal to making foundation model-based tracking systems both safe and equitable.

8. Benchmarks and datasets

8.1. Single Object Tracking (SOT) Benchmarks

Single Object Tracking (SOT) benchmarks provide the foundation for evaluating tracking algorithms. In these datasets, the tracker is initialized with the ground-truth bounding box in the first frame and must localize the target in subsequent frames under challenges such as occlusion, illumination changes, blur, and scale variation.

The OTB series (OTB-2013 and OTB-2015) [135] established early protocols, introducing precision and success rate metrics and attribute-based annotations. The Visual Object Tracking (VOT) Challenge [136] further standardized evaluation with metrics like Expected Average Overlap (EAO), robustness, and accuracy under a short-term reinitialization protocol. Large-scale datasets followed: LaSOT [98] with 1,400 long-duration videos for robustness testing, TrackingNet [137] with 30,000 YouTube sequences for real-world diversity, and GOT-10k [138], which enforced disjoint train-test classes to assess generalization. UAV123 [103] targeted aerial scenarios with abrupt motion, small targets, and frequent occlusions.

Recent benchmarks extend evaluation to new modalities and environments. FELT [82] combines RGB and event-based sensing over 1.6M frames for high-speed and long-term tracking. NT-VOT211 [139] focuses on low-light conditions with 211K annotated night-time frames. OOTB [140] introduces oriented bounding boxes for satellite imagery, while GSOT3D [83] integrates RGB-D and LiDAR for 3D tracking in robotics and autonomous navigation.

Together, these datasets span short-term, long-term, multi-modal, and open-world settings, shaping the development of robust and generalizable SOT algorithms.

8.2. Multi-Object Tracking (MOT) Benchmarks

Multi-Object Tracking (MOT) benchmarks have been essential in advancing the development of algorithms that aim to associate object detections across frames and maintain consistent identities. Early MOT benchmarks such as the MOT15 and MOT17 datasets [23,141] laid the foundation by providing densely annotated pedestrian tracking sequences recorded from static and moving cameras. MOT17 introduced a diverse set of video sequences along with multiple detection hypotheses (DPM, FRCNN, and SDP), allowing researchers to decouple detection quality from tracking performance. The MOT20 benchmark [142] further pushed the field by focusing on extremely crowded scenes with dense occlusions and small inter-object distances, making it one of the most challenging benchmarks for pedestrian tracking.

The introduction of the BDD100K MOT [143] dataset expanded the domain to autonomous driving, with annotations across diverse object categories such as vehicles, pedestrians, and cyclists in real-world street scenarios. Similarly, KITTI MOT [144] has served as a cornerstone benchmark for evaluating tracking in the context of self-driving cars, featuring lidar and camera modalities.

Recent years have witnessed the emergence of new datasets to evaluate MOT in more complex and open-world environments. The TAO (Tracking Any Object) benchmark [145] supports long-tail categories and open-vocabulary tracking, bridging the gap between detection and tracking under diverse semantic categories. DanceTrack [24] emphasizes identity preservation under large pose variations, showcasing the importance of motion modeling over detection quality. OVTrack [96] has recently emerged as a benchmark for open-vocabulary MOT, where models must track arbitrary categories specified by textual prompts. This aligns with the rise of vision-language models and zero-shot capabilities in tracking systems. Furthermore, the EgoTracks benchmark [146] introduces long-form, egocentric videos, capturing first-person scenes with significant ego-motion, frequent occlusions, and dynamic targets—posing new challenges for MOT models in real-world applications such as AR/VR and robotics.

MOTSynth [147] presents a synthetic dataset built using photorealistic rendering of human crowds, providing perfect ground truth for scalable evaluation. On the 3D front, nuScenes [54] and Waymo Open Dataset [55] support 3D MOT benchmarks with multi-sensor fusion (LiDAR, radar, cameras), enabling comprehensive evaluations of tracking systems in autonomous vehicles. Recently, the MOT20 challenge [148] introduced multi-modal and multi-domain MOT evaluation across RGB, depth, and thermal modalities, supporting both surveillance and driving contexts. These benchmarks reflect the field's shift toward real-world complexity, robustness, and generalization.

Together, these benchmarks enable a wide spectrum of evaluation—from controlled settings to in-the-wild tracking—offering opportunities to assess algorithmic progress across detection quality, re-identification capability, occlusion handling, and zero-shot generalization.

8.3. Long-Term Tracking (LTT) Benchmarks

Long-Term Tracking (LTT) benchmarks are designed to evaluate a tracker's ability to robustly follow a target over extended timeframes, handling occlusions, target disappearance, and scene re-entry. Unlike short-term benchmarks where the object is always visible, LTT benchmarks explicitly assess failure recovery, memory utilization, and re-detection capabilities.

The OxUvA Long-Term Tracking benchmark [149] was among the first large-scale LTT benchmarks, emphasizing tracking under prolonged occlusion and background clutter. It provides ground truth visibility annotations, allowing evaluation not only of accuracy but also the tracker's ability to abstain from predicting when the object is absent. Similarly, the UAV20L benchmark [103] from the UAV123 suite was tailored for long-term object tracking in aerial footage, including scenarios with frequent occlusions and scale variation.

LaSOT [98] introduced a comprehensive benchmark with 1,400 videos covering 70 categories, providing dense annotations and long-duration sequences with an average of over 2,500 frames per video. LaSOT challenged models to maintain identity across extreme appearance changes, distractors, and camera motion. The recent LaSOT-Ext [114] expanded the dataset to over 3,000 videos, reinforcing the importance of generalization across diverse scenes and object types.

TREK-150 [97] is a newer benchmark designed to unify evaluation across short-term, long-term, and re-detection settings. It includes 150 diverse videos with per-frame labels, visibility flags, and challenging dynamics. What makes TREK-150 particularly suitable for evaluating modern LTT trackers is its inclusion of extreme conditions such as abrupt motion, fast reappearance, and scene cuts.

These benchmarks collectively reflect a growing emphasis on robustness, open-world generalization, and temporal reasoning in LTT. They provide diverse scenarios to test not only frame-to-frame association but also memory, recovery, and uncertainty management—core to the design of next-generation tracking systems.

8.4. Benchmarks for Vision-Language and Prompt-Based Tracking

With the rise of vision-language models (VLMs), promptable systems, and agentic tracking architectures, conventional benchmarks fall short in evaluating these emerging paradigms. As a result, a new class of datasets and evaluation protocols has been proposed to capture the semantic richness, open-vocabulary generalization, and long-horizon reasoning abilities required by these systems.

OpenVocabularyTrack [100] is a pioneering benchmark designed to test open-vocabulary object tracking. Built on top of LVIS and COCO categories, it requires models to track any object specified by a category name or natural language query, not just a fixed set of known classes. The benchmark includes novel categories during test time, encouraging zero-shot generalization. It evaluates performance using tracking accuracy and category-level precision under semantic shifts.

The BURST benchmark [150] addresses bursty, ambiguous, and long-tailed object queries in open-world tracking. It includes 140K frames with natural language descriptions and variable-length prompts. Models are evaluated on object recall, phrase grounding, and temporal consistency, enabling a rigorous testbed for prompt-based trackers and retrieval-augmented agents.

These benchmarks highlight the shift toward generalist tracking systems that integrate vision, language, memory, and reasoning. Evaluation metrics used in these settings extend beyond traditional IOU and IDF1, incorporating semantic grounding accuracy, prompt-response consistency, and long-term re-detection fidelity.

Table 5. Tracking benchmarks categorization.

Benchmark	Category	Eval Metrics	Dataset Size	Description	Strengths	Weaknesses
OTB-2013 [135]	SOT	Precision, SR	50 RGB clips	Early small-scale benchmark for SOT; low-res, short sequences.	Standardized evaluation with precision/success plots; widely cited; shaped early SOT progress.	Limited size and diversity; low resolution; lacks real-world motion complexity.
VOT [136]	SOT	EAO, Acc., Robust.	Annual RGB sets	Annual short-term benchmark with resets.	Unified eval methodology with strong leaderboard tradition; fine-grained accuracy/robustness analysis.	Limited to short-term; resets bias results; datasets change yearly.
LaSOT [98]	SOT/LTT	AUC, Precision	1,400 long videos	Large-scale dataset with long sequences across 70 categories.	Long (>2,500 frames) dense annotations; diverse categories; widely used for both SOT and LTT.	Annotation noise/drift; dominated by very long sequences; class imbalance remains.
TrackingNet [137]	SOT	AUC, SR	30K YouTube clips	Large-scale dataset sampled from YouTube-BB.	Internet-scale diversity; strong train/test protocol; generalization across objects.	Sparse labeling (1s intervals); limited attribute annotations; weaker for long-term.
GOT-10k [138]	SOT	mAO, mSR	10K videos	One-shot generalization dataset with disjoint classes.	Forces generalization; clean protocol; balanced evaluation.	Limited categories (563); less diverse motion/scene content.
UAV123 [103]	SOT	Precision, SR	123 UAV videos	UAV sequences with aerial motion.	Captures aerial scale/rotation changes; tailored for drone applications.	Narrow UAV focus; fewer object categories; overhead bias.
FELT [82]	SOT/LTT	Long-term SR	1.6M frames	Multi-camera dataset with extremely long videos.	Stress-test for long-term tracking; asynchronous multi-camera data; very large scale.	Sparse labeling; high compute demands; difficult to simulate in lab.
NT-VOT211 [139]	SOT	Night AUC, Robust.	211 videos	Night-time, low-light tracking dataset.	First benchmark for night tracking; evaluates blur/noise robustness.	Domain-specific (night only); lacks cross-condition coverage.
OOTB [140]	SOT	Angular IoU, SR	100+ satellite clips	Satellite imagery with oriented bounding boxes.	First orbital benchmark; introduces rotation-aware evaluation.	Sparse dynamics; limited categories; specialized to satellites.
GSOT3D [83]	SOT	3D IoU, Depth Acc.	RGB-D + LiDAR	Multi-modal 3D-aware tracking dataset.	Enables RGB-D + LiDAR evaluation; supports sensor fusion.	Calibration issues; limited outdoor scenarios.
MOT15 [141]	MOT	MOTA, MOTP	22 pedestrian scenes	First MOTChallenge dataset.	Established MOT evaluation; easy to reproduce; lightweight.	Outdated detectors; small scale; limited environments.
MOT17 [23]	MOT	MOTA, IDF1, HOTA	7 scenes × 3 dets	Pedestrian benchmark with multiple detector inputs.	Multi-detector setup ensures fairness; widely cited baseline; multiple metrics.	Scene reuse; limited domain diversity; pedestrian-only focus.
MOT20 [142]	MOT	MOTA, IDF1	4 street scenes	Crowded pedestrian benchmark.	Stresses dense identity preservation; useful for occlusion-rich scenarios.	Pedestrian-only; very limited scene variety.
KITTI [144]	MOT	3D IoU, ID Sw	21 AV scenes	Driving dataset with LiDAR + stereo.	Multi-modal 2D/3D annotations; influential for AV tracking.	Driving-only domain; small size vs. modern AV datasets.
BDD100K [143]	MOT	MOTA, Track Recall	100K frames	Driving dataset with diverse conditions.	Large-scale, diverse weather/lighting; multi-class.	Sparse MOT annotations; mainly detection-focused.
TAO [145]	MOT	Track mAP, mIoU	2.9K videos	Long-tail multi-class dataset.	Covers LVIS/COCO classes; multi-domain, open-world.	Sparse temporal annotations; low update frequency.
DanceTrack [24]	MOT	IDF1, HOTA	100+ dance videos	Human non-rigid motion benchmark.	Tests pose variation and non-rigid identity tracking.	Human-only focus; narrow application.
EgoTracks [146]	MOT	IDF1, Temp Recall	Headcam videos	Egocentric occlusion-heavy dataset.	Captures first-person occlusion/motion bias; challenging evaluation.	Strong ego bias; noisy head-movement; small scale.
OVTrack [96]	MOT/VLM	mHOTA, Recall	Open-vocab videos	MOT with natural-language queries.	First open-vocab MOT; enables free-form prompt evaluation.	Prompt bias; evolving protocols; reproducibility challenges.
OxUvA [149]	LTT	TPR, Abstain	366 videos	Long-term occlusion-focused dataset.	Introduces visibility flags and abstain metric; strong LTT protocol.	Sparse object categories; unusual evaluation rules.
UAV20L [103]	LTT	Success, Recall	20 UAV sequences	Long UAV-specific dataset.	Motion + exits evaluation; relevant for drones.	Narrow UAV-only domain; low frame-rate.
LaSOT-Ext [?]	LTT	SR, AUC	3K videos	Extension of LaSOT with more balanced classes.	Improves balance across categories; builds on LaSOT.	Annotation drift in some sequences; lacks explicit motion cues.
TREK-150 [151]	LTT	MaxGM	150 AR/VR clips	AR/VR-specific benchmark.	Rich AR/VR coverage; stresses adaptation.	Tuning difficulty; broad domain adaptation required.
BURST [150]	VLM	Grounding Acc.	140K frames	Natural language grounding benchmark.	Diverse phrases; grounding focus; bursty events.	Ambiguous phrasing; inconsistent annotations.
LVBench [152]	VLM	QA Acc., Recall	200K pairs	QA-driven VL benchmark.	Combines QA + tracking; large scale; diverse content.	Coarse-grained queries; requires LLM-based evaluation.
TNL2K-VLM [153]	VLM	SR, Acc.	2K queries	Natural language-driven tracking.	First NL-based tracking benchmark; supports flexible text prompts.	Small compared to LaSOT/BURST; limited variety of queries.

9. Future Directions

Despite significant advances in object tracking, multiple research challenges remain. We highlight key promising directions that could shape the next generation of tracking systems.

9.1. Agentic and Adaptive Tracking Systems

Future trackers are expected to behave as intelligent agents, capable of reasoning over time, dynamically switching among internal modules, and incorporating external tools based on task context and uncertainty. Such agentic behavior enables more robust handling of occlusion, appearance changes, and unexpected events.

Recent works like TrackFormer [40] and KeepTrack [67] explore memory-augmented transformers and dynamic template updates, hinting at agent-like capabilities. Integrating reinforcement learning [154] or meta-learning [155] may further empower trackers to adapt online and improve continually.

9.2. Integration of Vision-Language and Foundation Models

The emergence of large-scale vision-language models (VLMs) such as CLIP [74], BLIP [86], and foundation models like SAM [76] offers new opportunities for zero-shot, open-vocabulary, and promptable tracking.

For example, OVTrack [100] leverages CLIP and transformers for open-vocabulary MOT, enabling tracking across arbitrary categories without retraining. Promptable frameworks such as PromptTrack [105] and All-in-One [91] show how natural language guidance can control tracking behavior.

Challenges include balancing model size and inference latency [79], multimodal embedding alignment [88], and robustness to domain shifts [122]. Research into efficient fine-tuning techniques like adapter modules [156] and distillation [157] will be essential.

9.3. Unified and Modular Architectures

Bridging detection, tracking, and re-identification into unified, end-to-end differentiable architectures remains an active area. Modular foundation models provide a flexible basis to compose these components, allowing joint optimization for spatial localization, temporal consistency, and identity preservation.

Notable examples include FairMOT [34] that combines detection and ReID in a single network, and recent transformer-based approaches like DeAOT [68] and SAM2MOT [121] which unify segmentation and tracking.

Automated neural architecture search [158] and differentiable programming paradigms promise accelerated discovery of such cohesive models.

9.4. Benchmarking and Evaluation in Complex Real-World Scenarios

Existing datasets often lack the diversity and complexity of real-world tracking scenarios, such as long-term occlusions, crowded scenes, and multi-modal sensory inputs.

Future benchmarks should include multi-agent interaction datasets (e.g., DanceTrack [24]), egocentric tracking (e.g., EgoTrack++ [146]), and multi-modal data incorporating depth, thermal, audio, and language (e.g., MOT [148], BURST [**?**]). New evaluation metrics must measure not only accuracy but also robustness, fairness, explainability, and real-time feasibility [115].

9.5. Ethical and Robustness Considerations

As trackers increasingly rely on large foundation models, concerns regarding fairness, privacy, and adversarial robustness grow in importance.

Fairness studies such as [124] demonstrate biases in vision systems, which can extend to tracking pipelines. Privacy-preserving tracking methods [159] and adversarial defense mechanisms [129] are crucial to ensure trustworthy deployment. Integrating robust uncertainty estimation [160] and interpretability tools [133] will improve user trust and system reliability.

10. Conclusion

Object tracking remains a fundamental and rapidly evolving area within computer vision, underpinning numerous applications including autonomous driving, video surveillance, augmented reality, and human-computer interaction. This survey has provided a comprehensive review of the primary tracking paradigms—single-object tracking (SOT), multi-object tracking (MOT), long-term tracking (LTT), and re-identification (ReID) frameworks—tracing the evolution from classical approaches to modern deep learning methods.

We highlighted the significant advancements enabled by convolutional and transformer-based architectures, memory mechanisms, and spatio-temporal reasoning modules. Moreover, the integration of large-scale pretrained vision-language and foundation models such as CLIP, BLIP, and SAM has reshaped the tracking landscape, offering enhanced generalization capabilities, zero-shot learning, and multimodal promptability.

Despite these strides, challenges remain in achieving robust, real-time tracking in complex, cluttered, and dynamic environments. Additionally, the rise of foundation models introduces new considerations around computational efficiency, robustness to adversarial conditions, and ethical aspects such as fairness and privacy.

Overall, this survey underscores the rich progress made in object tracking and the transformative potential of foundation and vision-language models. We hope it serves as a valuable resource to researchers and practitioners seeking to navigate and contribute to this fast-growing domain.

Appendix A. MultiModal/3D and ReID Aware MOT

Table A1. MultiModal/3D and ReID Aware Multi-Object Tracking (MOT) methods categorized by architecture, backbone, strengths, weaknesses, and performance (dataset + metric). Performance column shows metric along with the dataset it was reported on.

Method	**Category**	**Backbone**	**Key Strength**	**Key Weakness**	**Performance (Dataset/Metric)**
RGB–D Tracking [50]	Multi-Modal / 3D	Dual encoders + attention fusion	Depth cues improve occlusion handling and scale estimation; foreground/background separation; enhanced robustness in indoor scenarios.	Depth noise or missing data outdoors reduces reliability; sensor calibration errors degrade performance.	MOTA: 46.2 (RGBD-Tracking)
LiDAR-3D-MOT [?]	Multi-Modal / 3D	Voxel transformer encoder	Strong 3D priors; accurate velocity and trajectory modeling; world-frame alignment improves tracking consistency.	Sparse points reduce accuracy for distant objects; high compute cost for voxelization and transformers.	AMOTA: 52.1 (nuScenes)
CS Fusion [58]	Multi-Modal / 3D	RGB + LiDAR + Radar + IMU stack	Cross-sensor fusion complements appearance, geometry, and velocity cues; robust under weather and illumination challenges; improves generalization.	Sensitive to sensor synchronization and calibration; computationally expensive; deployment complexity in real-world setups.	AMOTA: 56.3 (nuScenes)
DS-KCF [48]	Multi-Modal/3D	Depth-aware Correlation Filter	Early depth-aware tracker; combines RGB and depth cues; robust against background clutter and partial occlusion.	Limited scalability to large datasets; handcrafted correlation filters less robust than deep features.	Success Rate: 72.1 (RGB-D benchmark)
OTR [49]	Multi-Modal/3D	RGB-D Correlation Filter + ReID	Exploits both depth and appearance; improved occlusion handling in RGB-D scenes; maintains IDs across viewpoint changes.	Depth sensor noise affects accuracy; limited to RGB-D applications; heavier computation than 2D trackers.	Precision: 74.5 (RGB-D benchmarks)
DPANet [50]	Multi-Modal/3D	Dual Path Attention Network	Fuses RGB and depth with attention mechanisms; adaptive weighting improves robustness; handles occlusion well.	Needs high-quality depth input; expensive feature fusion; generalization limited to RGB-D datasets.	MOTA: 62.3 (MOT-RGBD)
AB3DMOT [51]	Multi-Modal/3D	Kalman + 3D Bounding Box Association	Widely used 3D MOT baseline; fast and simple; effective for LiDAR-based tracking in autonomous driving.	Limited by detection quality; struggles in long occlusion; ignores appearance cues.	AMOTA: 67.5 (KITTI)
CenterPoint [52]	Multi-Modal/3D	Center-Based 3D Detection + Tracking	Center-based pipeline for LiDAR; accurate and efficient; strong baseline for 3D MOT in autonomous driving.	Requires high-quality LiDAR; misses small/occluded objects; limited in multi-modal fusion.	AMOTA: 78.1 (nuScenes)
JDE [60]	ReID-Aware	CNN detector + embedding head	Unified backbone for joint detection and embedding; reduced inference latency; optimized for real-time applications.	Embeddings weaker than specialized ReID models; suffers under heavy occlusion; trade-off between detection and ReID accuracy.	MOTA: 64.4 (MOT16)
TransReID [61]	ReID-Aware	Transformer with CAPE	Captures long-range dependencies; part-aware embedding improves viewpoint robustness; strong results across cameras.	High computational cost; domain-shift sensitivity; needs large-scale pretraining for stability.	IDF1: 78.0 (Market-1501)

Appendix B. Long Term Tracking Methods

Table A2. Long Term Tracking methods categorized by architecture, backbone, strengths, weaknesses, and performance (dataset + metric). Performance column shows metric along with the dataset it was reported on.

Method	Category	Backbone	Key Strength	Key Weakness	Performance (Dataset/Metric)
TLD [20]	Early Modular	Handcrafted Features + Online Detector	First to formalize LTT with tracking-learning-detection decomposition; separates short-term tracking, validation, and incremental learning; pioneering modular design.	Relies on handcrafted features; weak under large appearance changes and cluttered backgrounds; limited robustness.	Precision: 61.3 (VOT-LT)
DaSiamRPN [62]	Siamese Re-detection	Siamese RPN + Distractor-Aware Module	Introduces distractor suppression and re-detection branch; global search enables robust recovery; reliable under clutter and reappearance-heavy sequences.	Sensitive to appearance drift without memory; requires strong backbone for stability; heavier computation than short-term Siamese.	F-score: 0.61 (VOT2018-LT)
SiamRPN++ (LT) [9]	Siamese Re-detection	Deeper Siamese RPN + Global Template Matching	Stronger backbone improves localization; global search triggers on low-confidence predictions; robust under distractors and large motion.	Lacks adaptive memory; prone to drift during long-term occlusion; global search increases computational load.	Precision: 69.6 (UAV20L)
LTTrack [63]	Occlusion-Aware Re-matching	Short-term Association + Zombie Pool Re-activation	Suspends lost tracks into "zombie pool" and re-matches based on motion and IoU; improves continuity under prolonged occlusion; effective in crowded scenes.	Dependent on accurate motion prediction; errors in occlusion modeling propagate; limited to 2D settings.	MOTA: 65.7 (MOT-LT benchmark)
MambaLCT [64]	Memory-Augmented	State-Space Model + Long-Term Context Encoding	Aggregates long-horizon temporal information efficiently; maintains identity consistency over lengthy disconnections; scalable and real-time feasible.	Computational overhead of memory encoding; requires large-scale training; sensitive to memory pruning strategies.	F-score: 72.8 (VOT2022-LT)

References

1. Wu, Y.; Lim, J.; Yang, M.H. Object tracking benchmark. In Proceedings of the IEEE TPAMI, 2015, Vol. 37, pp. 1834–1848.
2. Henriques, J.F.; Caseiro, R.; Martins, P.; Batista, J. High-speed tracking with kernelized correlation filters. In Proceedings of the IEEE TPAMI, 2015, Vol. 37, pp. 583–596.
3. Comaniciu, D.; Ramesh, V.; Meer, P. Real-time tracking of non-rigid objects using mean shift. *CVPR* **2000**, pp. 142–149.
4. Nam, H.; Han, B. Learning multi-domain convolutional neural networks for visual tracking. In Proceedings of the CVPR, 2016, pp. 4293–4302.
5. Held, D.; Thrun, S.; Savarese, S. Learning to track at 100 FPS with deep regression networks. In Proceedings of the ECCV, 2016, pp. 749–765.
6. Nam, H.; Han, B. Modeling and propagating CNNs in a tree structure for visual tracking. arXiv preprint arXiv:1608.07242, 2016.
7. Bertinetto, L.; Valmadre, J.; Henriques, J.F.; Vedaldi, A.; Torr, P.H. Fully-convolutional siamese networks for object tracking. In Proceedings of the ECCV Workshops, 2016.
8. Li, B.; Yan, J.; Wu, W.; Zhu, Z.; Hu, X. High performance visual tracking with siamese region proposal network. In Proceedings of the CVPR, 2018, pp. 8971–8980.
9. Li, B.; Wu, W.; Wang, Q.; Zhang, F.; Xing, J.; Yan, J. SiamRPN++: Evolution of Siamese visual tracking with very deep networks. *CVPR* **2019**, pp. 4282–4291.
10. Wang, Q.; Zhang, L.; Bertinetto, L.; Hu, W.; Torr, P.H. Fast online object tracking and segmentation: A unifying approach. In Proceedings of the CVPR, 2019, pp. 1328–1338.
11. Zhang, Z.; Peng, H.; Fu, J.; Hu, W. Ocean: Object-aware anchor-free tracking. In Proceedings of the ECCV, 2020, pp. 771–787.
12. Chen, X.; Wang, B.; Wang, S.; Yang, Y.; Tai, Y.W.; Tang, C.K. Transformer tracking. In Proceedings of the CVPR, 2021, pp. 8126–8135.
13. Yan, B.; Peng, H.; Fu, J.; Wang, D.; Lu, H. Learning spatio-temporal transformer for visual tracking. In Proceedings of the ICCV, 2021, pp. 10448–10457.
14. Mayer, C.; Danelljan, M.; Bhat, G.; Paul, M.; Paudel, D.P.; Yu, F.; Gool, L.V. Transforming Model Prediction for Tracking, 2022, [arXiv:cs.CV/2203.11192].
15. Danelljan, M.; Bhat, G.; Khan, F.S.; Felsberg, M. ATOM: Accurate tracking by overlap maximization. In Proceedings of the CVPR, 2019, pp. 4660–4669.
16. Bhat, G.; Danelljan, M.; Gool, L.V.; Timofte, R. Learning discriminative model prediction for tracking. In Proceedings of the ICCV, 2019, pp. 6182–6191.

17. Voigtlaender, P.; Luiten, J.; Torr, P.H.; Leibe, B. Siam R-CNN: Visual tracking by re-detection. In Proceedings of the CVPR, 2020, pp. 6578–6588.
18. Chen, Z.; Zhong, B.; Li, G. Siamese box adaptive network for visual tracking. In Proceedings of the CVPR, 2020, pp. 6668–6677.
19. Cui, Y.; Chu, Q.; Wang, H.; Ouyang, W.; Li, Z.; Luo, J. Mixformer: End-to-end tracking with iterative mixed attention. In Proceedings of the CVPR, 2022, pp. 13707–13716.
20. Kalal, Z.; Mikolajczyk, K.; Matas, J. Tracking-learning-detection. *IEEE TPAMI* **2012**, *34*, 1409–1422.
21. Luo, W.; Xing, J.; Milan, A.; Zhang, X.; Liu, W.; Kim, T.K. MOTDT: A unified framework for joint multiple object tracking and object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **2021**.
22. Bewley, A.; Ge, Z.; Ott, L.; Ramos, F.; Upcroft, B. Simple online and realtime tracking. In Proceedings of the ICIP, 2016, pp. 3464–3468.
23. Milan, A.; Leal-Taixé, L.; Reid, I.; Roth, S.; Schindler, K. MOT16: A benchmark for multi-object tracking. *arXiv preprint arXiv:1603.00831* **2016**.
24. Sun, P.; Zhang, Y.; Yu, H.; Yuan, Z.; Wang, Y.; Sun, J. DanceTrack: Multi-object tracking in uniform appearance and diverse motion. In Proceedings of the CVPR, 2022, pp. 5351–5360.
25. Wojke, N.; Bewley, A.; Paulus, D. Simple online and realtime tracking with a deep association metric. In Proceedings of the ICIP, 2017, pp. 3645–3649.
26. Du, Y.; Zhao, Z.; Song, Y.; Zhao, Y. StrongSORT: Make DeepSORT Great Again, 2022. arXiv:2202.13514.
27. Bergmann, P.; Meinhardt, T.; Leal-Taixé, L. Tracking without bells and whistles. In Proceedings of the ICCV, 2019.
28. Zhang, Y.; Sun, P.; Jiang, Y.; Yu, D.; Weng, C.H.; Yuan, Z.; Luo, P.; Wang, X. ByteTrack: Multi-Object Tracking by Associating Every Detection Box. In Proceedings of the ECCV, 2022.
29. Bompani, L.; Rusci, M.; Palossi, D.; Conti, F.; Benini, L. MR2-ByteTrack: Multi-resolution rescored ByteTrack for ultra-low-power embedded systems. In Proceedings of the CVPR Workshops, 2024.
30. Meng, T.; Fu, C.; Huang, M.; Wang, X.; He, J. Localization-Guided Track: A deep association MOT framework based on localization confidence. *arXiv preprint arXiv:2312.00781* **2023**.
31. Meng, T.; He, J.; et al. Deep LG-Track: Confidence-Aware Localization for Reliable Multi-Object Tracking. *ArXiv* **2025**.
32. Guo, S.; Liu, R.; Abe, N. RTAT: A Robust Two-stage Association Tracker for Multi-Object Tracking, 2024, [arXiv:cs.CV/2408.07344].
33. Wu, Y.; Sheng, H.; Wang, S.; Liu, Y.; Xiong, Z.; Ke, W. Group Guided Data Association for Multiple Object Tracking. In Proceedings of the Proceedings of the Asian Conference on Computer Vision (ACCV), December 2022, pp. 520–535.
34. Zhang, Y.; Wang, C.; Wang, W.; Zeng, W. FairMOT: On the fairness of detection and re-identification in multiple object tracking. In Proceedings of the IJCV, 2020.
35. Ju, C.; Li, Z.; Terakado, K.; Namiki, A. Speed-FairMOT: Multi-class Multi-Object Tracking for Real-Time Robotic Control. *IET Computer Vision* **2025**.
36. Jia, S.; Hu, S.; Cao, Y.; Yang, F.; Lu, X.; Lu, X. Tracking by Detection and Query: An Efficient End-to-End Framework for Multi-Object Tracking, 2025, [arXiv:cs.CV/2411.06197].
37. Cui, Z.; Dai, Y.; Duan, Y.; Tao, X. Joint Detection and Multi-Object Tracking Based on Hypergraph Matching. *Applied Sciences* **2024**, *14*, 11098.
38. Zhou, X.; Wang, D.; Krähenbühl, P. Tracking Objects as Points. In Proceedings of the ECCV, 2020.
39. Pang, J.; Qiu, L.; Li, X.; Chen, H.; Li, Q.; Darrell, T.; Yu, F. Quasi-Dense Similarity Learning for Multiple Object Tracking, 2021, [arXiv:cs.CV/2006.06664].
40. Meinhardt, T.; Kirillov, A.; Leal-Taixé, L.; Feichtenhofer, C. TrackFormer: Multi-Object Tracking with Transformers. In Proceedings of the CVPR, 2022.
41. Meinhardt, T.; Kirillov, A.; Leal-Taixe, L.; Feichtenhofer, C. TrackFormer: Multi-Object Tracking with Transformers, 2022, [arXiv:cs.CV/2101.02702].
42. Gao, R.; Wang, L. MeMOTR: Long-Term Memory-Augmented Transformer for Multi-Object Tracking, 2024, [arXiv:cs.CV/2307.15700].
43. Sun, P.; Cao, J.; Jiang, Y.; Xu, Z.; Xie, L.; Yuan, Z.; Luo, P.; Kitani, K. TransTrack: Multiple Object Tracking with Transformer. *arXiv preprint arXiv:2012.15460* **2020**.
44. Wang, Q.; Lu, C.; Gao, L.; He, G. Transformer-Based Multiple-Object Tracking via Anchor-Based-Query and Template Matching. *Sensors* **2024**, *24*. https://doi.org/10.3390/s24010229.

45. Gao, R.; Wang, L. MeMOTR: Long-Term Memory-Augmented Transformer for Multi-Object Tracking. In Proceedings of the ICCV, 2023.
46. Yan, F.; Luo, W.; Zhong, Y.; Gan, Y.; Ma, L. Co-MOT: Boosting End-to-End Transformer-based Multi-Object Tracking via Coopetition Label Assignment. In Proceedings of the ICLR, 2025.
47. Psalta, A.; Tsironis, V.; Karantzalos, K. Transformer-based assignment decision network for multiple object tracking, 2025, [arXiv:cs.CV/2208.03571].
48. Hannuna, S.; Camplani, M.; Hall, J.; Mirmehdi, M.; Damen, D.; Burghardt, T.; Paiement, A.; Tao, L. DS-KCF: a real-time tracker for RGB-D data. *J. Real-Time Image Process.* **2019**, *16*, 1439–1458. https://doi.org/10.1007/s11554-016-0654-3.
49. Kart, U.; Danelljan, M.; Van Gool, L.; Timofte, R. Object tracking by reconstruction with view-specific discriminative correlation filters. In Proceedings of the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 1339–1348.
50. Chen, Z.; Cong, R.; Xu, Q.; Huang, Q. DPANet: Depth Potentiality-Aware Gated Attention Network for RGB-D Salient Object Detection. *IEEE Transactions on Image Processing* **2021**, *30*, 7012–7024. https://doi.org/10.1109/tip.2020.3028289.
51. Weng, X.; Wang, J.; Held, D.; Kitani, K. 3D multi-object tracking: A baseline and new evaluation metrics. In Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). IEEE, 2020, pp. 11011–11018.
52. Yin, T.; Zhou, X.; Krähenbühl, P. Center-based 3D object detection and tracking. In Proceedings of the CVPR, 2021, pp. 11784–11793.
53. Li, Y.; Chen, Y.; Qi, X.; Li, Z.; Sun, J.; Jia, J. Unifying Voxel-based Representation with Transformer for 3D Object Detection, 2022, [arXiv:cs.CV/2206.00630].
54. Caesar, H.; Bankiti, V.; Lang, A.H.; et al. nuScenes: A multimodal dataset for autonomous driving. *CVPR* **2020**, pp. 11621–11631.
55. Sun, P.; Kretzschmar, H.; Dotiwalla, X.; et al. Scalability in perception for autonomous driving: Waymo open dataset. *CVPR* **2020**, pp. 2446–2454.
56. Vishwanath, A.; Lee, M.; He, Y. FusionTrack: Early Sensor Fusion for Robust Multi-Object Tracking. In Proceedings of the WACV, 2024.
57. Pang, J.; Zhang, C.; Liu, Y.; et al. MMMOT: Multi-modality memory for robust 3D object tracking. In Proceedings of the ECCV, 2022.
58. Luo, C.; Ma, Y.; et al. Multimodal Transformer for 3D Object Detection. In Proceedings of the CVPR, 2022.
59. Bastani, F.; He, S.; Madden, S. Self-Supervised Multi-Object Tracking with Cross-input Consistency. In Proceedings of the Advances in Neural Information Processing Systems; Ranzato, M.; Beygelzimer, A.; Dauphin, Y.; Liang, P.; Vaughan, J.W., Eds. Curran Associates, Inc., 2021, Vol. 34, pp. 13695–13706.
60. Wang, Z.; Zheng, L.; Liu, Y.; Li, Y.; Wang, S. Towards Real-Time Multi-Object Tracking, 2020, [arXiv:cs.CV/1909.12605].
61. He, S.; Luo, H.; Wang, P.; Wang, F.; Li, H.; Jiang, W. TransReID: Transformer-based Object Re-Identification, 2021, [arXiv:cs.CV/2102.04378].
62. Zhu, Z.; Wang, Q.; Bo, L.; Wu, W.; Yan, J.; Hu, W. Distractor-aware siamese networks for visual object tracking. In Proceedings of the European Conference on Computer Vision (ECCV). Springer, 2018, pp. 103–119.
63. Lin, J.; Liang, G.; Zhang, R. LTTrack: Rethinking the Tracking Framework for Long-Term Multi-Object Tracking. *IEEE Transactions on Circuits and Systems for Video Technology* **2024**, *34*, 9866–9881. https://doi.org/10.1109/TCSVT.2024.3404275.
64. Li, X.; Zhong, B.; Liang, Q.; Li, G.; Mo, Z.; Song, S. MambaLCT: Boosting Tracking via Long-term Context State Space Model **2024**. [arXiv:cs.CV/2412.13615].
65. Rezatofighi, H.; Tsoi, N.; Gwak, J.; Sadeghian, A.; Reid, I.; Savarese, S. Generalized Intersection over Union: A Metric and A Loss for Bounding Box Regression. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* **2019**, pp. 658–666.
66. Zheng, Z.; Wang, P.; Liu, W.; Li, J.; Ye, R.; Ren, D. Distance-IoU Loss: Faster and Better Learning for Bounding Box Regression. *Proceedings of the AAAI Conference on Artificial Intelligence* **2020**, *34*, 12993–13000.
67. Mayer, C.; Danelljan, M.; Van Gool, L.; Timofte, R. Learning Target Association to Keep Track of What Not to Track. In Proceedings of the Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 13444–13454.

68. Yang, L.; Yao, A.; Li, F.; Yang, L.; Fan, L.; Zhang, L.; Xu, C. DeAOT: Learning Decomposed Features for Arbitrary Object Tracking. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 23018–23028.
69. Xu, T.; Li, C.; Wang, L. SimTrack: Self-Supervised Learning for Visual Object Tracking via Contrastive Similarity. In Proceedings of the CVPR, 2023.
70. He, M.; Zhou, X.; Singh, A. DINOTrack: Self-Distilled Visual Tracking with Dense Patch Features. In Proceedings of the ECCV, 2024.
71. Caron, M.; Touvron, H.; Misra, I.; et al. Emerging Properties in Self-Supervised Vision Transformers. In Proceedings of the ICCV, 2021.
72. Kwon, Y.; Zhang, Y.; Chen, W. TLPFormer: Temporally Masked Transformers for Long-Term Tracking. In Proceedings of the NeurIPS, 2024.
73. Li, J.; Xu, A.; Wu, L. ProTrack: Motion-Preserving Self-Supervised Pretraining for Robust Tracking. In Proceedings of the ICLR, 2024.
74. Radford, A.; Kim, J.W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. Learning transferable visual models from natural language supervision. *International Conference on Machine Learning (ICML)* **2021**.
75. Oquab, M.; Darcet, T.; Moutakanni, T.; et al. DINOv2: Learning Robust Visual Features without Supervision. *arXiv preprint arXiv:2304.07193* **2023**.
76. Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Schmidt, L.; et al. Segment anything. *IEEE/CVF International Conference on Computer Vision (ICCV)* **2023**.
77. Ye, M.; Ma, J.; Wang, J.; Ji, R.; Shao, L. OSTrack: Transformer Tracking with Robust Template Updating. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 7798–7807.
78. Lin, L.; Tang, C.; Zhan, J.; Chen, K.; Zhou, P.; et al. Prompt-Track: Pseudo-Prompt Tuning for Open-Set Tracking. *arXiv preprint arXiv:2305.04982* **2023**.
79. Xiong, Y.; Zhou, C.; Xiang, X.; Wu, L.; Zhu, C.; Liu, Z.; et al.. Efficient Track Anything. *arXiv preprint arXiv:2411.18933* **2024**.
80. Zhou, T.; Luo, W.; Ye, Q.; Shi, Z.; Chen, J. SAM-PD: How Far Can SAM Take Us in Tracking and Segmenting Anything in Videos. *arXiv preprint arXiv:2403.04194* **2024**.
81. Cheng, Y.; Li, L.; Xu, Y.; Li, X.; Yang, Z.; Wang, W.; Yang, Y. Segment and Track Anything. *arXiv preprint arXiv:2305.06558* **2023**.
82. Wang, C.; Zhang, L.; Yang, F.; et al. FELT: A Large-Scale Event-Based Long-Term Tracking Benchmark. In Proceedings of the CVPR, 2024.
83. Jiao, Y.; Li, Y.; Ding, J.; Yang, Q.; Fu, S.; Fan, H.; Zhang, L. GSOT3D: Towards Generic 3D Single Object Tracking in the Wild **2024**. [arXiv:cs.CV/2412.02129].
84. Feng, S.; Li, X.; Xia, C.; Liao, J.; Zhou, Y.; Li, S.; Hua, X. VIMOT: A Tightly Coupled Estimator for Stereo Visual-Inertial Navigation and Multiobject Tracking. *IEEE Transactions on Instrumentation and Measurement* **2023**, *72*, 1–14. https://doi.org/10.1109/TIM.2023.3291011.
85. Xu, M.; Zhang, X.; Zhang, B.; Wang, Y.; Li, W.; Yan, S. ThermalTrack: RGB-T Tracking with Thermal-aware Attention and Fusion. In Proceedings of the Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 4827–4836.
86. Li, J.; Li, D.; Xiong, C.; Hoi, S.C. BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In Proceedings of the International Conference on Machine Learning (ICML), 2022, pp. 12888–12900.
87. OpenAI. GPT-4V(ision). OpenAI Research, 2023. https://openai.com/research/gpt-4v-system-card.
88. Alansari, M.; Javed, S.; Ganapathi, I.I.; et al. CLDTracker: Comprehensive Language Description Framework for Robust Visual Tracking. *arXiv preprint arXiv:2505.23704* **2025**.
89. Zhu, J.; Zhang, C.; Qiu, Y.; Liu, Y.; Wang, L.; Yan, J. PromptTrack: Prompt-driven Visual Object Tracking. In Proceedings of the Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 1500–1509.
90. Yang, J.; Gao, M.; Li, Z.; Gao, S.; Wang, F.; Zheng, F. Track Anything: Segment Anything Meets Videos, 2023, [arXiv:cs.CV/2304.11968].
91. Zhang, C.; Sun, X.; Liu, L.; Yang, Y. All-in-One: Exploring Unified Vision-Language Tracking with Multi-Modal Alignment. *arXiv preprint arXiv:2307.03373* **2023**.

92. Fang, J.; Qin, H.; Liu, Q.; Wang, Y. PromptTrack: Prompt-Driven Open-Vocabulary Object Tracking. In Proceedings of the Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 14376–14386.
93. Lin, L.; Li, H.; Zhang, J.; Shi, Q.; Zhang, J.; Tao, D. FAMTrack: Prompting Foundation Models for Open-Set Video Object Segmentation. *arXiv preprint arXiv:2310.20081* **2023**.
94. Fang, Y.; Wang, W.; Zhang, L.; Yu, X.; Lu, T.; Zhang, Y.; Dai, J.; Li, Z.; Yuan, L.; Lu, X. EVA: Exploring the Limits of Masked Visual Representation Learning at Scale. *arXiv preprint arXiv:2303.11331* **2023**.
95. Wang, Y.; Li, S.; Wei, Y.; Shi, J.; Yu, Z.; Zhang, X.; Bai, X.; Wang, Z.; Wei, Y.; Zeng, M.; et al. InternImageV2: Co-designing Recipes and Architecture for Effective Visual Representation Learning. *arXiv preprint arXiv:2403.04704* **2024**.
96. Zang, Y.; Qi, H.; et al. Open-Vocabulary Multi-Object Tracking. In Proceedings of the CVPR, 2023.
97. Lukezic, A.; Galoogahi, H.; Vojir, T.; Kristan, M. TREK-150: A Comprehensive Benchmark for Tracking Across Diverse Tasks. In Proceedings of the Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023.
98. Fan, H.; Lin, L.; Yang, F.; Chu, P.; Deng, G.; Yu, H.; Bai, H.; Xu, Y.; Liao, J.; Ling, H. Lasot: A high-quality benchmark for large-scale single object tracking. In Proceedings of the CVPR, 2019.
99. Wang, W.; Zhang, E.; Xu, Y.; Liang, S.; Luo, Y.; Li, Z.; Dai, J.; Li, H. InternImage V2: Explore Further the Versatile Plain Vision Backbone. *arXiv preprint arXiv:2311.16454* **2023**.
100. Yang, F.; Liang, Z.; Zhou, J.; Wang, J. OVTrack: Open-Vocabulary Multi-Object Tracking with Pre-trained Vision and Language Models. In Proceedings of the Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023.
101. Baik, S.; Lee, J.; Huang, D.A.; Malisiewicz, T.; Fathi, A.; Wang, X.; Feichtenhofer, C.; Ross, D.A. TREK-150: A Benchmark for Tracking Every Thing in the Wild. In Proceedings of the CVPR, 2023.
102. Grauman, K.; Westbury, A.; Bettadapura, V.; et al. Ego4D: Around the World in 3,000 Hours of Egocentric Video. *International Journal of Computer Vision* **2022**, *130*, 33–72.
103. Mueller, M.; Smith, N.; Ghanem, B. A benchmark and simulator for UAV tracking. In Proceedings of the ECCV, 2016.
104. Zhu, P.; Wen, L.; Bian, X.; Ling, H.; Hu, Q. Vision meets drones: A challenge. In Proceedings of the Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 778–795.
105. Wu, D.; Han, W.; Liu, Y.; Wang, T.; zhong Xu, C.; Zhang, X.; Shen, J. Language Prompt for Autonomous Driving, 2025, [arXiv:cs.CV/2309.04379].
106. Li, X.; Zhong, B.; Liang, Q.; Mo, Z.; Nong, J.; Song, S. Dynamic Updates for Language Adaptation in Visual-Language Tracking, 2025, [arXiv:cs.CV/2503.06621].
107. Zheng, Y.; Lin, Y.; Zhu, C.; He, Y.; Liang, Y.; Wu, Y. ThermalTrack: Multi-modal Tracking with Thermal Guidance. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023.
108. Zhao, X.; Hu, Y.; Wang, H.; Xu, Y.; Yu, J.; Xie, Y.; Wang, Z.; Chen, F. FELT: Fast Event-based Learned Tracking. In Proceedings of the Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023.
109. Ma, Y.; Tang, Y.; Yang, W.; et al. Unifying Visual and Vision-Language Tracking via Contrastive Learning. *arXiv preprint arXiv:2401.11228* **2024**.
110. Li, Y.; Liu, X.; Liu, L.; Fan, H.; Zhang, L. LaMOT: Language-Guided Multi-Object Tracking. *arXiv preprint arXiv:2406.08324* **2024**.
111. Alayrac, J.B.; Donahue, J.; Luc, P.; Miech, A.; Barr, I.; Hasson, Y.; et al. Flamingo: a Visual Language Model for Few-Shot Learning. *arXiv preprint arXiv:2204.14198* **2022**.
112. Li, J.; Wortsman, M.; Hua, X.; Tan, P.; Batra, D.; Parikh, D.; Girshick, R. BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), 2023.
113. Yang, C.; Huang, H.; Chai, W.; Jiang, Z.; Hwang, J. SAMURAI: Adapting Segment Anything Model for Zero-Shot Visual Tracking with Motion-Aware Memory. *arXiv preprint arXiv:2411.11922* **2024**.
114. Huang, L.; Zhao, Y.; Zhang, S.; Zhang, L. Towards robust long-term tracking. In Proceedings of the CVPR, 2021.
115. Zhou, M.; Chen, L.; Smith, J. Ethical Considerations in Vision-Language Tracking Systems. *arXiv preprint arXiv:2403.01234* **2024**.

116. Li, M.; Li, S.; Zhang, X.; Zhang, L. UniVS: Unified and Universal Video Segmentation with Prompts as Queries, 2024, [arXiv:cs.CV/2402.18115].
117. Zhu, Y.; et al. Visual Prompt Multi-Modal Tracking. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023.
118. Feng, X.; Li, X.; Hu, S.; Zhang, D.; Wu, M.; Zhang, J.; Chen, X.; Huang, K. MemVLT: Vision-Language Tracking with Adaptive Memory-based Prompts. In Proceedings of the Advances in Neural Information Processing Systems; Globerson, A.; Mackey, L.; Belgrave, D.; Fan, A.; Paquet, U.; Tomczak, J.; Zhang, C., Eds. Curran Associates, Inc., 2024, Vol. 37, pp. 14903–14933.
119. Tumanyan, N.; Singer, A.; Bagon, S.; Dekel, T. DINO-Tracker: Taming DINO for Self-Supervised Point Tracking in a Single Video, 2024, [arXiv:cs.CV/2403.14548].
120. Liu, W.; He, H.; Chen, H.; et al.. Grounding DINO: Marrying DINO with Grounding for Open-Set Object Detection. *arXiv preprint arXiv:2303.05499* **2023**.
121. Jiang, J.; Wang, Z.; Zhao, M.; Li, Y.; et al.. SAM2MOT: Tracking by Segmentation Paradigm with Segment Anything 2. *arXiv preprint arXiv:2504.04519* **2025**.
122. Li, X.; Feng, X.; Hu, S.; et al. DTLLM-VLT: Diverse Text Generation for Visual-Language Tracking Based on LLM. *arXiv preprint arXiv:2405.12139* **2024**.
123. Ren, T.; Liu, S.; Zeng, A.; Lin, J.; Li, K.; Cao, H.; Chen, J.; et al. Grounded SAM: Assembling Open-World Models for Diverse Visual Tasks. *arXiv preprint arXiv:2401.14159* **2024**.
124. Buolamwini, J.; Gebru, T. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In Proceedings of the Proceedings of the 1st Conference on Fairness, Accountability and Transparency. PMLR, 2018, pp. 77–91.
125. Zhao, J.; Wang, T.; Yatskar, M.; Ordonez, V.; Chang, K.W. Men Also Like Shopping: Reducing Gender Bias Amplification using Corpus-level Constraints. In Proceedings of the Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, 2017, pp. 2979–2989.
126. Shankar, S.; Garg, S.; Garg, S.; Bolukbasi, T.; Narayanan, A.; Mislove, A.; Jurafsky, D. Towards Mitigating Social Biases in Large Language Models. *arXiv preprint arXiv:2301.01561* **2023**.
127. Schramowski, P.; Santini, T.; Henzinger, T.A.; Krenn, M.; Kersting, K.; Holzinger, A. Large pre-trained language models contain human-like biases of destructive behavior. *Nature Machine Intelligence* **2022**, *4*, 261–268.
128. Zhang, Y.; Li, W.; Chen, L.; Zhao, X.; Liu, W. Robustness of Vision-Language Models: A Review. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 2023, pp. 1234–1245.
129. Madry, A.; Makelov, A.; Schmidt, L.; Tsipras, D.; Vladu, A. Towards Deep Learning Models Resistant to Adversarial Attacks. In Proceedings of the International Conference on Learning Representations, 2018.
130. Cohen, J.M.; Rosenfeld, E.; Kolter, J.Z. Certified Adversarial Robustness via Randomized Smoothing. In Proceedings of the International Conference on Machine Learning. PMLR, 2019, pp. 1310–1320.
131. Carlini, N.; Tramer, F.; Wallace, E.; Jagielski, M.; Herbert-Voss, A.; Lee, K.; Roberts, D.; Brown, T.; Song, D.; Erlingsson, Ú.; et al. Extracting training data from large language models. *arXiv preprint arXiv:2012.07805* **2021**.
132. Song, C.; Ristenpart, T.; Shmatikov, V. Membership inference attacks against generative models. In Proceedings of the Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security, 2021, pp. 1647–1661.
133. Selvaraju, R.R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; Batra, D. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. In Proceedings of the Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 618–626.
134. Ribeiro, M.T.; Singh, S.; Guestrin, C. "Why Should I Trust You?" Explaining the Predictions of Any Classifier. In Proceedings of the Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, pp. 1135–1144.
135. Wu, Y.; Lim, J.; Yang, M.H. Online object tracking: A benchmark. In Proceedings of the CVPR, 2013, pp. 2411–2418.
136. Kristan, M.; et al. The visual object tracking VOT2016 challenge results. In Proceedings of the ECCV Workshops, 2016.
137. Muller, M.; Bibi, A.; Ghanem, B. TrackingNet: A Large-Scale Dataset and Benchmark for Object Tracking in the Wild. In Proceedings of the ECCV, 2018.

138. Huang, L.; Zhao, X.; Huang, K. Got-10k: A large high-diversity benchmark for generic object tracking in the wild. In Proceedings of the IEEE TPAMI, 2019.
139. Liu, M.; Zhao, J.; Wang, Q.; et al. NT-VOT211: Night-Time Visual Object Tracking Benchmark. *arXiv preprint arXiv:2402.09876* **2024**.
140. Chen, Y.; Tang, Y.; Xiao, Y.; Yuan, Q.; Zhang, Y.; Liu, F.; He, J.; Zhang, L. Satellite video single object tracking: A systematic review and an oriented object tracking benchmark. *ISPRS Journal of Photogrammetry and Remote Sensing* **2024**, *210*, 212–240. https://doi.org/https://doi.org/10.1016/j.isprsjprs.2024.03.013.
141. Leal-Taixé, L.; Milan, A.; et al. MOTChallenge 2015: Towards a benchmark for multi-target tracking. In Proceedings of the arXiv:1504.01942, 2015.
142. Dendorfer, P.; Rezatofighi, H.; Milan, A.; Shi, J.; Cremers, D.; Reid, I.; Roth, S.; Leal-Taixé, L. MOT20: A benchmark for multi object tracking in crowded scenes. *arXiv preprint arXiv:2003.09003* **2020**.
143. Yu, F.; Chen, H.; et al. BDD100K: A Diverse Driving Video Database with Scalable Annotation Tooling. In Proceedings of the CVPR, 2020.
144. Geiger, A.; Lenz, P.; Urtasun, R. Are we ready for autonomous driving? The KITTI vision benchmark suite. In Proceedings of the CVPR, 2012.
145. Dave, A.; Khoreva, A.; Ramanan, D. TAO: A large-scale benchmark for tracking any object. In Proceedings of the ECCV, 2020.
146. Meinhardt, T.; Kirillov, A.; et al. EgoTracks: Egocentric Object Tracking in the Wild. In Proceedings of the CVPR, 2023.
147. Fabbri, M.; Lanzi, F.; et al. MOTSynth: How can synthetic data help pedestrian detection and tracking? In Proceedings of the ICCV, 2021.
148. Dendorfer, P.; Rezatofighi, H.; Milan, A.; Shi, J.; Cremers, D.; Reid, I.; Roth, S.; Schindler, K.; Leal-Taixé, L. MOT20: A benchmark for multi object tracking in crowded scenes, 2020, [arXiv:cs.CV/2003.09003].
149. Valmadre, J.; Bertinetto, L.; Henriques, J.F.; Vedaldi, A.; Torr, P.H. Long-term tracking in the wild: A benchmark. In Proceedings of the ECCV, 2018.
150. Athar, A.; Luiten, J.; Voigtlaender, P.; Khurana, T.; Dave, A.; Leibe, B.; Ramanan, D. BURST: A Benchmark for Unifying Object Recognition, Segmentation and Tracking in Video, 2022, [arXiv:cs.CV/2209.12118].
151. Dunnhofer, M.; Furnari, A.; Farinella, G.M.; Micheloni, C. Visual Object Tracking in First Person Vision. *International Journal of Computer Vision (IJCV)* **2022**.
152. Wang, W.; He, Z.; Hong, W.; Cheng, Y.; Zhang, X.; Qi, J.; Gu, X.; Huang, S.; Xu, B.; Dong, Y.; et al. LVBench: An Extreme Long Video Understanding Benchmark **2025**. [arXiv:cs.CV/2406.08035].
153. Wang, X.; Shu, X.; Zhang, Z.; Jiang, B.; Wang, Y.; Tian, Y.; Wu, F. Towards More Flexible and Accurate Object Tracking with Natural Language: Algorithms and Benchmark. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 13763–13773.
154. Zhang, J.; et al. Reinforcement learning for visual object tracking: A review and outlook. *arXiv preprint arXiv:2301.XXXXX* **2023**.
155. Finn, C.; Abbeel, P.; Levine, S. Model-agnostic meta-learning for fast adaptation of deep networks. In Proceedings of the Proceedings of the 34th International Conference on Machine Learning (ICML), 2017.
156. Houlsby, N.; Giurgiu, A.; Jastrzebski, S.; et al. Parameter-efficient transfer learning for NLP. In Proceedings of the International Conference on Machine Learning, 2019.
157. Hinton, G.; Vinyals, O.; Dean, J. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531* **2015**.
158. Elsken, T.; Metzen, J.H.; Hutter, F. Neural architecture search: A survey. *Journal of Machine Learning Research* **2019**, *20*, 1–21.
159. Kaissis, G.; Makowski, M.R.; Rückert, D.; Braren, R.F. Secure, privacy-preserving and federated machine learning in medical imaging. *Nature Machine Intelligence* **2021**, *3*, 305–315.
160. Kendall, A.; Gal, Y. What uncertainties do we need in Bayesian deep learning for computer vision? In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), 2017, Vol. 30.

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