Deep Learning-Based Object Tracking for Augmented Reality: A System-Level Survey of Methods, Constraints, and Challenges

Duoduo Mou[1]*

[1] Faculty of Computing, Universiti Teknologi Malaysia, Skudai, Johor 81310, Malaysia
* Correspondence: mouduoduo@graduate.utm.my

https://doi.org/10.53104/j.acad.res.adv.2026.03001

Abstract: The immersiveness and usability of augmented reality (AR) systems rely on accurate, temporally stable, and computationally efficient object tracking. Recent advances in deep learning have reshaped visual tracking and enabled increasingly complex AR applications on mobile and edge platforms. As AR progresses toward large-scale consumer and industrial deployment, tracking has become a system-critical perception module that directly affects visual stability, interaction latency, and user trust. This survey reviews deep learning based object tracking for AR from 2018 to 2025, focusing on algorithmic paradigms and system-level constraints. We analyze AR-specific requirements such as tight latency budgets, limited energy, long-term operation, and perceptual stability, and examine four representative paradigms (Siamese networks, deep discriminative correlation filters, Transformer based models, and long-term frameworks) with their design rationales and deployment challenges. We further discuss lightweight architectures, state-space temporal models, and diffusion-based approaches, along with integration strategies involving efficiency optimization, hardware-aware design, 6-DoF pose tracking, SLAM coupling, neural scene representations, and multimodal fusion. Representative datasets and evaluation protocols are analyzed from an AR deployment viewpoint, and open challenges and future research directions are identified. We argue that AR-oriented tracking constitutes a distinct research domain where algorithmic accuracy, perceptual stability, and system efficiency must be jointly optimized to support trustworthy and immersive next-generation AR experiences.

Keywords: augmented reality; object tracking; deep learning; transformer; lightweight models; multimodal fusion; 6-dof pose estimation

Received: 30 January 2026

Revised: 12 March 2026

Accepted: 16 March 2026

Published: 23 April 2026

Citation: Mou, D. (2026) 'Deep Learning-Based Object Tracking for Augmented Reality: A System-Level Survey of Methods, Constraints, and Challenges', *Journal of Academic Research and Advances*, 2(1), pp. 1-14.

1. Introduction

Augmented reality (AR) enhances human perception by superimposing digital content onto the physical world, enabling transformative applications in human–computer interaction, industrial inspection and maintenance, medical intervention, education, and consumer entertainment. A compelling AR experience fundamentally relies on the seamless spatial alignment between virtual content and the physical environment, which in turn requires continuous, accurate, and temporally stable perception. Among the various perception components in an AR pipeline, object tracking plays a central role by continuously estimating the position, scale, and, in many cases, the full six-degree-of-freedom (6-DoF) pose of target objects across time (Danelljan et al., 2020).

Accurate tracking ensures that virtual objects remain correctly anchored to their physical counterparts during user interaction, while even minor localization jitter or drift can immediately degrade perceptual quality and induce discomfort. With the rapid evolution of mobile computing platforms such as smartphones, tablets, and head mounted AR displays and major advances in deep learning, object tracking has progressed from handcrafted feature based methods to data driven approaches with substantially improved robustness and accuracy (Dai et al., 2020). Modern trackers exhibit strong resilience to challenging factors such as illumination variation, motion blur, occlusion, background clutter, and nonrigid deformation.

Accordingly, this survey aims to reposition object tracking within the broader context of AR system design. Rather than treating tracking as an isolated vision problem, we adopt a system-oriented perspective that explicitly accounts for perceptual stability, latency, energy efficiency, and long-term deployment constraints. By synthesizing recent advances across algorithms, system optimization, and evaluation methodologies, this survey seeks to provide both a structured reference for researchers and a practical guide for practitioners developing next-generation AR systems.

1.1 Contributions of this Survey

This survey differs from existing reviews in several important aspects. First, we explicitly frame AR-oriented object tracking as a system-level perception problem rather than an isolated vision task. Second, we provide a unified analysis of major deep learning tracking paradigms through the lens of AR-specific constraints, systematically comparing their algorithmic characteristics, temporal stability, and deployment efficiency on mobile and edge platforms. Third, we emphasize recent developments particularly relevant to AR deployment, including lightweight Transformer architectures, state-space temporal modeling, multimodal perception fusion, and tight integration with SLAM and scene understanding. Finally, we identify open challenges and future research directions that must be addressed to bridge the gap between benchmark-level tracking performance and reliable, long-term AR deployment.

2. Object Tracking in AR: Definitions, Challenges, and Evaluation

Object tracking in augmented reality (AR) differs fundamentally from generic visual tracking tasks studied in traditional computer vision. Rather than serving as an isolated perception module, tracking in AR operates within a tightly coupled, closed-loop system that integrates sensing, rendering, interaction, and user perception. As a result, the design objectives, evaluation criteria, and failure modes of AR-oriented tracking are significantly reshaped by system-level and perceptual constraints. This section formalizes the task definition of object tracking in AR systems, analyzes the unique challenges imposed by AR deployment scenarios, and reviews evaluation criteria that more accurately reflect real-world AR performance.

2.1 Task Definition and System-Level Integration

In AR systems, object tracking aims to continuously estimate the spatial state of a target object across time, typically represented as a two-dimensional bounding box in image coordinates or a full six-

degree-of-freedom (6-DoF) pose in camera or world coordinates (Labbé et al., 2020). Unlike offline tracking benchmarks, AR tracking outputs are consumed immediately by downstream modules, including rendering engines, physics simulators, and interaction logic, forming a closed perception-action loop.

A typical AR tracking pipeline consists of several tightly coupled stages. Initialization may be triggered by automatic object detection, user interaction, or prior scene knowledge. Once initialized, the tracker extracts visual features to model the target's appearance and estimates motion using frame-to-frame correspondence, temporal filtering, or inertial priors (Hodan et al., 2020). Online adaptation mechanisms update the appearance model to accommodate gradual changes, while failure recovery strategies handle occlusion, out-of-view events, or tracking loss.

Crucially, these stages are not independent. Tracking estimates influence rendering and interaction, while rendering feedback and user motion affect subsequent observations. This bidirectional coupling means that tracking errors propagate directly to perceptual artifacts, such as visual jitter, spatial drift, or delayed response, which are immediately noticeable to users (Chen et al., 2021). Consequently, AR tracking must be designed and evaluated as a system-level capability rather than a standalone algorithm.

2.2 Unique Challenges in AR Scenarios

AR applications impose stringent real time constraints that exceed generic tracking benchmarks, requiring 60–90 Hz tracking for visual comfort and end to end latency below 20 ms (Wang et al., 2021). Mobile, battery powered platforms further impose tight energy budgets, shifting the design objective from pure accuracy to a balance among accuracy, latency, and energy efficiency. Perceptual stability is also critical; small frame to frame fluctuations can produce visible jitter and motion sickness, motivating temporally smooth trajectories with low velocity and acceleration variance (Yan et al., 2021). AR sessions often last minutes to hours, demanding long term adaptation, memory management, and reliable reinitialization without drift under appearance variation, deformation, occlusion, and field of view interruptions (Yu et al., 2021). Many applications involve arbitrary user defined objects, creating an open world setting that favors category agnostic and few shot approaches over category specific models (Mayer et al., 2021). Finally, AR environments are highly interactive, with frequent occlusions and visually similar distractors that challenge appearance modeling and motion prediction, particularly in cluttered scenes (Wen et al., 2021).

2.3 Evaluation Metrics for AR-Oriented Tracking

Standard tracking benchmarks typically emphasize metrics such as center location error, intersection-over-union (IoU), success plots, and precision curves. While informative, these metrics fail to capture several aspects that are critical to AR deployment (Zhao et al., 2021). For applications involving 6-DoF tracking, pose accuracy metrics such as Average Distance of Model Points (ADD) and ADD(-S) are widely used. However, in AR contexts, spatial consistency across time is often more important than instantaneous pose error, particularly for maintaining stable virtual overlays (Mayer et al., 2022). Temporal stability can be quantified using trajectory-based metrics, including velocity variance, acceleration variance, or jerk. These measures provide a more direct proxy for perceptual smoothness than frame-wise accuracy and are essential for evaluating AR tracking quality (Cui et al., 2022). The time required to achieve stable tracking after initialization or re-detection is critical for user experience. Prolonged initialization delays or unstable recovery behavior can significantly disrupt interaction (Gao et al., 2022). Energy usage measured on target hardware platforms is an increasingly important evaluation criterion. Sustained high power consumption may lead to thermal throttling, degraded performance, or reduced battery life, all of which negatively impact AR usability.

2.4 Why Object Tracking for AR Is Fundamentally Different from Generic Tracking

Although AR tracking builds upon advances in generic visual tracking, several fundamental differences distinguish it as a separate research problem (Chen et al., 2022). First, AR tracking operates in a closed-loop system where tracking errors directly influence rendering and user interaction. In contrast, generic tracking benchmarks are open-loop and offline, with no perceptual feedback. Second, AR prioritizes perceptual stability over benchmark accuracy. A tracker that achieves high IoU scores but exhibits temporal jitter may perform poorly in AR, while a slightly less accurate but smoother tracker may deliver superior user experience. Third, AR deployment emphasizes end-to-end system constraints rather than algorithmic efficiency in isolation. Latency, power consumption, and hardware compatibility are first-order design objectives, not secondary considerations. Finally, AR tracking must generalize to open-world scenarios and arbitrary objects, often under prolonged and interactive use. These requirements fundamentally reshape both algorithm design and evaluation methodology (Paul et al., 2022). Together, these distinctions motivate treating AR-oriented object tracking as a distinct research domain, in which algorithmic accuracy, temporal stability, and system-level efficiency must be jointly optimized.

3. Deep Learning Tracking Paradigms and Their Suitability for AR

Deep learning has fundamentally transformed visual object tracking by enabling data-driven representation learning and robust modeling under complex real-world conditions. Over the past decade, several major tracking paradigms have emerged, each reflecting different design philosophies regarding efficiency, adaptability, temporal modeling, and robustness (Chen et al., 2023). When deployed in augmented reality (AR) systems, however, the performance of these paradigms must be evaluated not only in terms of tracking accuracy on standard benchmarks, but also with respect to latency, energy consumption, perceptual stability, and long-term reliability.

In this section, we systematically review four major deep learning based tracking paradigms: Siamese networks, deep discriminative correlation filters, Transformer based architectures, and long term tracking frameworks, and analyze their respective strengths and limitations from an AR oriented, system level perspective.

Table 1. Comparison of Major Deep Learning Tracking Paradigms under AR-Specific Constraints

Tracking Paradigm	Representative Methods	Latency & Throughput	Energy Efficiency	Temporal Stability	Long-Term Robustness	Adaptability	AR Deployment Suitability
Siamese Networks	SiamFC; SiamRPN(++); SiamMask; Ocean; LightTrack	Very low latency; high FPS; predictable inference	High, especially with lightweight backbones	Moderate; may suffer from drift and jitter	Limited without explicit re-detection	Low; primarily offline trained	Well-suited for short-term, interaction-intensive AR scenarios where responsiveness is critical
Deep DCF-based Trackers	ECO; ATOM; DiMP;	Moderate latency due	Medium;	High if updates are	Strong under gradual	High via online learning	Suitable for long-duration AR sessions,

	PrDiMP	to online updates	update cost is non-negligible	well-regularized	appearance change		but requires stability-aware update control
Transformer-based Trackers	TransT; STARK; OSTrack; MixFormer; SimTrack	Historically high, now moderate with lightweight designs	Medium to low; improving with efficient attention	High; global context modeling reduces jitter	Strong under occlusion and clutter	Medium; mostly offline with limited adaptation	Promising for complex AR scenes, requires hardware-aware optimization
Long-Term Tracking Frameworks	LTMU; SPLT; GlobalTrack variants	Variable; re-detection introduces latency spikes	Medium to low	Variable; depends on memory and update strategy	Very high; handles target disappearance	Medium to high	Essential for persistent AR anchoring, should be tightly integrated with SLAM
Emerging Architectures	State-space models; diffusion-based trackers	Uncertain; mostly prototype-level	Unclear	Potentially high	Potentially high	Medium	Research-stage; real-time AR applicability remains open

3.1 Siamese Network Paradigm: Efficiency-Oriented Tracking

The Siamese network paradigm represents one of the most influential and widely adopted approaches in modern object tracking. Early Siamese trackers reformulated tracking as a similarity learning problem between a target template extracted from an initial frame and a candidate search region in subsequent frames (Wen et al., 2023a), enabling efficient feed forward inference without expensive online updates. SiamFC established the foundational framework by showing that a fully convolutional Siamese network could achieve real time performance with competitive accuracy, while methods such as SiamRPN and SiamRPN++ integrated region proposal networks to improve scale and aspect ratio estimation, enhancing robustness under scale variation and fast motion common in interactive AR scenarios (Kristan et al., 2023). Recent variants explored anchor free formulations, lightweight backbones, and mobile friendly architectures (e.g., Ocean, LightTrack) to reduce computational overhead and facilitate deployment on resource constrained AR platforms such as smartphones and head mounted displays. From an AR oriented perspective, Siamese trackers benefit from predictable low latency and the absence of online learning, which simplifies system integration and avoids unstable updates (Chen et al., 2023). However, reliance on offline trained representations limits their adaptability to illumination changes, deformation, and prolonged occlusion in long duration AR sessions, and the lack of explicit recovery mechanisms hinders re acquisition after target disappearance. Consequently, Siamese trackers are well suited for short term, interaction intensive scenarios where responsiveness and stability outweigh long term adaptability, including object manipulation, gesture based interaction, and rapid scene exploration. In practice, they are often paired with higher level system components such as periodic reinitialization, detector based recovery, or SLAM constraints to mitigate drift and improve operational robustness.

3.2 Deep Discriminative Correlation Filters: Online Adaptation and Stability

Deep discriminative correlation filter (DCF) based trackers represent a complementary design philosophy that emphasizes online adaptation and discrimination between target and background. Classical DCF methods exploited frequency domain optimization, while modern deep DCF trackers integrate convolutional neural network features to improve representational power. ECO introduced efficient convolution operators and factorized convolution to enable deep feature usage without prohibitive cost, ATOM decoupled target classification from bounding box estimation for independent optimization (Wen et al., 2024), and DiMP and its variants further learned the optimization process itself, improving robustness and adaptability. Recent extensions incorporating memory mechanisms and distractor modeling allow DCF based trackers to maintain discriminative power in cluttered environments (Xie et al., 2024), a capability particularly relevant to AR scenarios involving interactive occlusions and visually similar backgrounds. The primary advantage of DCF based trackers lies in their strong adaptability; by continuously updating the appearance model during inference, they accommodate gradual changes in appearance, lighting, and context (Hong et al., 2024), making them suitable for long duration AR applications. However, online adaptation introduces challenges, as noisy or occluded observations can corrupt the model, leading to drift or failure, and the computational cost of continual optimization increases latency and energy consumption, conflicting with AR's strict real time constraints. In AR deployment, DCF based trackers are thus well suited for persistent object interaction scenarios, such as industrial inspection or maintenance, but perceptual stability demands careful regulation of update strategies through confidence estimation, temporal smoothing, or conservative scheduling to prevent perceptual jitter that is immediately noticeable to users.

3.3 Transformer-Based Tracking: Global Context and Robust Interaction

Transformer-based architectures have recently emerged as a strong alternative to convolution-centric tracking approaches. By leveraging self-attention and cross-attention, they explicitly model global dependencies between the target and its surrounding context. Early works such as TransT and STARK demonstrated the effectiveness of attention mechanisms under occlusion and background clutter (Wang et al., 2024), while later methods including OSTrack, MixFormer, and SimTrack adopted one-stream designs that unify feature extraction and matching within a single Transformer backbone for improved efficiency. Recent studies further explored lightweight and hybrid variants that combine convolutional layers with efficient attention mechanisms, token pruning, and reduced-resolution representations, lowering memory access and computational cost for mobile and edge AR platforms. Transformer-based trackers offer advantages for AR deployment, including improved robustness under partial occlusion and clutter, as well as more stable localization through broader spatial cues. However, they remain more resource-intensive than Siamese or DCF-based approaches (Ravi, et al., 2024), and attention operations may exceed latency or power budgets on mobile AR devices. In practice, Transformer-based trackers are most effective in visually complex environments that demand contextual reasoning, and real-world deployment benefits from hardware-aware attention design, selective token processing, and hybrid CNN–Transformer pipelines. With such optimizations, Transformers represent a promising direction for next-generation AR tracking.

3.4 Long-Term Tracking and Occlusion-Aware Architectures

Long-term tracking frameworks address scenarios in which targets may undergo extended occlusion, leave the field of view, or reappear after prolonged absence. These methods typically integrate short-term local tracking with global re-detection mechanisms and explicit memory components. Representative approaches combine a fast local tracker with a detector or retrieval module that

searches for the target at a global scale. Memory banks storing historical target representations are often employed to maintain identity consistency over time. Long-term tracking capabilities are essential for persistent AR content anchoring, where virtual objects must remain associated with physical counterparts across interruptions. However, global re-detection and memory management introduce additional latency and complexity, which may conflict with real-time AR constraints. In AR systems, long-term tracking frameworks are most effective when tightly integrated with SLAM and spatial mapping components. Spatial constraints provided by SLAM can significantly reduce the search space for re-detection, improving efficiency and stability. Conservative memory update strategies are also critical to avoid perceptual artifacts during re-acquisition.

3.5 Emerging Architectures and AR-Oriented Design Trends (2024–2025)

Recent studies have explored alternative temporal modeling paradigms, including state space models with linear computational complexity and diffusion based approaches for robust state refinement under uncertainty (Wu et al., 2024). While these architectures show promise in offline evaluation, their real time applicability to AR remains uncertain due to computational demands and integration challenges. Across tracking paradigms, no single architecture satisfies all AR requirements: Siamese trackers excel in efficiency and responsiveness, DCF based methods offer strong adaptability, Transformer based designs provide contextual reasoning, and long term frameworks enable persistent tracking. As a result, AR oriented systems increasingly adopt hybrid approaches and negotiate trade offs among latency, perceptual stability, and robustness (Li et al., 2024).

3.6 Representative Methods and Paradigm-Level Discussion

While Sections 3.1 to 3.4 categorize tracking methods by architectural paradigms, it is also important to highlight representative design trends and recurring principles that have shaped recent progress. Across paradigms, feature representation choices play a decisive role in balancing robustness and efficiency: early Siamese trackers adopted shallow convolutional features for real time operation, later designs incorporated deeper backbones and multilevel fusion to improve discrimination, and Transformer based methods extended this trend by modeling global context through attention. In AR systems, however, maintaining a single static template is often insufficient due to long term appearance variation, motivating controlled template update strategies such as storing a small set of historical templates or selectively updating based on confidence estimates, underscoring the importance of stability aware representation management (Liang et al., 2025). Another recurring theme is temporal reasoning beyond frame to frame matching; recent methods introduce recurrent models, temporal attention, or state space formulations, which are particularly useful for AR scenarios with rapid motion or intermittent occlusion, though real time constraints favor lightweight filters or predictive motion models over heavy recurrent architectures (Cai et al., 2025). Finally, robust failure handling remains a defining challenge, with long term frameworks integrating confidence estimation and redetection modules for recovery. In AR systems, failure detection is especially critical because incorrect tracking produces visible artifacts rather than silent errors, making conservative failure detection and graceful degradation, such as temporarily freezing virtual content or switching to lower confidence rendering modes, key design principles for AR oriented deployment.

3.7 Paradigm-Level Trade-offs and Design Considerations for AR

Although deep learning based object tracking methods can be broadly categorized into distinct paradigms, their practical performance in AR systems is ultimately governed by a set of recurring trade offs (Guo et al., 2025). A primary trade off lies between computational efficiency and

representational richness. Siamese based trackers prioritize streamlined architectures and predictable inference latency, making them attractive for latency sensitive AR applications, but their limited online adaptability can hinder robustness under long term appearance variation. In contrast, Transformer based trackers offer rich contextual reasoning and improved robustness in complex scenes, yet require careful architectural optimization to satisfy mobile deployment constraints (Liu et al., 2025). Another key trade off concerns adaptability versus stability. Trackers with aggressive online update mechanisms, such as deep DCF based approaches, can rapidly adapt to evolving appearance and background conditions, but in AR systems such adaptability must be regulated to avoid perceptual instability (Sun et al., 2025), motivating conservative update strategies, confidence aware learning, and bounded memory mechanisms. Long term tracking frameworks introduce an additional trade off between robustness and system complexity: global redetection and memory mechanisms enable persistent object anchoring across interruptions, but increase latency variance and integration complexity, costs that must be weighed against persistence benefits in interactive or safety critical AR applications (Videnovic, et al., 2025). Collectively, these trade offs highlight that AR oriented tracking design is inherently application dependent, and that effective systems tailor tracking architectures and update strategies to the specific latency, stability, and robustness requirements of the target application.

Table 2. Representative Deep Learning-Based Trackers for AR-Oriented Deployment

Method	Year	Paradigm	Key Design Characteristics	Strengths for AR	Limitations
SiamFC	2016	Siamese	Fully convolutional similarity matching	Ultra-low latency; predictable inference	Limited adaptability; weak long-term robustness
SiamRPN++	2019	Siamese	Deep backbone with region proposal	Accurate localization with high speed	Increased computation; offline-only adaptation
DiMP	2019	DCF-based	Learned discriminative model update	Strong adaptability to appearance change	Online updates may cause instability
ATOM	2019	DCF-based	Target classification and IoU prediction	Stable tracking under gradual variation	Moderate latency and energy cost
OSTrack	2022	Transformer	One-stream Transformer architecture	Robust under occlusion and clutter	Higher memory and compute demand
MixFormer	2022	Transformer	Iterative mixed attention mechanism	Strong global context modeling	Less suitable for low-power devices
FoundationPose	2024	6-DoF Tracking	Unified pose estimation and tracking	Accurate spatial anchoring for AR	Complex pipeline; heavy computation

4. AR-Oriented Optimization and System Integration

While advances in tracking algorithms form the foundation of AR perception, practical deployment hinges on system-level optimization and tight integration with other AR subsystems. Unlike offline tracking benchmarks, AR systems operate under strict real-time, energy, and stability constraints, requiring holistic co-design across algorithms, hardware, and software.

4.1 Extreme Model Efficiency Optimization

Achieving real-time performance on mobile and edge AR platforms necessitates aggressive efficiency optimization beyond conventional FLOP reduction. In practice, end-to-end latency, memory access patterns, and power consumption are often more informative objectives than theoretical computational complexity. Hardware-aware neural architecture search has emerged as an effective strategy for tailoring tracking models to specific AR devices (Shim et al., 2025). By incorporating latency and energy measurements into the optimization objective, such approaches can produce architectures that outperform manually designed networks under real deployment conditions. Dynamic inference techniques, including adaptive resolution processing, early exiting, and confidence-aware computation, further enable trackers to allocate resources selectively based on scene complexity. Platform-specific optimization is equally critical. Efficient utilization of mobile GPUs, NPUs, and heterogeneous computing units can substantially reduce latency and energy consumption. For AR systems, these optimizations must be evaluated holistically, as improvements in one subsystem may introduce bottlenecks or synchronization issues elsewhere in the pipeline (Alansari et al., 2025).

4.2 End-to-End 6-DoF Object Pose Tracking

Many AR applications require full 6-DoF object pose tracking rather than two-dimensional localization. Recent work has explored end-to-end learning approaches that directly regress object pose from image sequences, as well as hybrid methods that combine explicit geometric models with learned feature representations (Kristan et al., 2024). End-to-end approaches offer conceptual simplicity and robustness to appearance variation, but often incur high computational cost (Wen et al., 2023b). Hybrid methods leverage geometric constraints to improve efficiency and stability, making them more suitable for real-time AR deployment. Nonetheless, maintaining sub-pixel accuracy and temporal smoothness under strict latency constraints remains a significant challenge.

4.3 Tight Coupling with SLAM and Scene Understanding

Object tracking in AR rarely operates in isolation. Tight integration with simultaneous localization and mapping (SLAM) systems enables improved global consistency and robustness. By sharing spatial constraints, tracking estimates can be stabilized using camera pose information, while tracked objects can serve as semantic landmarks for mapping. However, such coupling introduces new challenges. Errors in one module may propagate to others, amplifying failure modes (Zhang et al., 2023). Synchronization between tracking, SLAM, and rendering pipelines is also non-trivial, particularly under asynchronous sensor input. Designing robust interfaces and fallback strategies is therefore essential for reliable AR systems.

4.4 Multimodal Perception Fusion

Modern AR platforms increasingly incorporate multiple sensing modalities, including RGB-D cameras, inertial measurement units (IMUs), and event-based sensors. Multimodal fusion enhances robustness under challenging conditions such as motion blur, low illumination, and rapid motion. Effective fusion strategies must account for differing noise characteristics, temporal resolutions, and failure modes across modalities (Hodan et al., 2023). Probabilistic and confidence-aware fusion frameworks are particularly promising, as they

allow the system to weigh sensor inputs dynamically based on reliability estimates.

4.5 System-Level Design Guidelines for AR-Oriented Tracking

Several system level guidelines emerge for AR oriented tracking. End to end latency must be minimized and predictable, favoring architectures with deterministic execution rather than data dependent computation spikes. Perceptual stability should be treated as a primary objective through temporal smoothing, confidence aware updates, and trajectory regularization while accounting for interactions between tracking, rendering, and user motion. Adaptation is essential for long term robustness but should be controlled via selective updates, bounded memory, and confidence thresholds. Tight integration with SLAM and scene understanding improves global consistency, yet coupling must be designed to prevent cascading failures, motivating modular interfaces, redundancy, and fallback mechanisms. Finally, AR oriented tracking should be evaluated not only by algorithmic metrics but also by human centric criteria, as user studies and perceptual experiments reveal insights beyond benchmark performance.

4.6 System-Level Failure Modes and Practical Lessons from AR Deployment

Beyond algorithmic performance, real-world AR deployment exposes several system-level failure modes that are often underrepresented in tracking benchmarks. Delayed or unstable initialization may disrupt interaction, prompting practical systems to adopt fast and conservative initialization strategies. During prolonged operation, small errors can accumulate and cause spatial misalignment of virtual content, necessitating periodic re-alignment via SLAM or environmental anchors. Latency spikes constitute another critical failure mode, as data-dependent computation or scheduling delays may break the real-time rendering loop; in such cases, worst-case latency matters more than average throughput and motivates bounded-complexity designs. Finally, failure recovery behavior strongly affects perceived robustness: abrupt re-initialization or incorrect re-detection can be visually disturbing, while graceful degradation techniques such as freezing virtual content, fading overlays, or signaling low confidence help maintain a stable and user-centric AR experience.

5. Datasets and Evaluation Protocols

Datasets and evaluation protocols play a central role in shaping the development of object tracking algorithms. However, a critical gap remains between widely used tracking benchmarks and the practical requirements of augmented reality (AR) deployment. This section reviews representative datasets from both generic tracking and AR-related domains, and analyzes their limitations from an AR-oriented perspective.

5.1 Generic Object Tracking Benchmarks

Classical object tracking benchmarks, including OTB, VOT, LaSOT, GOT-10k, and TrackingNet, have been instrumental in driving progress in deep learning–based tracking. These datasets typically emphasize short- to medium-length video sequences, diverse object categories, and standardized evaluation metrics such as precision and success curves. As a result, they provide a useful testbed for measuring algorithmic robustness under challenges such as occlusion, fast motion, and background clutter. Nevertheless, from an AR deployment perspective, these benchmarks exhibit several limitations. First, most sequences are relatively short and do not capture the long-term temporal dynamics characteristic of extended AR sessions. Second, evaluation is conducted offline and open-loop, without accounting for feedback from rendering or user interaction. Third, the metrics prioritize frame-wise accuracy rather than perceptual stability, making it difficult to assess jitter, drift, or latency-induced artifacts that are critical in AR systems.

5.2 6-DoF Object Pose Estimation and Tracking Datasets

For AR applications that require spatial anchoring of virtual content to physical objects, 6-DoF object pose tracking is often more relevant than two-dimensional bounding box tracking. The BOP benchmark series has become a standard evaluation framework for object pose estimation, providing multiple datasets, unified metrics, and reproducible evaluation protocols. While BOP-style datasets are valuable for benchmarking pose accuracy, they often assume controlled capture conditions and limited interaction. Occlusion patterns are typically synthetic or static, and long-term temporal consistency is not a primary evaluation target. As a result, high performance on pose benchmarks does not necessarily translate to perceptually stable AR experiences under real-world usage.

5.3 Emerging AR-Specific and Multimodal Datasets

Recent research has begun to address these gaps by introducing datasets tailored to AR deployment scenarios. Such datasets increasingly incorporate long sequences, handheld or head-mounted camera motion, multimodal sensing (e.g., RGB-D and IMU), and realistic interaction patterns involving hands and tools. These characteristics better reflect the challenges faced by AR tracking systems in practice. Despite this progress, publicly available AR-specific datasets remain limited in scale and diversity. Many datasets are collected using proprietary platforms or restricted environments, hindering reproducibility and fair comparison. Expanding the availability of open, large-scale AR tracking datasets therefore remains an important research priority.

5.4 Evaluation Beyond Accuracy: Toward AR-Relevant Metrics

From an AR-oriented perspective, evaluation protocols must move beyond conventional accuracy metrics. Temporal consistency measures, such as velocity or acceleration variance, provide a more direct proxy for perceptual smoothness. Initialization and recovery latency should also be evaluated explicitly, as delayed or unstable recovery can severely disrupt user interaction. Energy consumption and thermal behavior are additional factors that are rarely considered in tracking benchmarks but are critical for sustained AR operation. Incorporating such metrics into standardized evaluation protocols would significantly improve the relevance of benchmark results for real-world AR deployment.

6. Open Challenges and Future Research Directions

Despite substantial progress in deep learning based object tracking, reliable large scale AR deployment remains challenging. A central issue is robust generalization under open world conditions, where trackers must handle unseen objects, maintain identity over time, and avoid drift under occlusion and appearance variation. Beyond accuracy, AR systems require a joint optimization of accuracy, temporal stability, and computational efficiency, motivating multi objective training and evaluation frameworks that explicitly capture these trade offs. A significant gap also persists between state of the art tracking algorithms and deployable AR systems, as many methods are evaluated in isolation without considering synchronization, memory constraints, or interaction with SLAM and rendering pipelines. As AR applications move toward safety critical and professional domains, trustworthiness becomes essential, with an increasing need for confidence estimates, uncertainty modeling, and interpretable failure signals evaluated through human centric studies. Looking forward, integrating tracking into broader world models that capture dynamics, semantics, and physical constraints may enable predictive tracking, more robust recovery, and deeper interaction capabilities, suggesting that future progress will depend on holistic approaches that unify algorithmic advances with system level design.

7. Conclusion

Deep learning based object tracking has become a core capability for modern AR systems, determining how virtual content is spatially anchored,

temporally stabilized, and perceptually integrated with the physical world. Unlike generic tracking, AR oriented tracking must satisfy stringent system level constraints, including tight latency and energy budgets, long term operation, and strong sensitivity to perceptual instability. This survey has examined recent advances from a system oriented AR perspective and analyzed four major tracking paradigms, highlighting their respective strengths, limitations, and deployment trade offs. Our discussion underscores that benchmark accuracy alone is insufficient for AR deployment; temporal stability, interaction robustness, and predictable computational behavior are equally essential. No single paradigm currently satisfies the diverse requirements of AR, motivating hybrid designs, controlled adaptation strategies, and tighter coupling with complementary modules such as SLAM and multimodal sensing. We argue that AR oriented tracking constitutes a distinct research domain that requires holistic approaches integrating algorithmic accuracy, perceptual stability, and system efficiency, as well as closer collaboration between vision, systems, and HCI. By consolidating recent progress and outlining open challenges, this survey provides a structured reference and research roadmap for next generation AR tracking systems.

References

Alansari M, Hassan M, et al., 2025. Visual tracking by matching points using diffusion models. *Future Generation Computer Systems*, 152, pp. 98–110.

Cai W, Liu Y, et al., 2025. SPMTrack: Spatio-temporal parameter-efficient fine-tuning with mixture of experts for scalable visual tracking. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10452–10461.

Chen B, Li P, Bai L, et al., 2022. Backbone is all your need: A simplified architecture for visual object tracking. In: *Proceedings of the European Conference on Computer Vision*, pp. 375–391.

Chen X, Peng H, Wang D, et al., 2023. SeqTrack: Sequence to sequence learning for visual object tracking. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14535–14544.

Chen X, Yan B, Zhu J, et al., 2021. Transformer tracking. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8126–8135.

Chen Y H, Kristan M, et al., 2023. NeighborTrack: Single object tracking by bipartite matching with neighbor tracklets. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 1823–1832.

Cui Y, Jiang C, Wang L, et al., 2022. MixFormer: End-to-end tracking with iterative mixed attention. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13608–13617.

Dai K, Zhang Y, Wang D, et al., 2020. High-performance long-term tracking with meta-updater. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6298–6307.

Danelljan, M., Goutam, B. and Khan, F.S., 2020. Probabilistic regression for visual tracking. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7183–7192.

Gao S, Zhou C, Ma C, et al., 2022. AiATrack: Attention in attention for transformer visual tracking. In: *Proceedings of the European Conference on Computer Vision*, pp. 282–299.

Guo M, Zhang R, et al., 2025. DreamTrack: Dreaming the future for multimodal visual object tracking. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11234–11243.

Hodan T, Sundermeyer M, Drost B, et al., 2020. BOP Challenge 2020 on 6D object localization. In: *Proceedings of the European Conference on Computer Vision Workshops*, pp. 577–594.

Hodan T, Sundermeyer M, et al., 2023. BOP Challenge 2022 on detection, localization and pose estimation. *International Journal of Computer Vision*, 131(8), pp. 1940–1962.

Hong L, Wang Z, et al., 2024. OneTracker: Unifying visual object tracking with foundation models and efficient tuning. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16412–16421.

Kristan M, Leonardis A, Matas J, et al., 2023. The VOTS2023 challenge results. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pp. 312–331.

Kristan M, Leonardis A, Matas J, et al., 2024. The visual object tracking challenge 2024 results. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 1–15.

Labbé Y, Carpentier J, Aubry M, et al., 2020. CosyPose: Consistent multi-view multi-object 6D pose estimation. In: *Proceedings of the European Conference on Computer Vision*, pp. 574–591.

Li Y, Zhang H, et al., 2024. A transformer-based visual object tracker via learning immediate appearance change information in videos. *Pattern Recognition*, 148, pp. 110125.

Liang S, Chen Z, et al., 2025. Autoregressive sequential pretraining for visual tracking. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9231–9240.

Liu X, Wang Y, et al., 2025. MambaVLT: Time-evolving multimodal state space model for vision-language tracking. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12801–12810.

Mayer C, Danelljan M, Paudel D P, et al., 2021. Learning target candidate association to keep track of what not to track. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 13444–13454.

Mayer C, Danelljan M, Paudel D P, et al., 2022. Transforming model prediction for tracking. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8731–8740.

Paul M, Danelljan M, Van Gool L, et al., 2022. Robust visual tracking by segmentation. In: *Proceedings of the European Conference on Computer Vision*, pp. 571–588.

Ravi N, et al., 2024. SAM 2: Segment anything in images and videos. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4012–4021.

Shim K, Lee S, et al., 2025. Focusing on tracks for online multi-object tracking. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14903–14912.

Sun C, Zhao J, et al., 2025. Exploring historical information for RGBE visual tracking with Mamba. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13592–13601.

Videnovic J, et al., 2025. A distractor-aware memory for visual object tracking with SAM2. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14122–14131.

Wang N, Zhou W, Wang J, et al., 2021. Transformer meets tracker: Exploiting temporal context for robust visual tracking. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15711–15720.

Wang X, Zhang Y, et al., 2024. Event stream-based visual object tracking: A high-resolution benchmark dataset and evaluation. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11983–11992.

Wen B, Yang W, Birchfield S, et al., 2021. BundleTrack: 6D pose tracking for novel objects without instance or category-level 3D models. In: *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 806–813.

Wen B, Yang W, Kautz J, et al., 2023a. BundleSDF: Neural 6-DoF tracking and 3D reconstruction of unknown objects. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15861–15870.

Wen B, Yang W, Kautz J, et al., 2023b. Neural object-centric tracking and reconstruction in the wild. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 18456–18465.

Wen B, Yang W, Kautz J, et al., 2024. FoundationPose: Unified 6D pose estimation and tracking of novel objects. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14291–14300.

Wu J, Jiang Y, Liu Q, et al., 2024. General object foundation model for images and videos at scale. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5123–5132.

Xie F, Wang C, Wang G, et al., 2024. DiffusionTrack: Point set diffusion model for visual object tracking. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8751–8760.

Yan B, Peng H, Wu K, et al., 2021. Learning spatio-temporal transformer for visual tracking. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10448–10457.

Yu B, Tang M, Zheng L, et al., 2021. High-performance discriminative tracking with transformers. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9854–9863.

Zhang T, Huang L, et al., 2023. Unified multimodal tracking with event cameras and RGB sensors. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 19342–19351.

Zhao M, Okada K, Inaba M., 2021. TrTr: Visual tracking with transformer. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pp. 156–165.