WWW Companion '26 · April 13–17, Dubai, UAE

Rethinking Object Detection & Tracking

100+
Papers reviewed

4
Taxonomy axes

2022–2025
Coverage span

Arpit Sharma · Christ University, Bengaluru

Dr. Lei Wang · Griffith University & Data61/CSIRO

Dr. Yongsheng Gao · Griffith University

Presentation Roadmap

01

Motivation & Problem

Why a new survey? What's missing?

02

Our Taxonomy

4-axis framework for modern methods

03

Quantitative Trends

Datasets, metrics, challenge analysis

04

Key Insights

What patterns emerge across 100+ papers

05

Architectural Paradigms

Transformer · Diffusion · State-Space

06

Open Challenges

Five frontier problems that remain unsolved

Why Another Survey?

Challenges — not task labels — now define the research frontier.

01

Fragmented literature

100+ works across 2D, 3D, multimodal, vision-language — no unified cross-task lens

02

Task-centric surveys miss connections

Detection vs. Tracking boundaries dissolving: joint frameworks, end-to-end identity prediction, diffusion trackers

03

Shared challenges ignored across tasks

Occlusion · Identity switches · Fast motion · Domain shift — appear in every subfield yet studied in isolation

04

Evaluation lags deployment

mAP and HOTA omit long-term robustness, semantic correctness, and uncertainty — critical for real-world use

Three Core Contributions

What makes this survey different

01

Challenge-Driven Synthesis

Reorganize 100+ papers around real-world bottlenecks — occlusion, identity confusion, fast motion, domain shift — rather than traditional task labels.

Cross-task transferable insights

02

Quantitative Trend Analysis

Unified, manually-annotated statistics on task focus, architectures, sensing modalities, datasets, evaluation metrics, and challenge prevalence (2024–2025).

Exposes benchmark biases & gaps

03

Forward-Looking Perspective

Identifies paradigm shifts — unified detection-tracking, diffusion temporal reasoning, vision-language grounding, uncertainty-aware modeling — and open problems.

Roadmap for next-phase research

4-Axis Taxonomy of Modern Methods

All four axes show a consistent directional shift — the field is moving right on every dimension

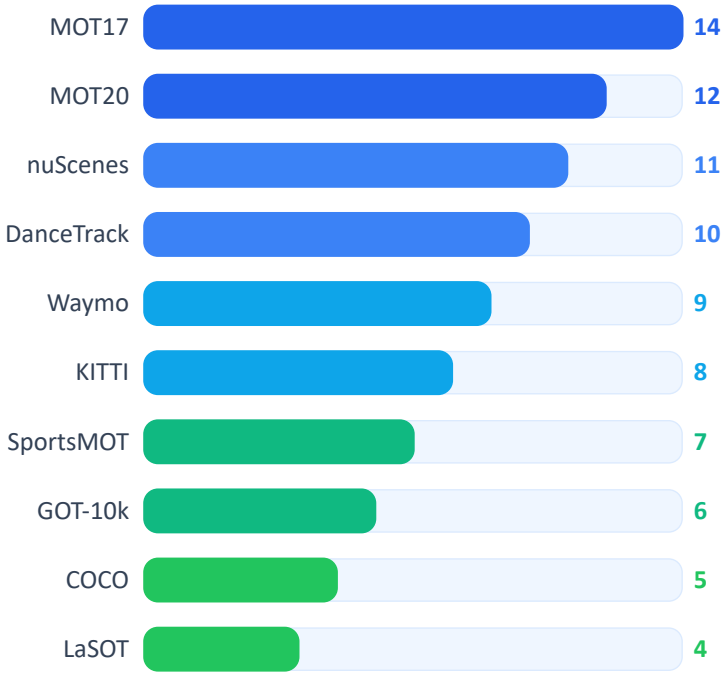
Task Formulation	Closed-set detection + tracking-by-detection	→ Joint detection-tracking + end-to-end trajectory prediction
Supervision & Openness	Fully supervised closed-world assumptions	→ Open-vocabulary / few-shot open-world + self-supervised
Spatial-Temporal Scope	2D frame-level perception short-term association	→ 3D/BEV + multi-view long-term memory tracking
Sensing Modality	RGB-only single sensor systems	→ LiDAR · Thermal · Event-camera Vision-Language fusion

One method can be transformer-based, open-vocabulary, and 3D-aware simultaneously

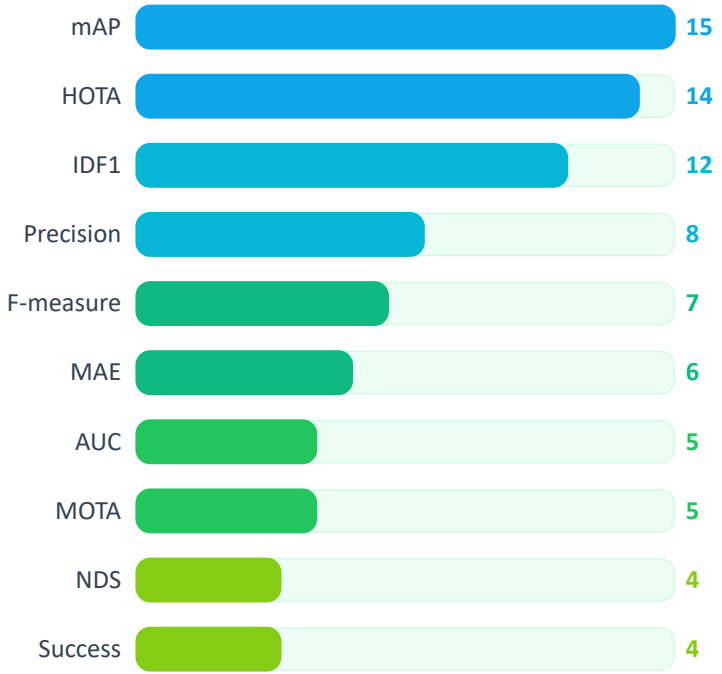
Quantitative Trends — Datasets & Evaluation Metrics

Manually annotated from 100+ peer-reviewed papers (2024–2025) · multi-label scheme

Top Benchmark Datasets



Evaluation Metric Frequency



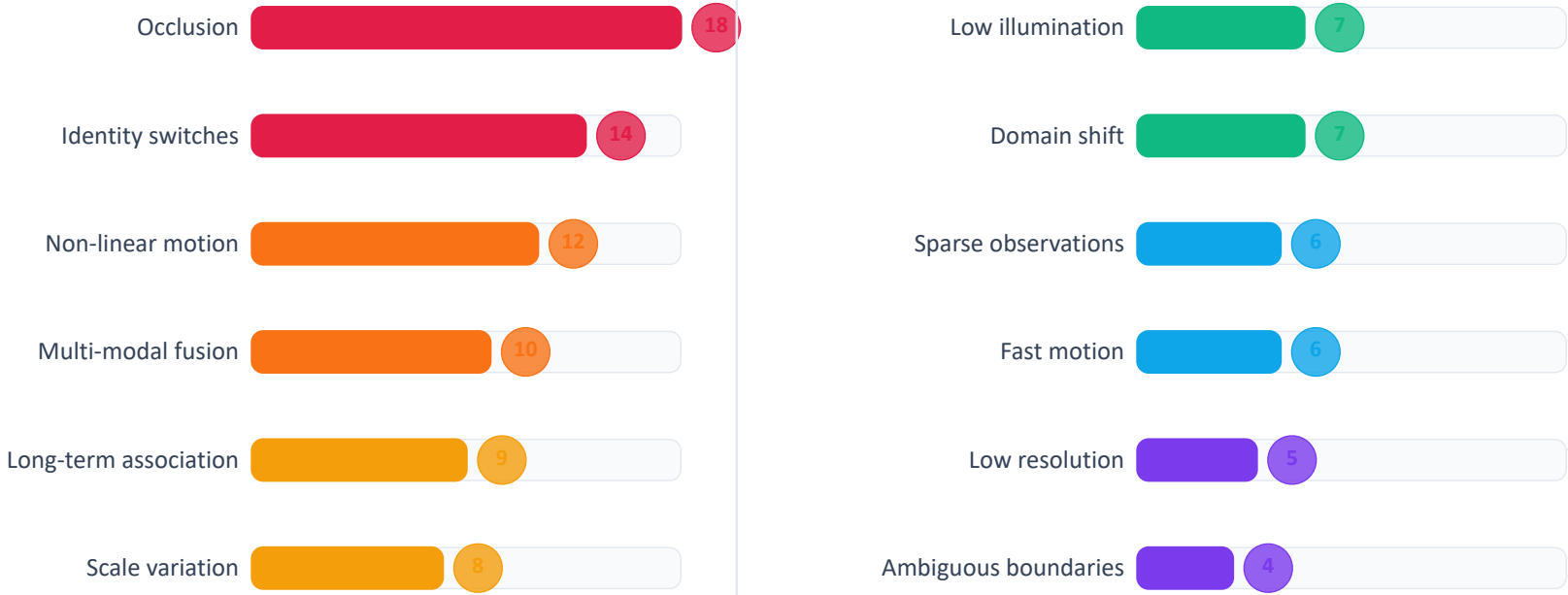
Top 3 → datasets account for >50% of all benchmark mentions

HOTA → fastest-rising tracker metric — identity-aware, not frame-local

VL methods → fastest growing modality subset enabled by large multimodal models

Research Challenge Landscape

Frequency of explicitly targeted challenges in 2024–2025 literature · extracted per-paper across 100+ works



⚠ Key Research Gap:

Uncertainty estimation, annotation efficiency, and long-term open-world scalability are severely under-addressed relative to their deployment importance — fewer than 5% of papers directly tackle these.

3 Key Insights

Derived from cross-task synthesis — these patterns hold across all 100+ surveyed works

1

Temporal structure — not per-frame accuracy — is the key performance lever

Methods explicitly encoding long-horizon attention, sequence modeling, or generative processes consistently outperform frame-local approaches under heavy occlusion and non-linear motion.

2

Semantic flexibility without uncertainty modeling shifts — not resolves — failure modes

Open-vocabulary detectors introduce fluctuating category assignments that directly destabilize downstream tracking. Success requires identity-consistency mechanisms alongside semantic grounding.

3

Benchmark optimization increasingly obscures real-world failure modes

Gains on MOT17/HOTA don't transfer to open-set conditions, domain shift, or sensor degradation. Reported progress is scenario-specific; challenge-aware evaluation is overdue.

Three Dominant Architectural Paradigms

Choose based on operational regime — no single architecture dominates universally

Transformer-Based

~60% of recent papers · Steady growth

Best for:

Long-term interactions dominate

Strengths

- ✓ Global context & long-range dependencies
- ✓ Joint detection-tracking & 3D perception
- ✓ Strong identity preservation under occlusion
- ✓ Flexible multi-modal fusion

Limitations

- ✗ High compute & memory cost
- ✗ Data-hungry; pretraining needed

Diffusion-Based

< 15% of papers · Fastest growing

Best for:

Uncertain or sparse observations

Strengths

- ✓ Probabilistic temporal generation
- ✓ Handles noise & irregular observations
- ✓ Emerging for motion estimation

Limitations

- ✗ High inference latency
- ✗ Complex training pipelines

State-Space / Recurrent

Stable at ~25% · Dominant in edge

Best for:

Real-time or resource-constrained

Strengths

- ✓ Efficient real-time tracking
- ✓ Hybrid: learned + classical motion
- ✓ Interpretable & stable

Limitations

- ✗ Limited long-range context modeling
- ✗ Less flexible to distribution shift

Five Open Challenges & Future Directions

Frontier problems that will shape the next generation of detection and tracking systems

Scalable Open-World Detection & Tracking

Novel categories, evolving scenes, long-term identity stability under semantic uncertainty. Needs continual learning + principled memory management.

Unified Evaluation Protocols

mAP / HOTA / IDF1 are task-incompatible. Need unified, challenge-aware frameworks measuring temporal consistency, semantic correctness, and uncertainty.

Long-Term Temporal Reasoning

Identity recovery after extended occlusion; sparse-observation reasoning. Requires generative temporal models + memory-augmented architectures.

Efficient Integration of Large Multimodal Models

VLMs powerful but costly. Lightweight prompting + modular architectures needed to balance open-vocabulary performance and temporal consistency.

Robustness, Uncertainty & Reliability

Safety-critical deployment (autonomous driving, medical) demands calibrated confidence at spatial, temporal, and semantic levels — currently absent.

How This Survey Differs From Prior Work

Comparison along six key dimensions with existing surveys

Dimension	Typical Prior Surveys	This Survey
Organization	Task-centric (detect vs. track)	**Challenge-centric (cross-task)**
Coverage	Single task or single modality	**2D, 3D, multi-view, VL, multimodal**
Time horizon	One-to-two year snapshots	**2022–2025, full paradigm shift**
Analysis type	Qualitative / descriptive	**Quantitative + manually annotated**
Evaluation lens	Benchmark performance only	**Benchmark + deployment gap analysis**
Forward view	Trend listing	**Open problem synthesis + roadmap**

Key Take-Aways

1

Challenges — not task labels — define the research frontier

Shared bottlenecks (occlusion, identity, motion) cut across ALL tasks and modalities

2

Temporal structure is the critical lever for tracking performance

Diffusion-based and transformer-based temporal modeling are the dominant growth stories

3

Evaluation protocols must urgently catch up to deployment reality

No single metric (mAP, HOTA, IDF1) suffices; challenge-aware evaluation is the path forward

4

Open-world scalability + uncertainty remain largely unsolved

Category discovery · Identity stability · Calibrated confidence = the grand unified challenge

Thank you · Questions? *arpit.sharma@mtech.christuniversity.in* · DOI: 10.1145/3774905.3794677