CVPR JUNE 3-7, 2026 DENVER COLORADO

Highlight 🌟

Real-World Point Tracking with Verifier-Guided Pseudo-Labeling

Görkay Aydemir[1] Fatma Güney[1,2,†] Weidi Xie[3,†]

[1] Koç University
[2] KUIS AI Center
[3] School of AI, Shanghai Jiao Tong University

[†] Equal supervision

Point Tracking

The task

Given a query point, predict its position over the rest of the video.

Query points

Tracks

Formally: given a query point q at time t_0, predict positions p_t and visibilities v_t for every later frame.

Applications: video editing · 4D scene understanding · robot manipulation · medical imaging

But trackers degrade on real videos

Trained on

Synthetic videos

Dense ground-truth trajectories (TAP-Vid Kubric)



Deployed on

Real videos

No dense annotations available, models degrade.

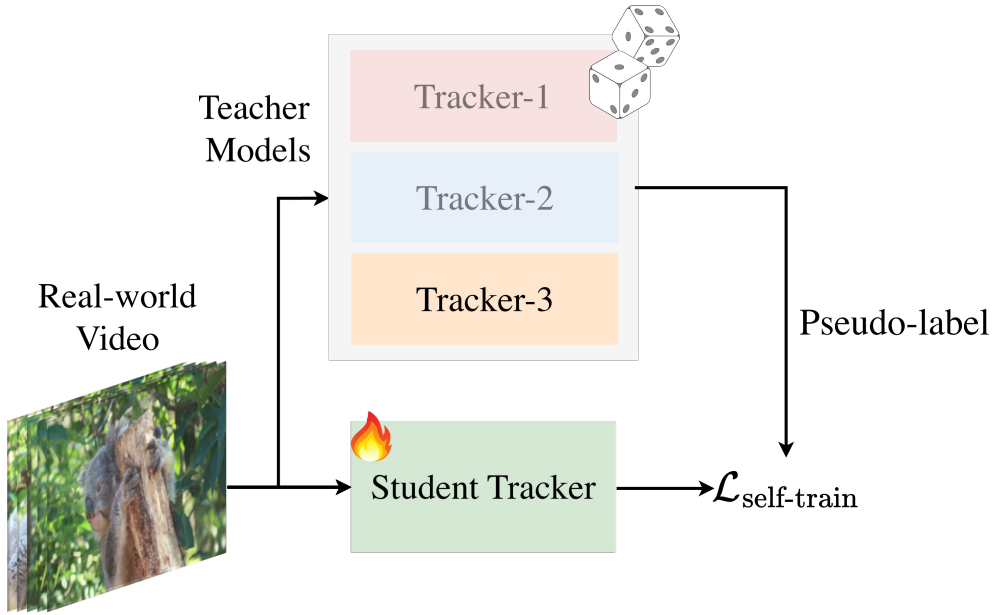

THE NEED

Train on real, unlabeled videos without dense ground-truth trajectories.

Naive pseudo-labeling: pick one teacher, hope it's right

Pretrained tracker → pseudo-labels → fine-tune

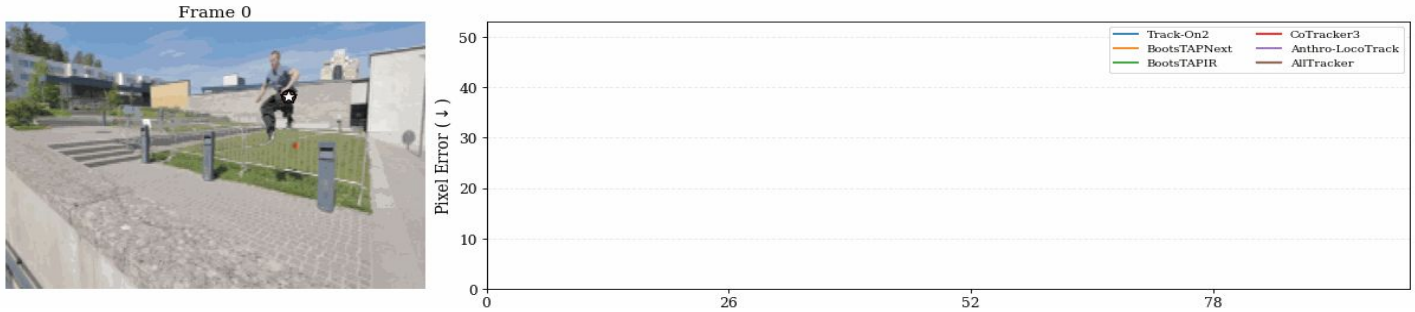
Prior-work pipeline



...but how reliable is a single dice-rolled teacher?

Same video, same point, yet teachers disagree

Multi-teacher disagreement: predictions on the same point



Predictions vary frame-to-frame

A teacher accurate on frame 50 may drift completely by frame 130.

Different trackers, different failures

Some resist fast motion but drift on low texture; others handle occlusion but flip identities.

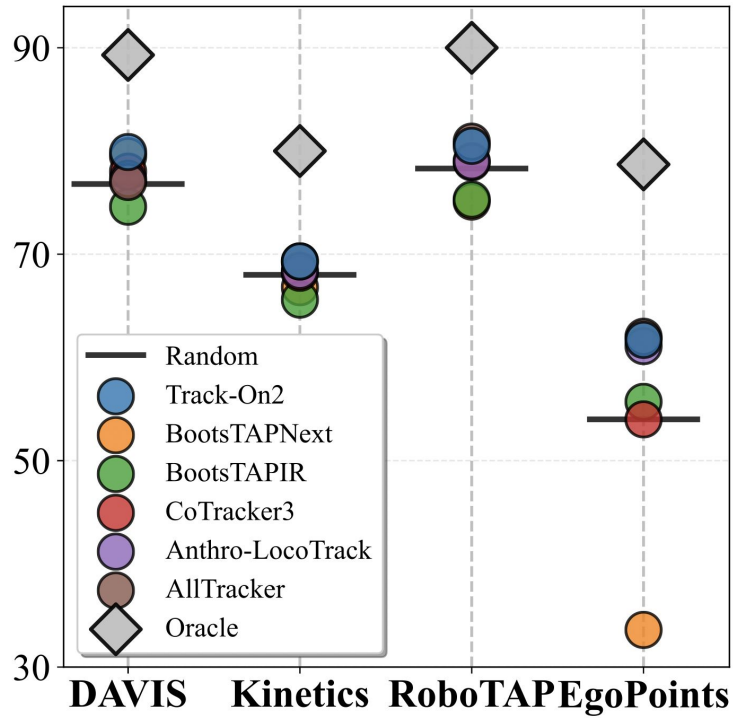
Random pick amplifies noise

Rolling a die over noisy teachers propagates errors instead of cancelling them out.

Key Insight

Per-frame best ≫ any single tracker

Oracle vs. individual teachers (δ_{avg} ↑)



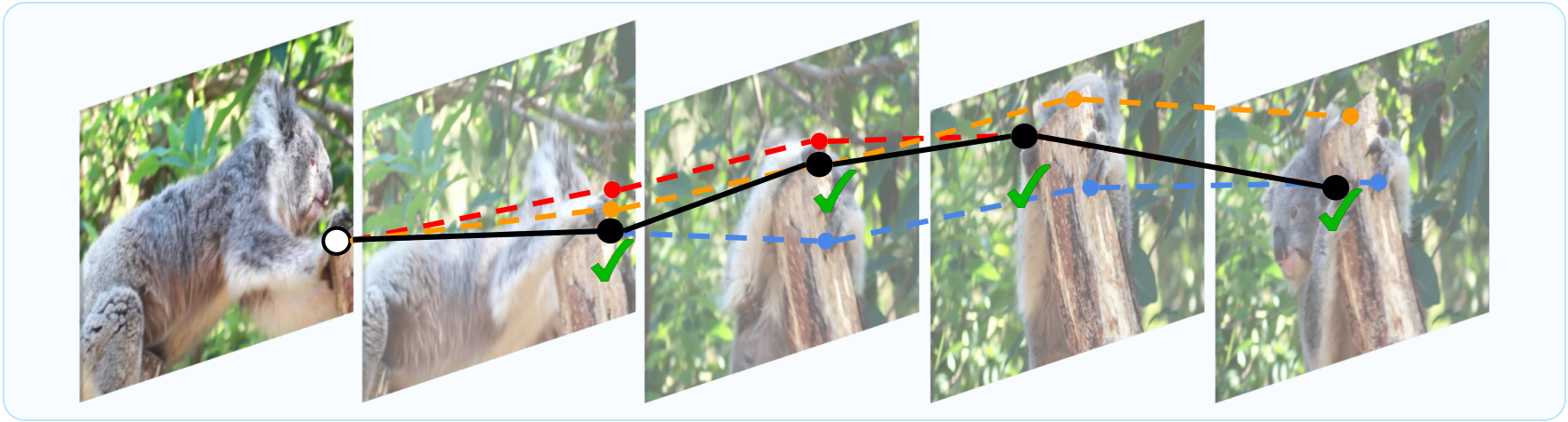
The best tracker changes per frame

Static or random selection is fundamentally suboptimal, the gap is real headroom.

Q.

Can we learn a model that approximates this oracle on real videos, without ground truth?

Verifier: a meta-model that scores per-frame reliability



Multi-teacher

M off-the-shelf trackers produce candidate trajectories

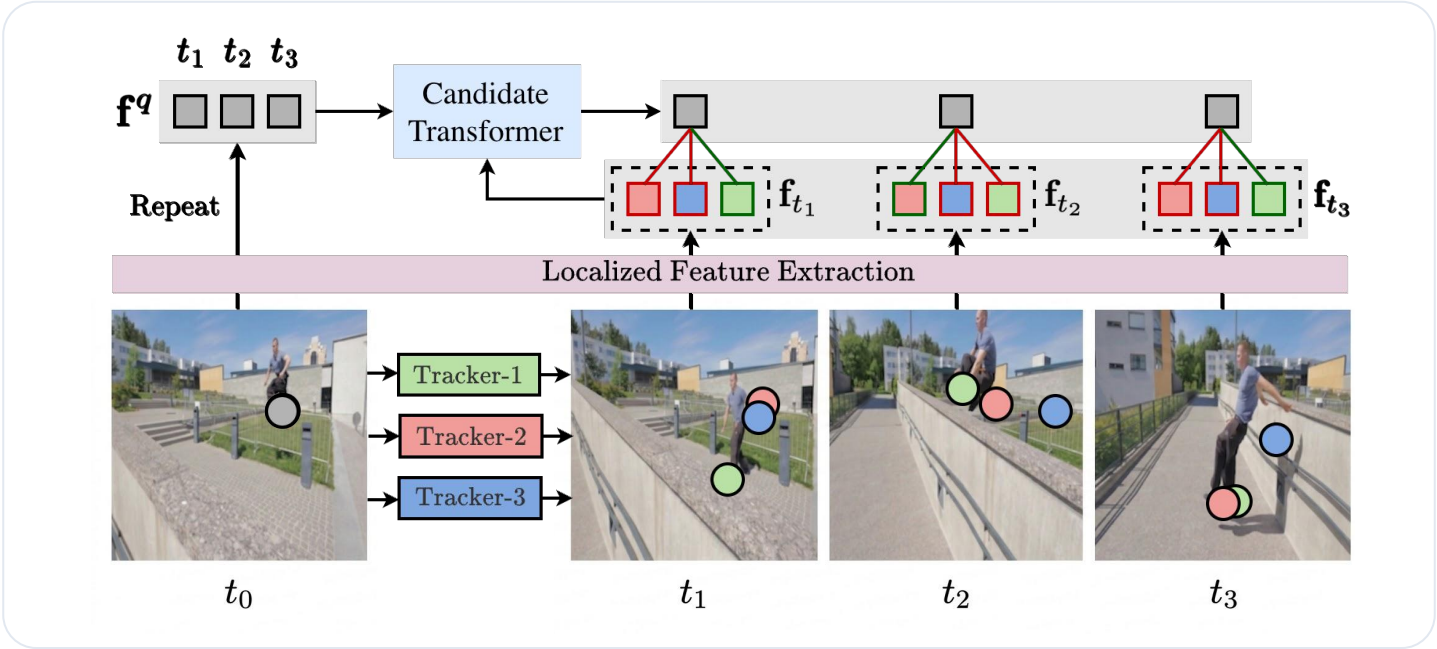
Per-frame reliability

Verifier scores each candidate at every frame

Two uses

Train-time pseudo-label selector
+ inference-time ensemble

Localized features + Candidate Transformer



1. **Localized features**
2. **Candidate transformer**
3. **Reliability scores**

Train the verifier on synthetic data only

STAGE 1

How the verifier learns: generate fake errors, rank candidates by proximity to GT

Synthetic only

Trained on K-EPIC (11K clips).

Realistic errors

Six perturbation types simulate drift, jumps, jitter, occlusion, and identity switches.

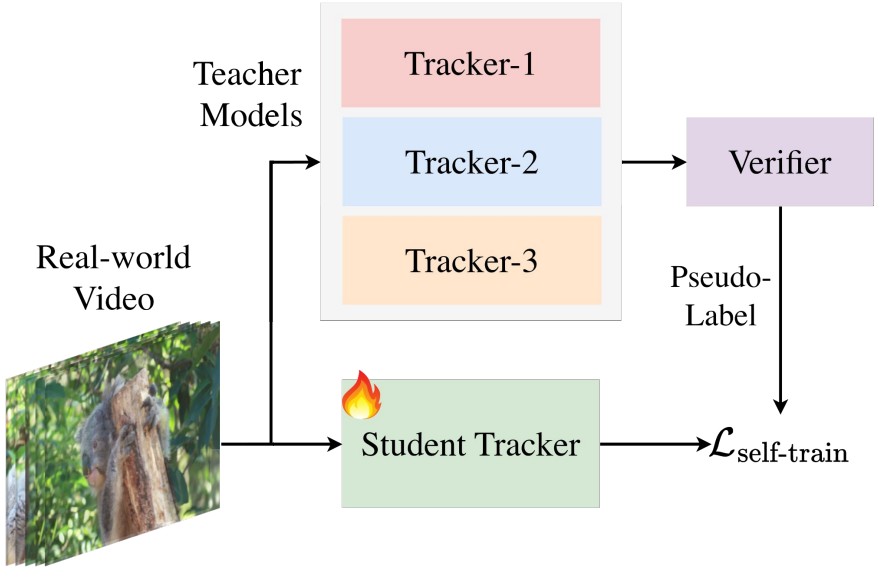
Ranking objective

Cross-entropy against soft targets based on per-frame distance to GT.

Use the verifier to label unlabeled real videos

STAGE 2

How we label real videos:
ensemble of teachers, verifier picks the most reliable per frame



Unlabeled real-world videos

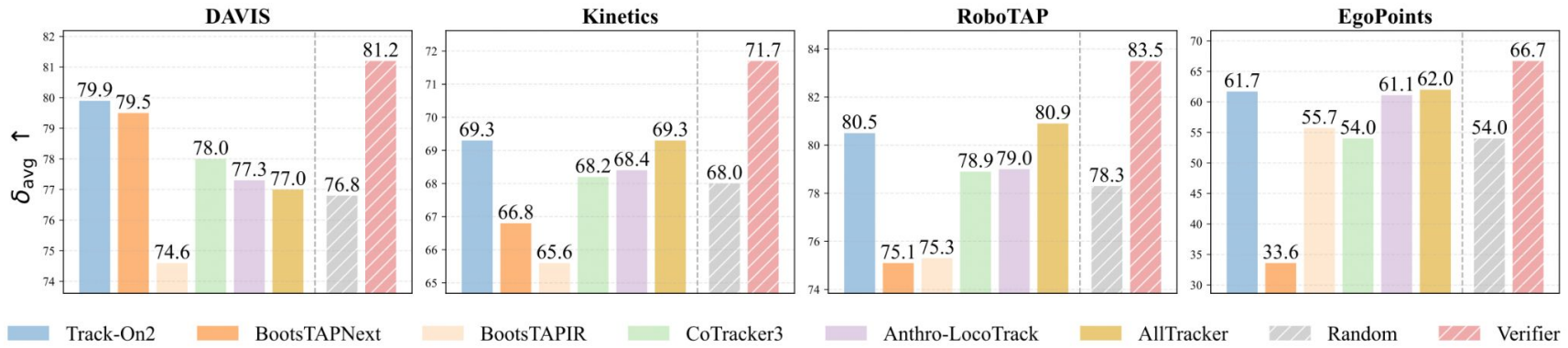
~ 5K TAO + OVIS + VSPW

Track-On2 → Track-On-R

Data efficient

Cleaner supervision

Verifier beats every individual teacher



Plug-and-play ensemble

Verifier turns model diversity into a strength by identifying frame-level expertise.

+4.7 δ_{avg}

vs. Random Teacher

Significant gain by intelligently selecting predictions.

+1.8 δ_{avg}

vs. Best Teacher

Outperforms even the strongest individual off-the-shelf model.

Track-On-R: benchmark comparison

δ_{avg} ↑ on benchmarks, fine-tuned with verifier-guided pseudo-labels

Model	Ego	RoboTAP	Kinetics	DAVIS	PO	DR
Track-On2 (baseline)	61.7	80.5	69.3	79.9	45.1	74.5
BootsTAPNext	33.6	75.0	70.6	78.5	9.9	46.2
CoTracker3	54.0	78.8	68.5	76.3	44.5	72.3
AllTracker	62.0	80.9	69.3	77.0	—	—
Track-On-R (ours)	**67.3**	**82.6**	**71.0**	**80.3**	**53.4**	**75.1**
Δ vs. Track-On2	**+5.6**	**+2.1**	**+1.7**	**+0.4**	**+8.3**	**+0.6**

Takeaways

Conclusion

01

A learned verifier

Scores per-frame reliability across multiple trackers: plug-and-play at inference, sharp pseudo-labels at training.

02

Annotation-free adaptation

Stage-1 train on synthetic only, then fine-tune on ~5K unlabeled real videos.

03

State-of-the-art

Across public benchmarks, outperforming every individual teacher model.

References & Acknowledgements

TAPVid: A Benchmark for Tracking Any Point in Video
Doersch et al.
NeurIPS 2022

CoTracker3: Simpler and Better Point Tracking by Pseudo-Labelling Real Videos
Karaev et al.
ICCV 2025

Tapnext: Tracking any point (tap) as next token prediction
Zholus et al.
ICCV 2025

AllTracker: Efficient dense point tracking at high resolution
Harley et al.
ICCV 2025

Tao: A large-scale benchmark for tracking any object
Dave et al.
ECCV 2020

Occluded video instance segmentation: A benchmark.
Qi et al.
IJCV 2022

Vspw: A large-scale dataset for video scene parsing in the wild
Miao et al.
CVPR 2021

BootsTAP: Bootstrapped training for tracking-any-point
Doersch et al.
ACCV 2024

Learning to track any points from human motion
Kim et al.
arXiv 2025

Track-on2: Enhancing online point tracking with memory
Aydemir et al.
TPAMI 2026

Thank You!

Project Page: kuis-ai.github.io/track_on_r

Code: github.com/gorkaydemir/track_on

Contact: gorkayaydemir@gmail.com