AnthroTAP: Learning Point Tracking with Real-World Motion

Inès Hyeonsu Kim[1,3*] Seokju Cho[1*] Jahyeok Koo[1] Junghyun Park[1] Jiahui Huang[2] Honglak Lee[3,4] Joon-Young Lee[2] Seungryong Kim[1]

[1]KAIST AI [2]Adobe Research [3]University of Michigan [4]LG AI Research

CVPR JUNE 3-7, 2026 DENVER COLORADO

TL;DR AnthroTAP automatically annotates **real human videos** into a large-scale, **non-proprietary point-tracking dataset**.

1.4K	1 day	SOTA
Real videos (Anthro-LD)	Training · 4 GPUs	TAP-Vid & RoboTAP

What is Point Tracking?

Input

Query Point

$q = (x_q, y_q, t_q) \in \mathbb{R}^3$

Video Sequence

$\mathcal{V} = \{I_t\}_{t=0}^{T-1},\ I_t \in \mathbb{R}^{H \times W \times 3}$

Goal

Trajectory

$\mathcal{T} = \{T_t\}_{t=0}^{T-1},\ T_t \in \mathbb{R}^2$

Occlusion Probability

$\mathcal{O} = \{O_t\}_{t=0}^{T-1},\ O_t \in [0,1]$

Task definition. Given a query point in a frame, point tracking predicts its **location** and **visibility** throughout the video, even under deformation, fast motion, and occlusion.

Real-world point tracks are difficult to annotate at scale!

Solution: A Pseudo-Labeling Pipeline for a Real-World Point Tracking Training Dataset

Existing Synthetic Training Data

Kubric Point Odyssey

- ✗ Unrealistic synthetic appearance
- ✗ Simple or rigid motion
- ✗ Limited / simple occlusions
- ✗ Large sim-to-real domain gap

AnthroTAP Annotated Data

Real human videos (Let's Dance)

- ✓ Non-rigid articulated motion
- ✓ Frequent self / inter-person occlusions
- ✓ Real-world appearance, blur, lighting
- ✓ Automatic pseudo-label generation

Motivation

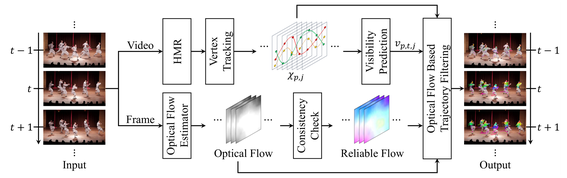
Human motion is naturally complex: bodies deform, limbs articulate, people occlude each other, and videos contain real-world blur, lighting, and appearance variation. Instead of manually annotating point tracks, we ask:

> **Can we distill reliable point trajectories from human mesh recovery models and use them to train general point trackers?**

Highlights

- **Automatic supervision.** Generates point-track pseudo-labels from real human videos using SMPL mesh vertices.
- **Reliability control.** Handles human occlusions with ray casting and filters noisy tracks via optical-flow consistency.
- **Data efficiency.** Improves state-of-the-art trackers with only **1.4K real videos** and **1 day of training on 4 GPUs.**
- **Generalization.** Training on human motion improves tracking not only on humans, but also on non-human objects.

AnthroTAP Pseudo-Labeling Pipeline



AnthroTAP converts real human videos into reliable point-track pseudo-labels by combining **human mesh recovery**, **ray-casting visibility**, and **optical-flow-based filtering**.

Quantitative Comparison — TAP-Vid & RoboTAP (best **bold**, Ours highlighted)

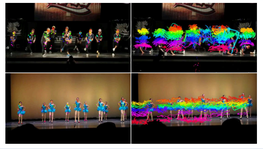
Method	Train Data	DAVIS First AJ	DAVIS First <δˣ	DAVIS First OA	DAVIS Strided AJ	DAVIS Strided <δˣ	DAVIS Strided OA	Kinetics First AJ	Kinetics First <δˣ	Kinetics First OA	RoboTAP First AJ	RoboTAP First <δˣ	RoboTAP First OA
Models evaluated at 256 × 256 resolution													
OmniMotion	–	–	–	–	51.7	67.5	85.3	–	–	–	–	–	–
Dino-Tracker	–	–	–	–	62.3	78.2	87.5	–	–	–	–	–	–
TAPNet	Kub	33.0	48.6	78.8	38.4	53.1	82.3	38.5	54.4	80.6	–	–	–
TAPIR	Kub	58.5	70.0	86.5	61.3	73.6	88.8	49.6	64.2	85.0	59.6	73.4	87.0
TAPTR	Kub	63.0	76.1	91.1	66.3	79.2	91.0	49.0	64.4	85.2	60.1	75.3	86.9
TAPTRv2	Kub	63.5	75.9	**91.4**	66.4	78.8	91.3	49.7	64.2	85.7	–	–	–
TAPTRv3	Kub	63.2	76.7	91.0	–	–	–	54.5	67.5	**88.2**	–	–	–
BootsTAPIR	Kub+15M	61.4	74.0	88.4	66.2	78.5	90.7	**54.6**	**68.4**	86.5	**64.9**	**80.1**	86.3
LocoTrack (baseline)	Kub	63.0	75.3	87.2	67.8	79.6	89.9	52.9	66.8	85.3	62.3	76.2	87.1
Anthro-LocoTrack (Ours)	Kub+1.4K	**64.8**	**77.3**	**89.1**	**69.0**	**81.0**	**90.8**	**53.9**	**68.4**	**86.4**	**64.7**	**79.2**	**88.4**
Improvement over baseline		*+1.8*	*+2.0*	*+1.9*	*+1.2*	*+1.4*	*+0.9*	*+1.0*	*+1.6*	*+1.1*	*+2.4*	*+3.0*	*+1.3*
TAPNext (baseline)	Kub	62.4	76.6	90.5	65.4	79.7	88.9	–	–	–	59.8	73.1	88.1
BootsTAPNext	Kub+15M	65.2	78.5	91.2	68.9	82.4	91.6	–	–	–	64.1	75.1	88.8
Anthro-TAPNext (Ours)	Kub+1.4K	66.1	79.3	91.7	71.4	83.5	92.4	–	–	–	**63.4**	**76.3**	90.2
Improvement over baseline		*+3.7*	*+2.7*	*+1.2*	*+6.0*	*+3.8*	*+3.5*	–	–	–	*+3.6*	*+3.2*	*+2.1*
Models evaluated at 384 × 512 resolution													
CoTracker2	Kub	62.2	75.7	89.3	65.9	79.4	89.9	48.8	64.5	85.8	–	–	–
Track-On	Kub	65.0	78.0	90.8	–	–	–	53.9	67.3	87.8	–	–	–
CoTracker3 (online)	Kub64+15K	**64.4**	76.9	**91.2**	–	–	–	54.7	67.8	87.4	–	–	–
CoTracker3 (offline)	Kub64+15K	63.8	76.3	90.2	–	–	–	**55.8**	68.5	**88.3**	–	–	–
LocoTrack (baseline)	Kub	64.8	77.4	86.2	69.4	81.3	88.6	52.3	66.4	82.1	–	–	–
Anthro-LocoTrack (Ours)	Kub+1.4K	**65.9**	**78.9**	**87.3**	**71.1**	**82.9**	**90.3**	**54.8**	**68.6**	**85.3**	–	–	–
Improvement over baseline		*+1.1*	*+1.5*	*+1.1*	*+1.7*	*+1.6*	*+1.7*	*+2.5*	*+2.2*	*+3.2*	–	–	–

Fine-tuning LocoTrack and TAPNext with AnthroTAP pseudo-labels **consistently improves** performance on TAP-Vid and RoboTAP. Despite using only **1.4K real videos**, AnthroTAP outperforms or matches methods trained with much larger real-video collections.

Qualitative Comparison — training-data trajectory richness

DriveTrack

Anthro-LD (Ours)



Anthro-LD contains complex, non-rigid, articulated human motion with frequent occlusions, while **DriveTrack** mainly captures rigid object motion in driving scenes. AnthroTAP generates richer real-world trajectories for training point trackers.

Generalization beyond humans

Benchmark	Method	AJ	<δˣ	OA
DAVIS (Human Only)	LocoTrack	50.7	42.4	57.8
	Anthro-LocoTrack (Ours)	**51.2** +0.5	**43.3** +0.9	**58.6** +0.8
DAVIS (Non-Human)	LocoTrack	58.8	73.1	83.9
	Anthro-LocoTrack (Ours)	**60.8** +2.0	**75.2** +2.1	**85.3** +1.4

Q: Does tracking learned from human points generalize beyond humans?

Yes. Although AnthroTAP is trained from human motion, it improves tracking on **both human and non-human regions**, with larger gains on non-human points.

Stronger supervision than self-training

	Model	Train Strategy	Dataset	AJ	<δˣ	OA
(I)	CoTracker3	Supervised	Kubric	63.8	76.3	90.2
(II)	CoTracker3	Self-training	Let's Dance	64.2 +0.4	76.5 +0.2	89.6 −0.6
(III)	**CoTracker3**	**AnthroTAP (Ours)**	**Anthro-LD**	**65.0** +1.2	**77.3** +1.0	**90.7** +0.5

Q: Is AnthroTAP more effective than self-training on the same videos?

Yes. On the identical Let's Dance videos and CoTracker3 baseline, AnthroTAP pseudo-labels provide **stronger supervision** than self-training.

Pseudo-label accuracy vs. human annotation

Method	<δ⁰	<δ¹	<δ²	<δ³	<δ⁴	<δˣ
LocoTrack prediction	13.2	36.2	62.3	81.0	87.7	56.1
AnthroTAP pseudo-label (Ours)	**18.0**	**43.2**	**74.8**	**92.4**	**94.1**	**64.5**

Position accuracy (%) against human-annotated trajectories on Let's Dance.

Q: Are AnthroTAP pseudo-labels accurate enough for supervision?

Yes. Compared with human annotations, AnthroTAP pseudo-labels are **substantially more accurate** than LocoTrack predictions across all position thresholds.

Richer real-world motion

Training Dataset	Data Type	Complexity ↑	Diversity
DriveTrack	Real	0.4396	0.0073
PointOdyssey	Synthetic	0.5222	0.1597
Kubric	Synthetic	0.1772	0.1165
Anthro-LD (Ours)	**Real**	**1.2492**	**0.1008**

Q: Do AnthroTAP trajectories capture richer real-world motion?

Yes. Anthro-LD achieves the **highest trajectory complexity** and much higher diversity than DriveTrack, showing that human motion provides richer training signals.