Carnegie Mellon University

Particle Video Revisited:

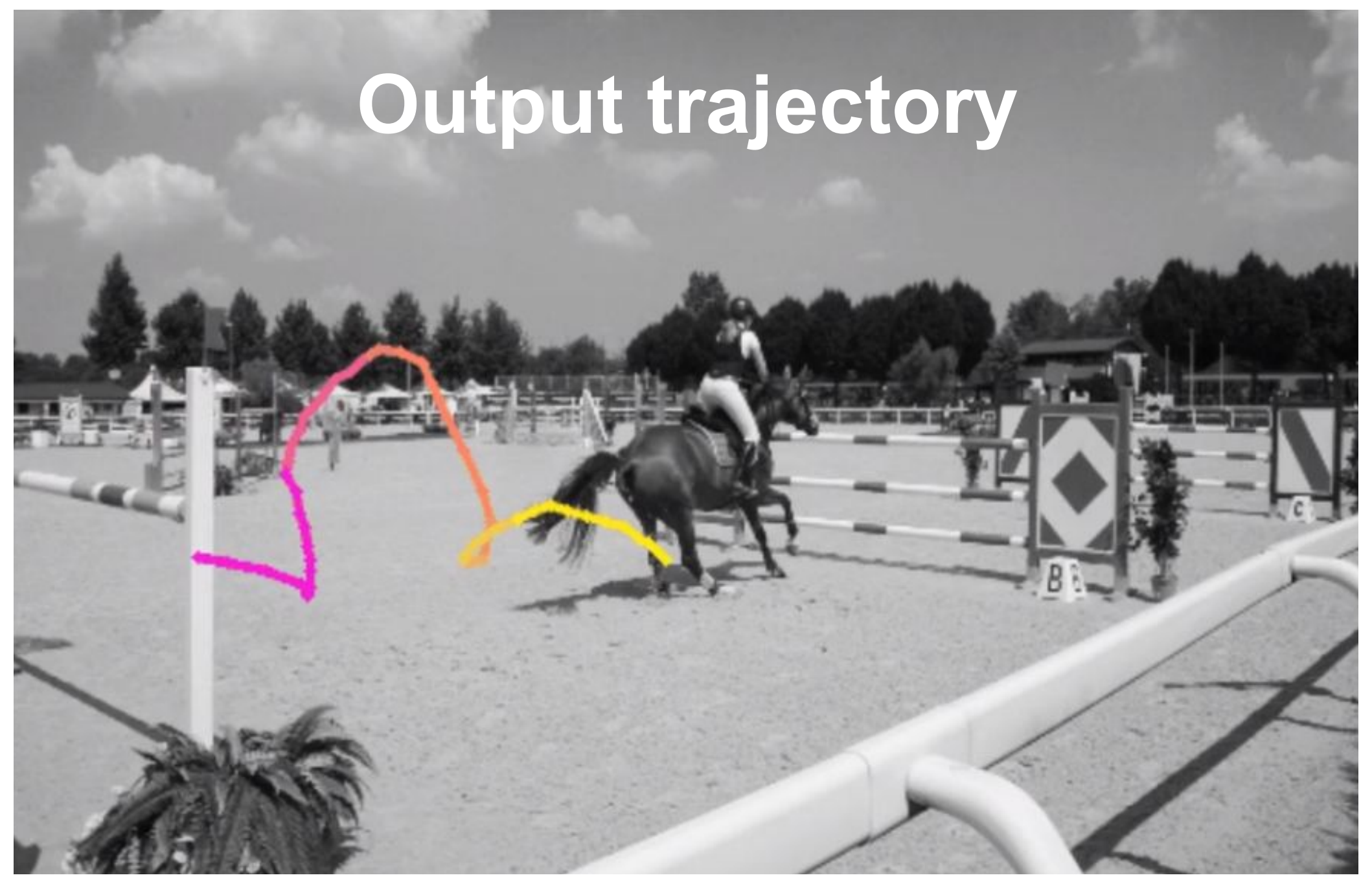
Tracking Through Occlusions Using Point Trajectories

Adam W. Harley, Zhaoyuan Fang, Katerina Fragkiadaki

https://particle-video-revisited.github.io
https://github.com/aharley/pips

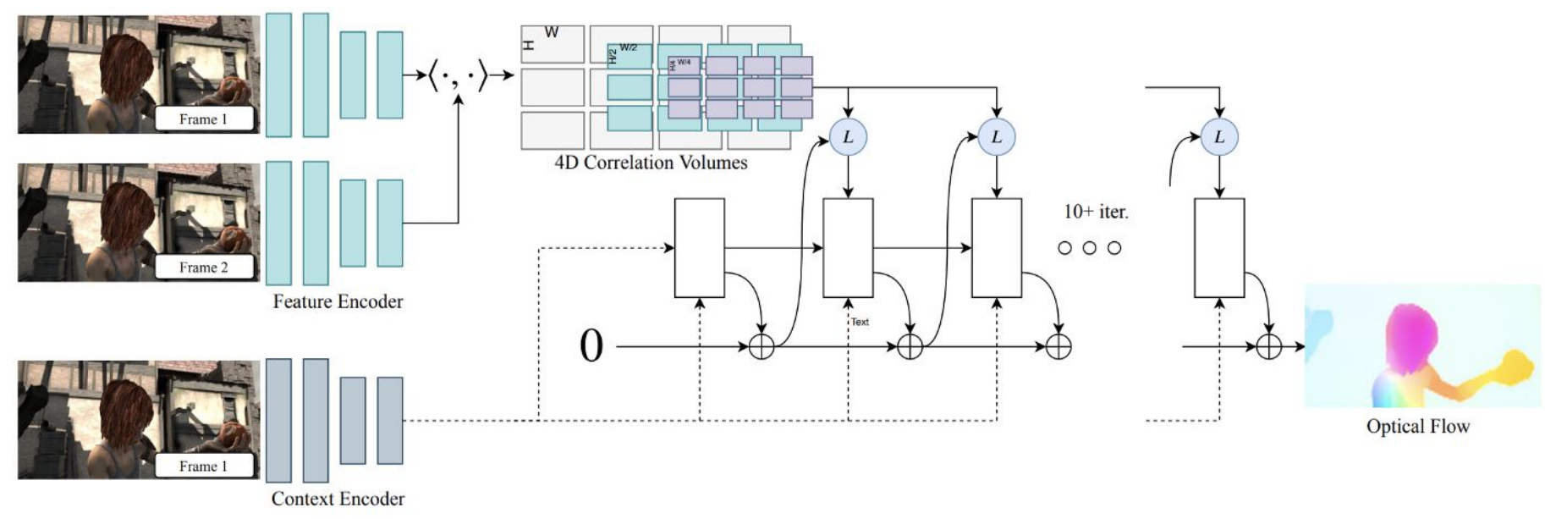
ECCV TEL AVIV 2022

Goal: Given a target pixel specified on the first frame, track that pixel across the whole video.



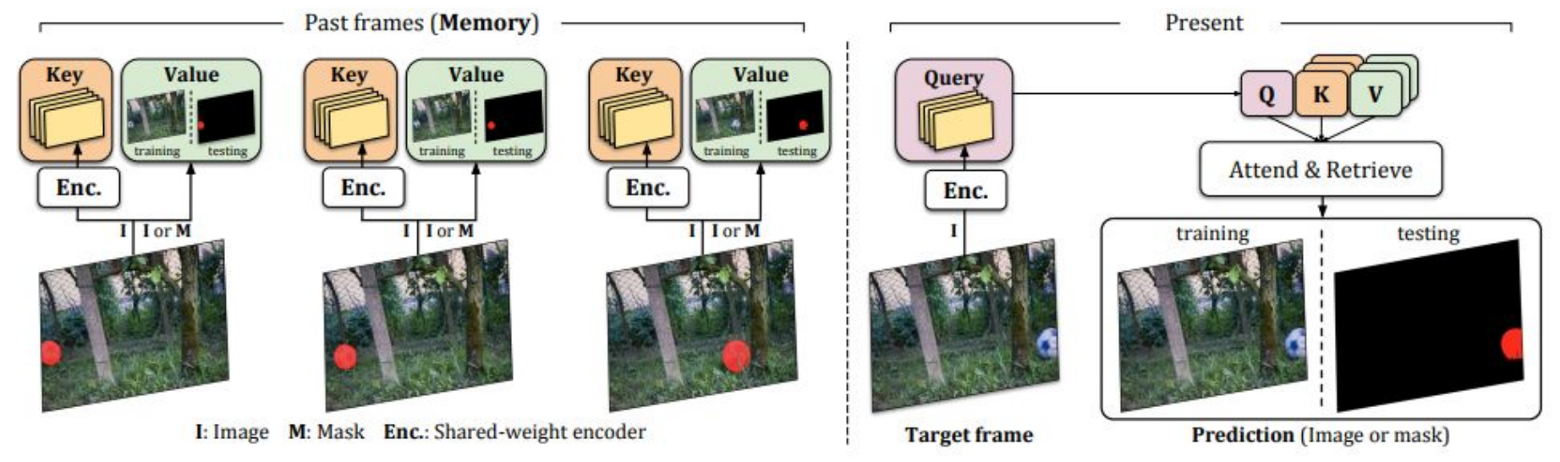
Related: Optical flow and feature matching are general-purpose, but have difficulty with occlusions.

Optical flow provides fine-grained motion, but with a narrow timespan.



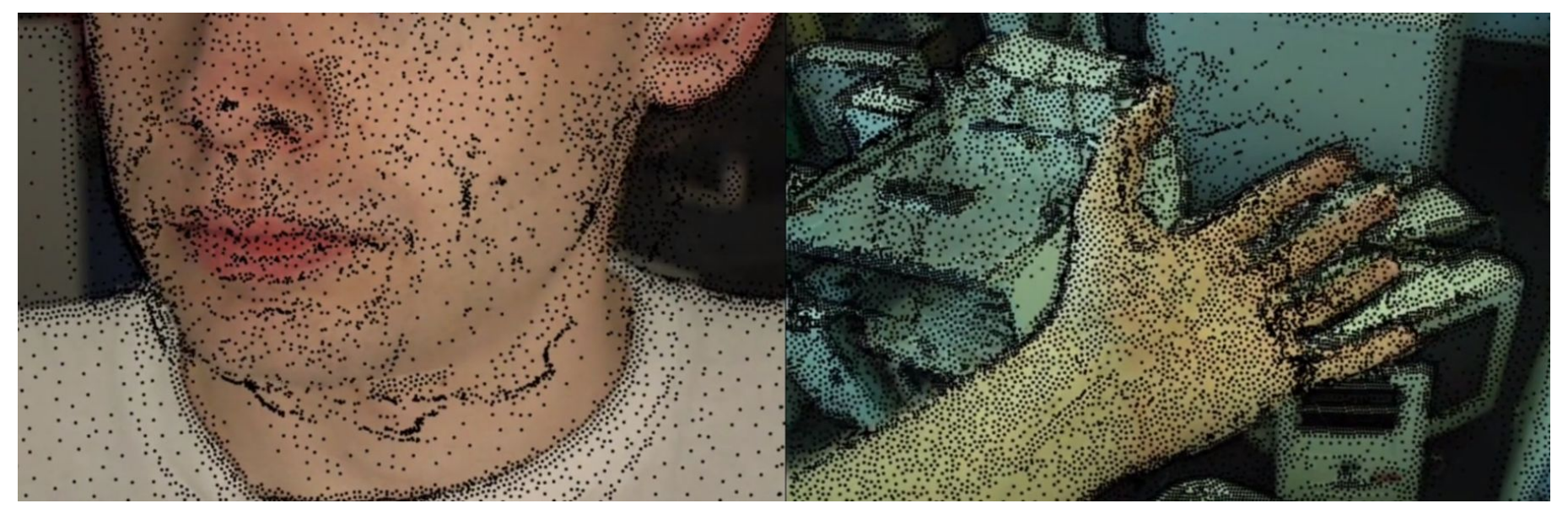
Teed & Deng. RAFT: Recurrent All Pairs Field Transforms for Optical Flow ECCV 2020

Feature matching provides long-range correspondence, but without a temporal prior.



Lai et al. MAST: A Memory-Augmented Self-Supervised Tracker. CVPR 2020.

Sand & Teller (2008) suggest to treat pixels like “particles”. Let’s revisit this!

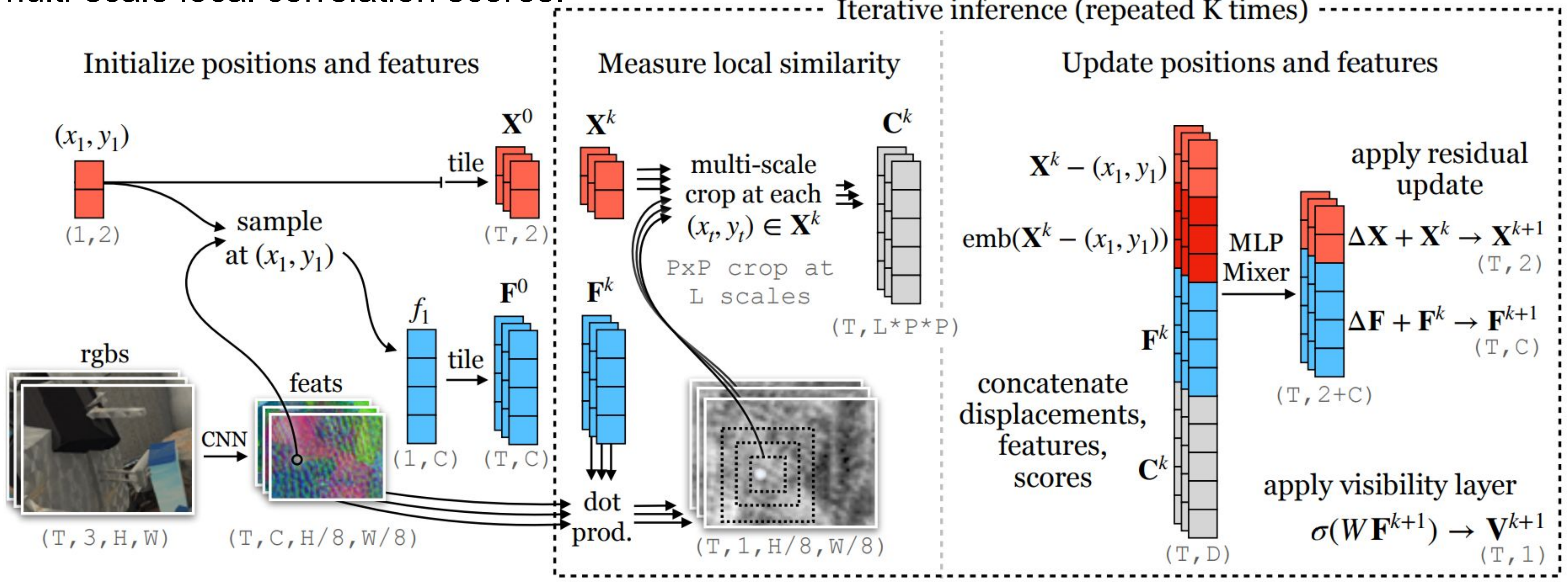


Sand & Teller. Particle Video: Long-Range Motion Estimation Using Point Trajectories. IJCV 2008.

Modern trackers for pedestrians and cars use strong temporal priors to help track through occlusions. Can we bring this power to pixel tracking?

Rajasegaran et al., Tracking People by Predicting 3D Appearance, Location & Pose. CVPR 2022.

Persistent Independent Particles (PIPs): Every particle is tracked independently. We initialize a zero-velocity trajectory for each particle, and iteratively refine it, inspecting multi-frame multi-scale local correlation scores.



Iterative inference (repeated K times)

Initialize positions and features

(x_1, y_1) (1,2)

tile $\mathbf{X}^0$ (T,2)

sample at (x_1, y_1)

rgbs (T,3,H,W)

CNN

feats (T,C,H/8,W/8)

f_1 (1,C)

tile $\mathbf{F}^0$ (T,C)

Measure local similarity

$\mathbf{X}^k$

multi-scale crop at each $(x_t, y_t) \in \mathbf{X}^k$

PxP crop at L scales

$\mathbf{C}^k$ (T,L*P*P)

$\mathbf{F}^k$

dot prod.

(T,1,H/8,W/8)

Update positions and features

$\mathbf{X}^k - (x_1, y_1)$

$\text{emb}(\mathbf{X}^k - (x_1, y_1))$

$\mathbf{F}^k$

$\mathbf{C}^k$

concatenate displacements, features, scores

(T,D)

MLP Mixer

(T,2+C)

apply residual update

$\Delta\mathbf{X} + \mathbf{X}^k \rightarrow \mathbf{X}^{k+1}$ (T,2)

$\Delta\mathbf{F} + \mathbf{F}^k \rightarrow \mathbf{F}^{k+1}$ (T,C)

apply visibility layer

$\sigma(W\mathbf{F}^{k+1}) \rightarrow \mathbf{V}^{k+1}$ (T,1)

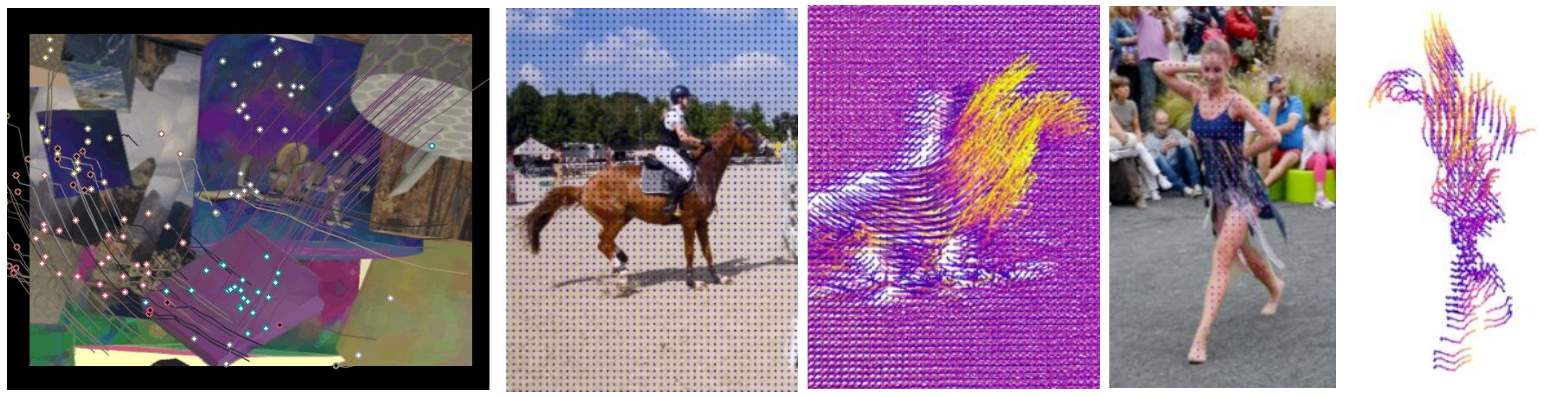
PIPs vs. RAFT

8-frame temporal window instead of 2-frame.

Deep temporal prior via MLP-Mixer, instead of shallow one by convGRU.

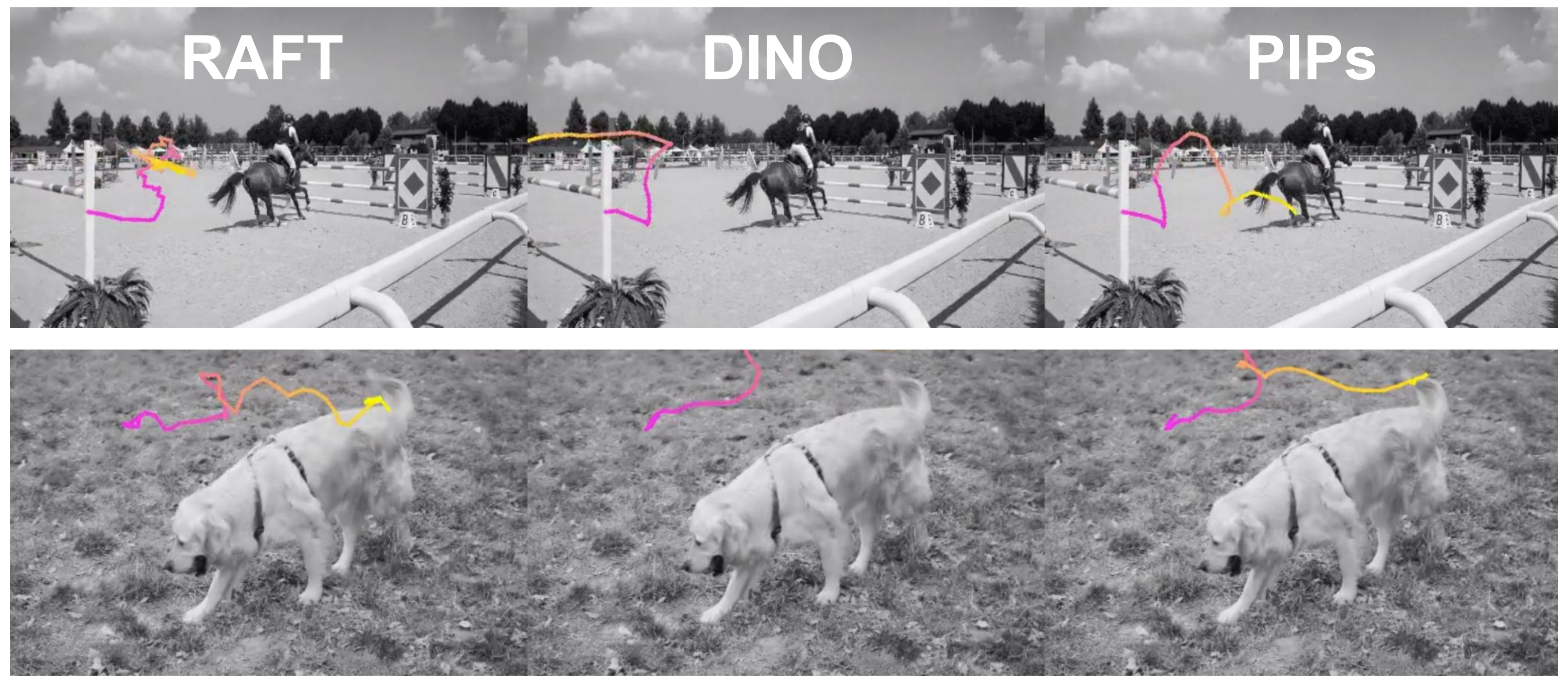
200ms for 8 frames for 256 particles, instead of 200ms/frame.

Results: Outperforms optical flow and feature matching. Despite training on synthetic data, generalizes to YouTube.

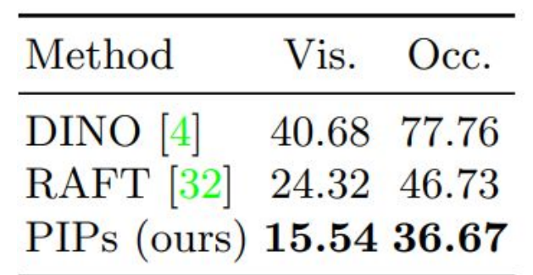


Training data consists of “Flying Things” with multi-frame occlusions. All experiments evaluate the same model.

At test time, we chain our 8-frame trajectories to track for arbitrary timespans.

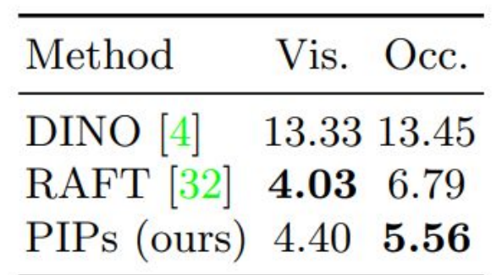


FlyingThings



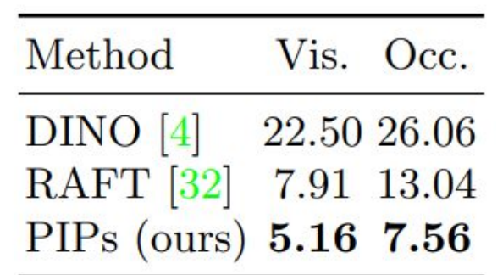
Method	Vis.	Occ.
DINO [4]	40.68	77.76
RAFT [32]	24.32	46.73
PIPs (ours)	**15.54**	**36.67**

KITTI



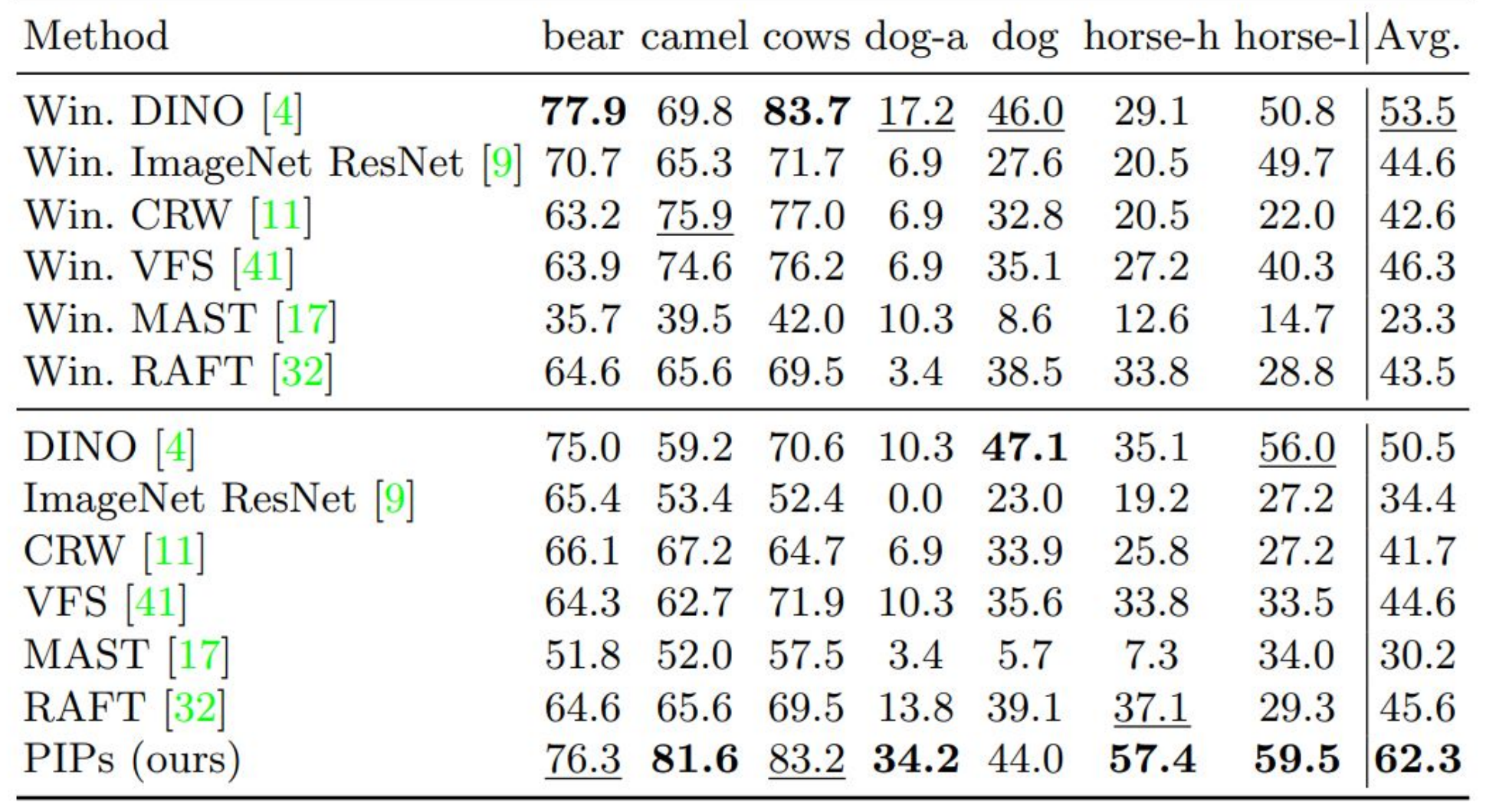
Method	Vis.	Occ.
DINO [4]	13.33	13.45
RAFT [32]	**4.03**	6.79
PIPs (ours)	4.40	**5.56**

CroHD (ATE)



Method	Vis.	Occ.
DINO [4]	22.50	26.06
RAFT [32]	7.91	13.04
PIPs (ours)	**5.16**	**7.56**

BADJA (PCK-T)



Method	bear	camel	cows	dog-a	dog	horse-h	horse-l	Avg.
Win. DINO [4]	**77.9**	69.8	**83.7**	17.2	46.0	29.1	50.8	53.5
Win. ImageNet ResNet [9]	70.7	65.3	71.7	6.9	27.6	20.5	49.7	44.6
Win. CRW [11]	63.2	75.9	77.0	6.9	32.8	20.5	22.0	42.6
Win. VFS [41]	63.9	74.6	76.2	6.9	35.1	27.2	40.3	46.3
Win. MAST [17]	35.7	39.5	42.0	10.3	8.6	12.6	14.7	23.3
Win. RAFT [32]	64.6	65.6	69.5	3.4	38.5	33.8	28.8	43.5
DINO [4]	75.0	59.2	70.6	10.3	**47.1**	35.1	56.0	50.5
ImageNet ResNet [9]	65.4	53.4	52.4	0.0	23.0	19.2	27.2	34.4
CRW [11]	66.1	67.2	64.7	6.9	33.9	25.8	27.2	41.7
VFS [41]	64.3	62.7	71.9	10.3	35.6	33.8	33.5	44.6
MAST [17]	51.8	52.0	57.5	3.4	5.7	7.3	34.0	30.2
RAFT [32]	64.6	65.6	69.5	13.8	39.1	37.1	29.3	45.6
PIPs (ours)	76.3	**81.6**	83.2	**34.2**	44.0	**57.4**	**59.5**	**62.3**

We outperform prior work, especially in long-range tracking, mostly due to handling occlusions with multi-frame inference.

Future work: multi-trajectory inference, recurrence.