JUNE 18-22, 2023
CVPR
VANCOUVER, CANADA

MotionTrack: Learning Robust Short-term and Long-term Motions for Multi-Object Tracking

Zheng Qin[1†] Sanping Zhou[1†] Le Wang[1*] Jinghai Duan[2] Gang Hua[3] Wei Tang[4]
[1]National Key Laboratory of Human-Machine Hybrid Augmented Intelligence,
National Engineering Research Center for Visual Information and Applications,
Institute of Artificial Intelligence and Robotics, Xi'an Jiaotong University
[2]School of Software Engineering, Xi'an Jiaotong University
[3]Wormpex AI Research [4]University of Illinois at Chicago

Quick preview

MotionTrack: Learning Robust Short-term and Long-term Motions for Multi-Object Tracking

Zheng Qin#, Sanping Zhou#, Le Wang*, Jinghai Duan, Gang Hua, Wei Tang
Institute of Artificial Intelligence and Robotics, Xi'an Jiaotong University

JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

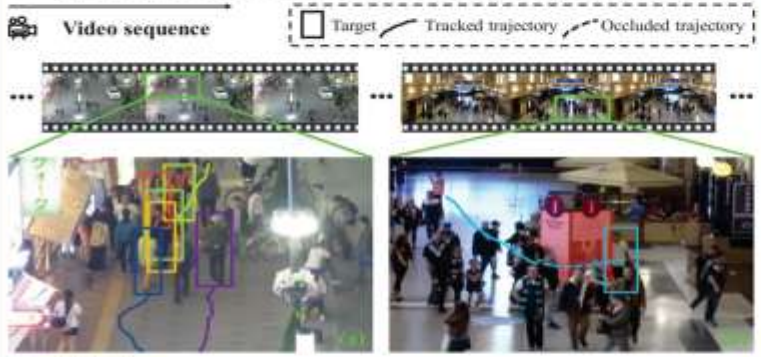
Introduction

Background

- Multi-Object Tracking (MOT) : jointly locate targets through bounding boxes and recognize their identities throughout a whole video.

Motivation

- (a) ***Dense crowds:*** Pedestrians do not move independently in this situation.
- (b) ***Extreme occlusion:*** Pedestrians are easily occluded by fixed facilities.

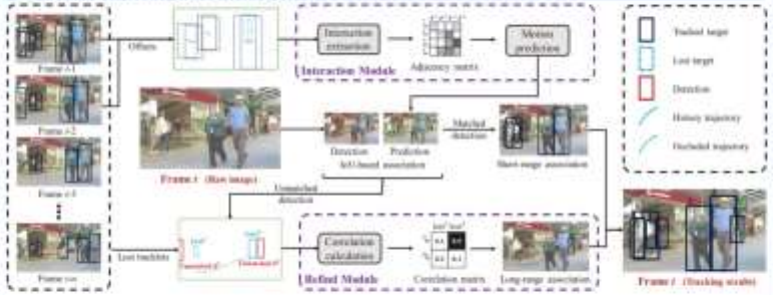


Contribution

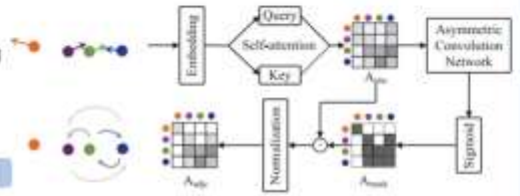
- We propose a simple yet effective multi-object tracker, MotionTrack, to address the short-range and long-range association problems.
- We design a novel Interaction Module to model the interaction between targets, which can handle complex motions in dense crowds.
- We design a novel Refind Module to learn discriminative motion patterns, which can re-identify the lost tracklets with current detections.

Methodology

Overview of MotionTrack

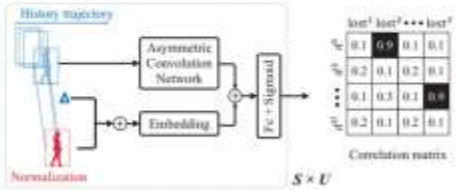


- ***Step 1: Short-range association***. Modeling the inter-tracklet interaction to obtain more accurate predictions and the short-range tracking results.
- ***Step 2: Long-range association***. Re-identifying lost tracklets based on the history trajectory and unmatched detections and then compensating the trajectory during occlusion.



Interaction Module

- To obtain more accurate tracklets, we capture the directed interactions between tracklets and then use them to estimate the offsets between two consecutive frames.



Refind Module

- To refind the lost tracklet, we first identify its matched detection and then refine the occluded trajectory.

Results

Comparison with SOTA

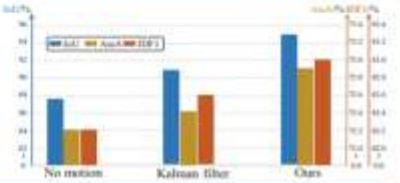
Tracker	Venue	IDF1 ↑	MOTA ↑	HOTA ↑	AssA ↑	DetA ↑	FP ↓	FN ↓	IDs ↓	Frag ↓
ReMOT [55]	IVC'21	72.0	77.0	59.7	57.1	62.8	53034	93612	2853	5304
QuasiDense [30]	CVPR'21	66.3	68.7	53.9	52.7	55.6	26589	146643	3378	8091
SOTMOT [63]	CVPR'21	71.9	71.0	-	-	-	39537	118983	5184	-
SiamMOT [37]	CVPR'21	72.3	76.3	-	-	-	-	-	-	-
CorrTracker [42]	CVPR'21	73.6	76.5	60.7	58.5	62.9	29808	99510	3369	6063
PermaTrackPr [17]	ICCV'21	68.9	73.8	55.5	53.1	58.5	28998	115104	3699	6132
PureMOT [35]	ICCV'21	72.5	73.7	59.3	58.0	60.9	27937	112477	3303	6477
CSTrack [24]	TIP'22	72.6	74.9	59.3	57.9	61.1	23647	114303	3567	7668
RelationTrack [49]	TMM'22	74.7	73.8	61.0	61.5	60.6	27999	118623	1374	2166
TrackFormer [28]	CVPR'22	68.0	74.1	-	-	-	34602	108777	2829	-
MeMOT [3]	CVPR'22	69.0	72.5	56.9	55.2	-	37221	115248	2724	-
MTrack [46]	CVPR'22	73.5	72.1	-	-	-	53501	101844	2028	-
MOTR [58]	ECCV'22	68.6	73.4	57.8	55.7	60.3	-	-	2439	-
ByteTrack [53]	ECCV'22	77.3	80.3	63.1	62.0	64.5	25491	83721	2196	2277
P3AFormer+W&R) [54]	IJCAI'22	78.1	**81.2**	-	-	-	**17281**	86861	1893	-
MotionTrack(ours)	-	**80.1**	81.1	**65.1**	**65.3**	**65.4**	23802	**81660**	**1140**	**1603**

Tracker	Venue	IDF1 ↑	MOTA ↑	HOTA ↑	AssA ↑	DetA ↑	FP ↓	FN ↓	IDs ↓	Frag ↓
PureMOT [35]	IJCV'21	67.3	61.8	54.6	54.7	54.7	103440	88901	8253	7874
CorrTracker [42]	CVPR'21	69.1	65.2	-	-	-	79429	95855	3183	-
SiamMOT [37]	CVPR'21	69.1	67.1	-	-	-	-	-	-	-
SOTMOT [63]	CVPR'21	71.4	68.6	57.4	57.3	57.7	57064	101154	4209	5368
CSTrack [24]	TIP'22	68.6	66.6	54.0	50.0	54.2	**25404**	144358	3196	7632
RelationTrack [49]	TMM'22	70.5	67.2	56.4	56.8	61134	104497	4243	8236	
MeMOT [3]	CVPR'22	66.1	63.7	54.1	55.0	-	47502	137982	1938	-
MTrack [46]	CVPR'22	69.2	63.5	-	-	-	96123	86064	6031	-
ByteTrack [53]	ECCV'22	75.2	77.8	61.3	59.6	63.4	26249	87594	1223	1460
P3AFormer+W&R) [54]	IJCAI'22	76.4	**78.1**	-	-	-	25613	86510	1332	-
MotionTrack(ours)	-	**76.5**	78.0	**62.8**	**61.8**	**64.0**	28629	**84152**	**1165**	**1321**

Comparison for handling occlusions

Setting	#	IDF1 ↑	MOTA ↑	HOTA ↑	AssA ↑	DetA ↑
IoU-based	30	80.9	79.8	69.0	70.6	68.1
	120	80.1	77.6	68.4	70.5	66.9
	Δ	-0.8	-2.2	-0.6	-0.1	-1.2
ReID-based	30	77.2	77.0	66.4	67.1	66.3
	120	70.4	67.5	60.6	60.3	61.6
	Δ	-6.8	-9.5	-5.8	-6.8	-4.7
Ours	30	82.6	80.4	70.2	72.4	68.7
	120	83.3	80.7	70.7	73.2	68.8
	Δ	+0.7	+0.3	+0.5	+0.8	+0.1

Results of Motion Models.

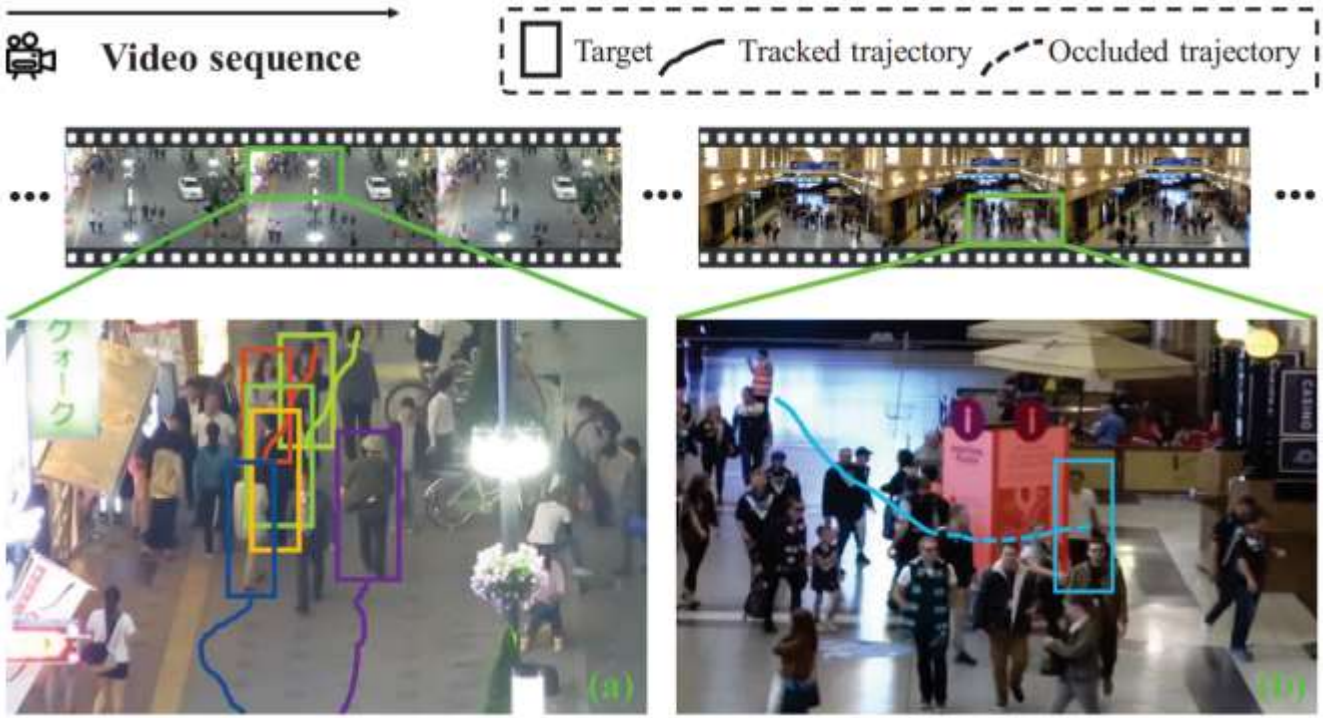


Visualization of Directed Interaction

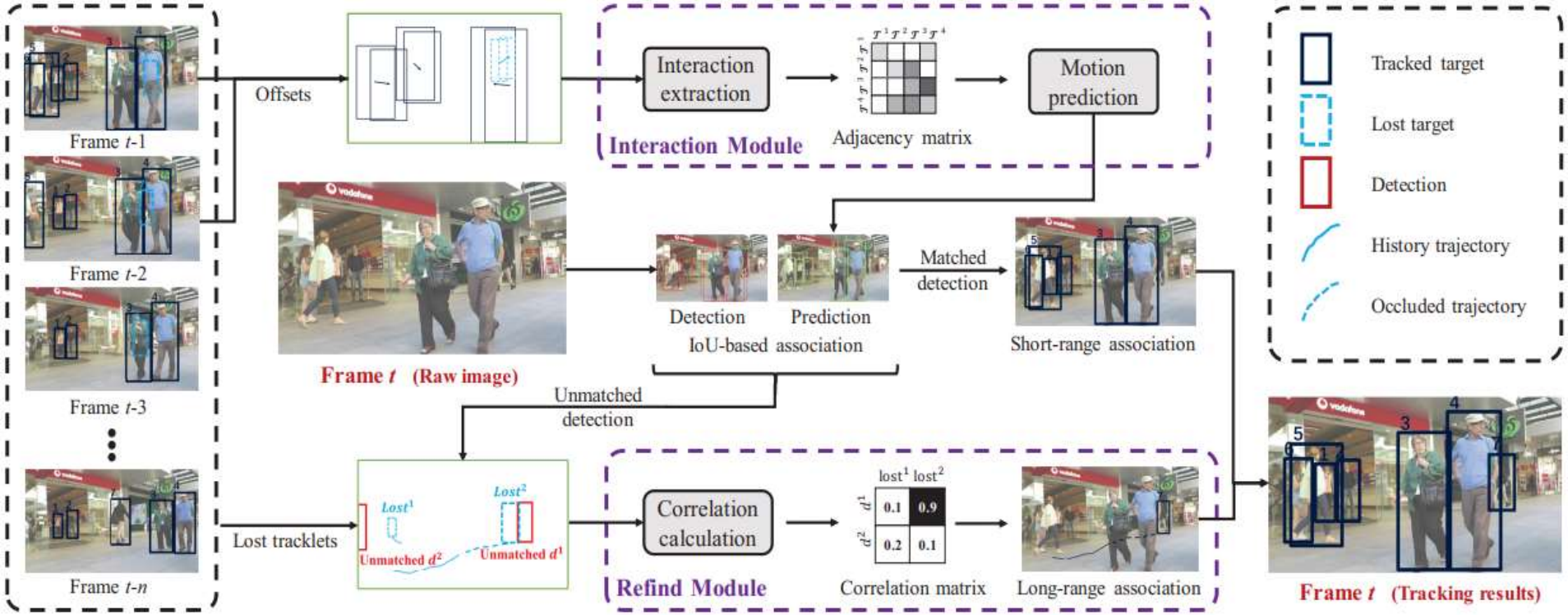
Visualization of Refinding Targets

Motivation

- (a) ***Dense crowds：*** Pedestrians do not move independently in this situation.
- (b) ***Extreme occlusion：*** Pedestrians are easily occluded by fixed facilities.



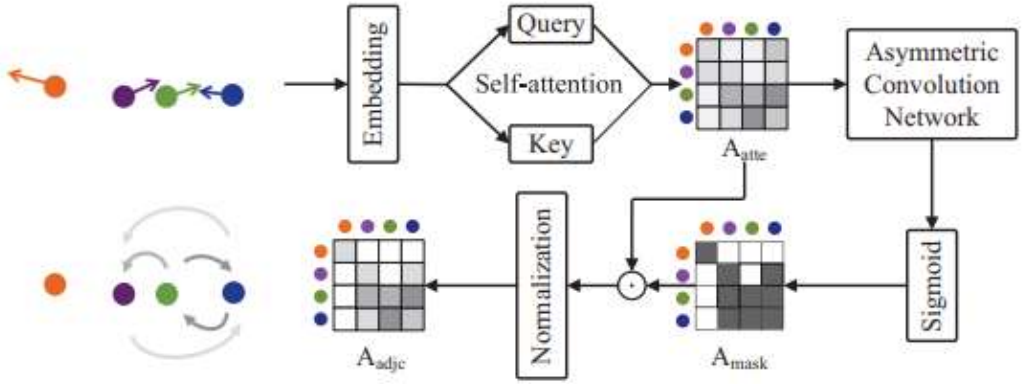
Overview of MotionTrack



- ***Step 1: Short-range association***. Modeling the inter-tracklet interaction to obtain more accurate predictions and the short-range tracking results.

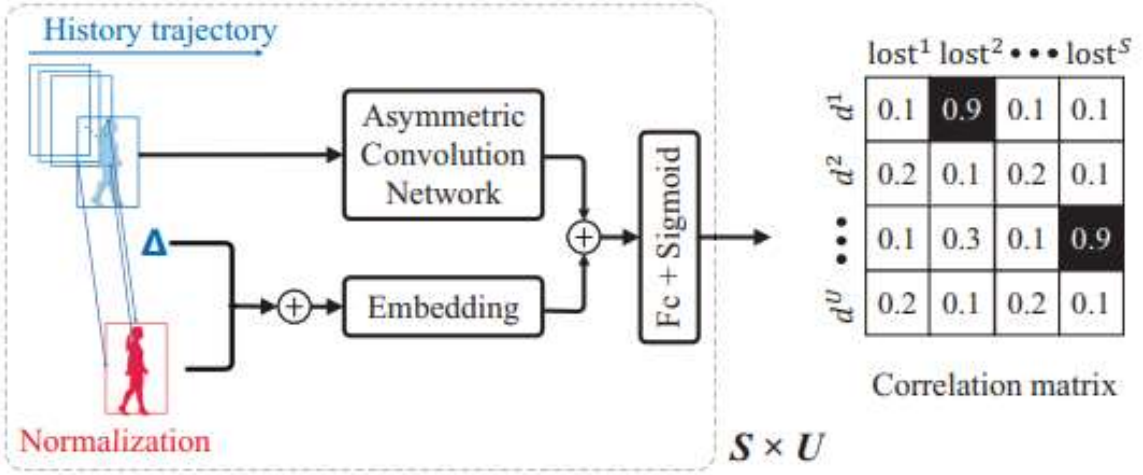
- ***Step 2: Long-range association***. Re-identifying lost tracklets based on the history trajectory and unmatched detections and then compensating the trajectory during occlusion.

Model architecture



Interaction Module

- To obtain more accurate tracklets, we capture the directed interactions between tracklets and then use them to estimate the offsets between two consecutive frames.



Refind Module

- To refind the lost tracklet, we first identify its matched detection and then refine the occluded trajectory.

Comparison with SOTA

MOT17

Tracker	Venue	IDF1 ↑	MOTA ↑	HOTA ↑	AssA ↑	DetA ↑	FP ↓	FN ↓	IDs ↓	Frag ↓
ReMOT [45]	IVC'21	72.0	77.0	59.7	57.1	62.8	33204	93612	2853	5304
QuasiDense [30]	CVPR'21	66.3	68.7	53.9	52.7	55.6	26589	146643	3378	8091
SOTMOT [55]	CVPR'21	71.9	71.0	-	-	-	39537	118983	5184	-
SiamMOT [23]	CVPR'21	72.3	76.3	-	-	-	-	-	-	-
CorrTracker [42]	CVPR'21	73.6	76.5	60.7	58.5	62.9	29808	99510	3369	6063
PermaTrackPr [37]	ICCV'21	68.9	73.8	55.5	53.1	58.5	28998	115104	3699	6132
FairMOT [53]	IJCV'21	72.3	73.7	59.3	58.0	60.9	27507	117477	3303	8073
CSTrack [24]	TIP'22	72.6	74.9	59.3	57.9	61.1	23847	114303	3567	7668
RelationTrack [48]	TMM'22	74.7	73.8	61.0	61.5	60.6	27999	118623	1374	2166
TrackFormer [26]	CVPR'22	68.0	74.1	-	-	-	34602	108777	2829	-
MeMOT [7]	CVPR'22	69.0	72.5	56.9	55.2	-	37221	115248	2724	-
MTrack [46]	CVPR'22	73.5	72.1	-	-	-	53361	101844	2028	-
MOTR [50]	ECCV'22	68.6	73.4	57.8	55.7	60.3	-	-	2439	-
ByteTrack [52]	ECCV'22	77.3	80.3	63.1	62.0	64.5	25491	83721	2196	2277
P3AFormer(+W&B) [54]	ECCV'22	78.1	**81.2**	-	-	-	**17281**	86861	1893	-
MotionTrack(ours)	-	**80.1**	81.1	**65.1**	**65.1**	**65.4**	23802	**81660**	**1140**	**1605**

MOT20

Tracker	Venue	IDF1 ↑	MOTA ↑	HOTA ↑	AssA ↑	DetA ↑	FP ↓	FN ↓	IDs ↓	Frag ↓
FairMOT [53]	IJCV'21	67.3	61.8	54.6	54.7	54.7	103440	88901	5243	7874
CorrTracker [42]	CVPR'21	69.1	65.2	-	-	-	79429	95855	5183	-
SiamMOT [23]	CVPR'21	69.1	67.1	-	-	-	-	-	-	-
SOTMOT [55]	CVPR'21	71.4	68.6	57.4	57.3	57.7	57064	101154	4209	7568
CSTrack [24]	TIP'22	68.6	66.6	54.0	50.0	54.2	**25404**	144358	3196	7632
RelationTrack [48]	TMM'22	70.5	67.2	56.5	56.4	56.8	61134	104597	4243	8236
MeMOT [7]	CVPR'22	66.1	63.7	54.1	55.0	-	47882	137982	1938	-
MTrack [46]	CVPR'22	69.2	63.5	-	-	-	96123	86964	6031	-
ByteTrack [52]	ECCV'22	75.2	77.8	61.3	59.6	63.4	26249	87594	1223	1460
P3AFormer(+W&B) [54]	ECCV'22	76.4	**78.1**	-	-	-	25413	86510	1332	-
MotionTrack(ours)	-	**76.5**	78.0	**62.8**	**61.8**	**64.0**	28629	**84152**	**1165**	**1321**

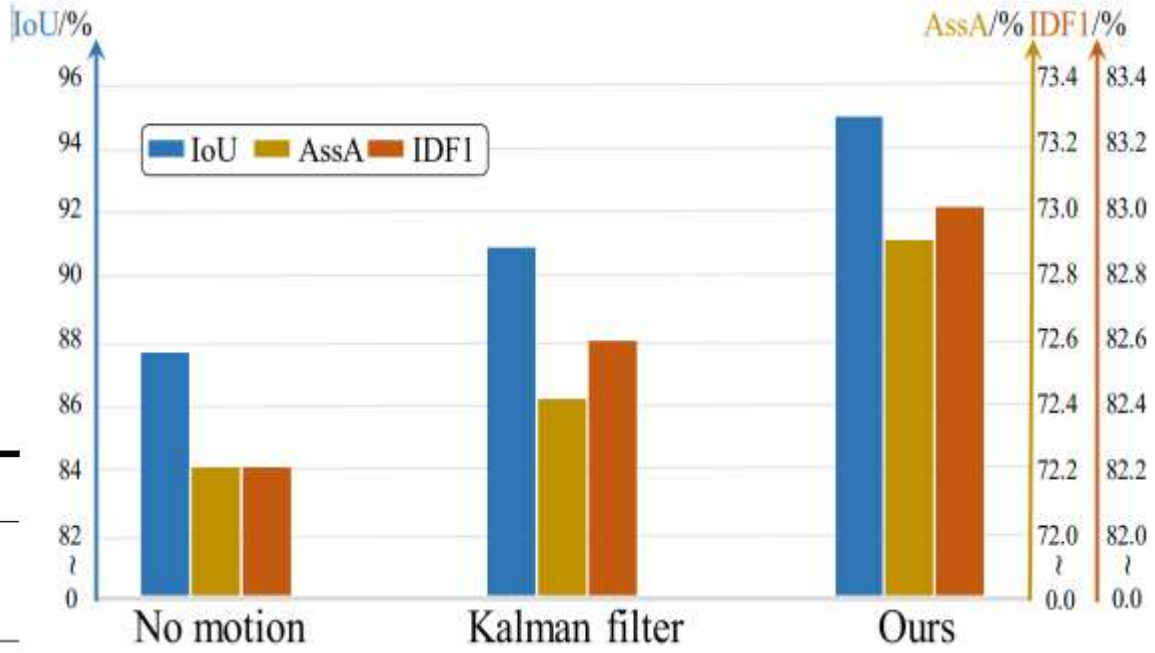
Ablation Study

Setting	IDF1 ↑	MOTA ↑	HOTA ↑	AssA ↑	DetA ↑	IDs ↓
Baseline	82.6	80.4	70.2	72.4	68.7	402
Baseline+I	83.0	80.5	70.5	72.9	68.8	390
Baseline+I+R	83.7	80.7	70.8	73.5	68.9	378

Comparison for handling occlusions

Setting	#	IDF1 ↑	MOTA ↑	HOTA ↑	AssA ↑	DetA ↑
IoU-based	30	80.9	79.8	69.0	70.6	68.1
	120	80.1	77.6	68.4	70.5	66.9
	Δ	**-0.8**	**-2.2**	**-0.6**	**-0.1**	**-1.2**
ReID-based	30	77.2	77.0	66.4	67.1	66.3
	120	70.4	67.5	60.6	60.3	61.6
	Δ	**-6.8**	**-9.5**	**-5.8**	**-6.8**	**-4.7**
Ours	30	82.6	80.4	70.2	72.4	68.7
	120	83.3	80.7	70.7	73.2	68.8
	Δ	**+0.7**	**+0.3**	**+0.5**	**+0.8**	**+0.1**

Results of Motion Model

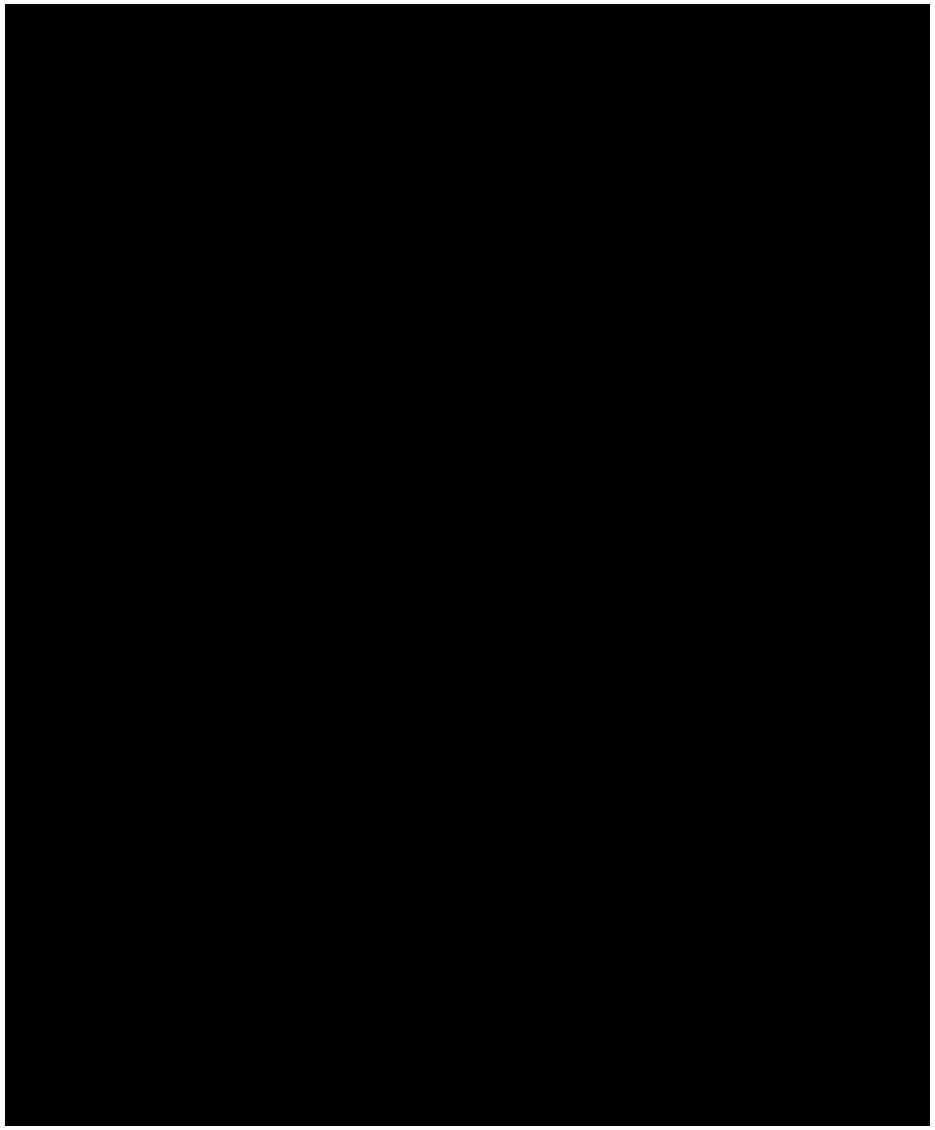


Evaluation for crowds and occlusions

Setting	⩾ 20	⩾ 40	⩾ 60	⩾ 80	⩾ 100
Baseline	77.2	73.6	75.2	73.3	71.5
Ours	78.3	75.1	76.8	74.1	72.4
Improvement	**+1.1**	**+1.5**	**+1.6**	**+0.8**	**+0.9**

video demos

Different colored boxes represent different identities and red bolded boxes represent the location during occlusion after the long-term refind. For the cases in our demo video (red bolded), almost all other methods fail to track them (tracklets before and after crowds or occlusion have different identities



西安交通大學
XI'AN JIAOTONG UNIVERSITY

IAIR Est. 1986
Institute of Artificial Intelligence and Robotics, XJTU

JUNE 18-22, 2023
CVPR
VANCOUVER, CANADA

Thank you for listening !

qinzheng@stu.xjtu.edu.cn