

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

 Zheng Qin<sup>1†</sup> Sanping Zhou<sup>1†</sup> Le Wang<sup>1\*</sup> Jinghai Duan<sup>2</sup> Gang Hua<sup>3</sup> Wei Tang<sup>4</sup>
<sup>1</sup>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
<sup>2</sup>School of Software Engineering, Xi'an Jiaotong University
<sup>3</sup>Wormpex AI Research <sup>4</sup>University of Illinois at Chicago





Institute of Artificial Intelligence and Robotics, XJTU



## Quick preview

### 新考え通大学 MotionTrack: Learning Robust Short-term and Long-term Motions for Multi-Object Tracking

### AIR 1995

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



### 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.



- 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.



#### Comparison with SOTA

Tracker	Versus	10011	SHOTA	HOTA †	AssA 1	DotA T	PP	#1N 2	HDs 1	Frag
BeMOT LOT	IVC.21	72.0	-71.8	- 38CT	. 97.1	42.8	93204	93612	2883	3,954
Quantificence [191]	CVPIC21	105.2	48.7	55.9	\$2.7	53.4	26599	146643	3329	20191
SOTMOT [12]	CVPR 21	78.30	71.0		1.0413		39533	11109883	5188	- 2040
Samb9OT [11]	CVPR:21	72.0	36.3		-					
CorrTracker [-11]	CVPR:21	73.6	36.5	463.7	28.5	\$2.9	29808	99510	3369	6053
Perrow Disck Pr [ 11]	3ECV231	65.9	75.8	55.5	33.5	58.5	209938	115104	3609	6132
PartNETTING	10CV21	72.5	78.7	358.3	38.0	60.9	27507	112422	3303	18677
CKTeset [24]	1101122	22.0	24.0	199.3	87.9	44.1	22647	114301	3567	TWebs
Relation/Timels [201]	TMMCD	74.2	23.8	443.00	41.5	60.6	211004	118621	1574	2164
Tites & Province 1 (1)	CVPR'22	108.17	24.6	10.00	11.64-1	1.0	34992	1082777	29(29)	12.22
MONDELL	CVP8*22	101.01	32.4	56.9	55.2		37221	115248	2724	- 22
MTrick [107]	CVPR:23	73.5	72.8		1.040	Gardel	5336i	101844	2028	
MOTH LULE	ECCV*22	68.6	73.4	87.8	35.7	60.3	100210	10000	24.99	1.4
Hote Truck 11.00	BCCV 22	22.8	40.1	848.4	6,7.0	44.5	25294	10.017211	21445	227
PAAFonmert-WARD [11]	BCCV'22	28.1	81.2				17281	36441	1104.8	
Metion(Track(met)		105.1	85.1	65.1	65.3	65.4	13000	31660	11.00	168
Tracker	Viene	IDP1 1	ARTES 1	HOTA 1	Ana A T	Det A. 7	12	251	ilbs /	Free
EnarMERT 11.11	LBC VCD	62.3	01.8	34.6	84.7	44.7	023440	888804	87.29	10.1
Conditional and the later	CVPR 24	69.1	65.2				794298	95655	3181	
Sincibility (1913)	CVPRIDE	69.1	67.1	13.545-1	A THE PARTY	200	24	5.255		1.00
SOUTHOR LINE	27VPR121	26.4	03.0	57.4	37.3	97.7	27004	101154	42199	7566
CS Drack 1141	TU* 22	69.0	00.0	54.0	1444.00	93.2	26,044	1.54150	31100	76.7
Relation Track Light	TMMC22	718.75	62.0	30.2	10.4	86.H	marine.	104297	4257	8214
MUNICEPT	CVPR <sup>1</sup> 22	00.1	63.7	54.1	55.0	0.000	479882	137992	1908	12
S.C.Telerk ( 191)	CVPR'32	100.7	015	1.1			145121	SECOLS.	6051	1.22
and a second by the second sec	and the second second	1000		1000	1000 A	WY	10,009	STAME.	1100	1.144
Boye Truck [37]	COLUMN 22	7.5	1 C C C C C C C C C C C C C C C C C C C							
Byie Truck [32]	BCCV32	75.0	28.1		1000		29413	68/510	1832	

16.9

61.6

48.3

Results

#### **Results of Motion Models** Comparison for handling occlusions





#### 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**

Tracker	Venue	IDF1↑	<b>MOTA</b> ↑	HOTA ↑	AssA ↑	<b>DetA</b> ↑	<b>FP</b> ↓	FN↓	IDs ↓	<b>Frag</b> ↓
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	-	-	87 <b>4</b> 8	39537	118983	5184	-
SiamMOT [23]	CVPR'21	72.3	76.3	-	-	<del></del>	-	-	55	-
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	2	1.72	0.5	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	-	-	73	17281	86861	1893	-
MotionTrack(ours)	-	80.1	81.1	65.1	65.1	65.4	23802	81660	1140	1605
Tracker	Venue	IDF1 ↑	MOTA ↑	HOTA ↑	AssA ↑	DetA ↑	<b>FP</b> ↓	FN↓	<b>IDs</b> ↓	<b>Frag</b> ↓
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	-		123	_	-		_
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	2
ByteTrack [52]	ECCV'22	75.2	77.8	61.3	59.6	63.4	26249	87594	1223	1460
$P_3 \wedge Former(\pm W \& B)$ [54]	ECCV'22	76.4	78 1	01.5	57.0	05.4	25413	86510	1332	1400
MotionTrack(ours)	LCC V 22	76.5	78.0	62.8	61.8	64.0	28620	84152	1165	1321
would mack (ours)	-	10.5	70.0	02.0	01.0	04.0	20029	04152	1105	1321

**MOT17** 

MOT20

## **Ablation Study**

Setting	IDF1 ↑	MOTA ↑	HOTA $\uparrow$	$\mathbf{AssA}\uparrow$	<b>DetA</b> ↑	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	#	<b>IDF1</b> ↑	<b>MOTA</b> ↑	HOTA $\uparrow$	AssA ↑	<b>DetA</b> ↑
	30	80.9	79.8	69.0	70.6	68.1
IoU-based	120	80.1	77.6	<u>68.4</u>	70.5	66.9
	Δ	-0.8	-2.2	-0.6	-0.1	-1.2
	30	77.2	77.0	66.4	67.1	66.3
ReID-based	120	70.4	67.5	60.6	60.3	61.6
	$\Delta$	-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
	$\Delta$	+0.7	+0.3	+0.5	+0.8	+0.1





# 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







Institute of Artificial Intelligence and Robotics, XJTU



## Thank you for listening !

qinzheng@stu.xjtu.edu.cn



