Towards Generalizable Multi-Object Tracking

Zheng Qin[1] Le Wang[1*] Sanping Zhou[1] Panpan Fu[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

西安交通大学
XI'AN JIAOTONG UNIVERSITY

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

VIPA

便利蜂 × WORMPEX AI RESEARCH LLC

UIC UNIVERSITY OF ILLINOIS CHICAGO

Quick preview

Towards Generalizable Multi-Object Tracking

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

Introduction

Background and Application Scenario

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

Tactical analysis of sports games

Dance movement analysis

Automatic driving

Contribution

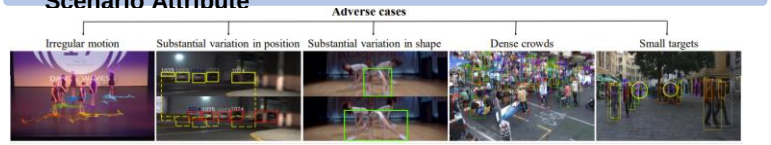
➢ We analyze the factors that hinder the generalizability of existing trackers and concretize them into tracking scenario attributes that can guide the design of trackers.

➢ We propose a "point-wise to instance-wise relation" framework for MOT. It first constructs point-wise relations through the multi-scale 4D correlation volume and then aggregates them into instance-wise associations through a novel "point-part-instance" hierarchy.

➢ Extensive evaluation of the GeneralTrack shows that it achieves the state-of-the-art performance on multiple MOT datasets. In addition, GeneralTrack experimentally demonstrates strong domain generalization capabilities.

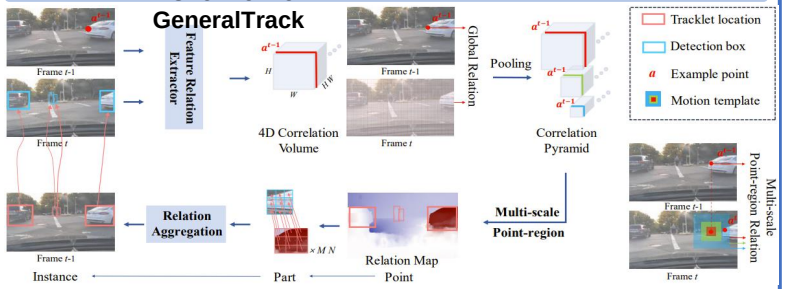
Methodology

Scenario Attribute



Motion Complexity reflects the irregularity and unpredictability of target motion within the scenario; ***Variation Amplitude*** reflects the target's variability, encompassing both shape and position; ***Target Density*** reflects the density of the crowds in the scenario; ***Small Target*** represents the average amount of small targets in the scenario; ***Frame Rate*** is the number of frames captured in

Overview of GeneralTrack



➢ ***Step 1:*** We use ***Feature Relation Extractor*** to construct global dense relations with frame t for each point in frame t − 1 by a 4D correlation volume.

➢ ***Step 2:*** We transform the global relations into ***Multi-scale Point-region Relations***, and form a relation map for frame t − 1.

➢ ***Step 3:*** We perform ***Hierarchical Relational Aggregation*** according to

Multi-scale point-region Relation

Analysis

Comparison with SOTA

	Venue	mHOTA↑	mIDF1↑	mMOTA↑	HOTA↑	IDF1↑	MOTA↑	IDs↓	MT↑	ML↓
validation										
QDTrack [33]	CVPR'21	-	50.8	36.6	-	71.5	63.5	**6262**	9481	3034
Unicorn [51]	ECCV'22	-	54.0	41.2	-	71.3	66.6	10876	10296	2505
MOTR [54]	ECCV'22	-	44.8	32.3	-	65.8	56.2	-	-	-
TETer [24]	ECCV'22	-	53.3	39.1	-	-	-	-	-	-
ByteTrack [57]	ECCV'22	45.3	54.8	45.2	61.3	70.4	**69.1**	9140	9626	3005
MOTRv2 [58]	CVPR'23	-	**56.5**	43.6	-	**72.7**	65.6	-	-	-
GHOST [41]	CVPR'23	45.7	55.6	44.9	61.7	70.9	68.1	-	-	-
GeneralTrack(Ours)		**46.9**	56.2	**46.4**	**63.1**	**72.7**	68.8	8496	**11830**	**2035**
test										
DeepBlueAI	-	-	38.7	31.6	-	56.0	56.9	25186	10296	12266
madamada	-	-	43.0	33.6	-	55.7	59.8	42901	16774	5004
QDTrack [53]	CVPR'21	41.9	52.4	35.7	60.5	72.5	64.6	**10790**	17353	5167
ByteTrack [57]	ECCV'22	-	55.8	**40.1**	-	71.3	**69.6**	15466	18057	5107
GHOST [41]	CVPR'23	46.8	**57.0**	39.5	62.2	72.0	68.9	-	-	-
GeneralTrack(Ours)		**47.9**	56.9	39.9	**63.7**	**73.6**	69.1	14489	**21281**	**3715**

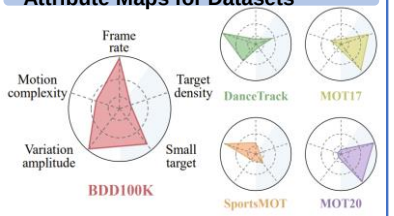
	Venue	HOTA↑	MOTA↑	IDF1↑	AssA↑	DetA↑
GTR [61]	CVPR'22	54.5	67.9	55.8	45.9	64.8
ByteTrack [57]	ECCV'22	64.1	95.9	71.4	52.3	78.5
OC-SORT [6]	CVPR'23	73.7	96.5	74.0	61.5	88.5
MixSort-Byte* [8]	ICCV'23	65.7	96.2	74.1	54.8	78.8
MixSort-OC* [8]	ICCV'23	**74.1**	96.5	74.4	**62.0**	88.5
GeneralTrack(Ours)	-	**74.1**	**96.8**	**76.4**	61.7	**89.0**

Domain Generalization

Training (Source)	Inference (Target)	HOTA↑	MOTA↑	IDF1↑	AssA↑	DetA↑
SportsMOT	SportsMOT	75.0	95.6	77.9	63.6	88.4
BDD100K	SportsMOT	73.8	95.7	76.7	61.6	88.4
DanceTrack	DanceTrack	56.9	90.1	57.5	41.1	79.1
BDD100K	DanceTrack	54.9	89.2	55.3	38.4	78.7

Class	Car	Peds	Rider	Bus	Truck	Train	Motocy	Bycicle
Setting	Source & Target							
HOTA↑	66.2	50.4	47.3	62.1	55.0	0	47.6	48.0
IDF1↑	75.7	60.8	60.7	70.9	60.6	0	59.3	60.4
MOTA↑	73.0	55.8	48.9	58.4	47.2	-0.6	42.4	43.6
IDs↓	5917	2209	29	45	192	0	8	143
Setting	Source	Target						
HOTA↑	65.8	48.9	45.6	61.8	54.6	0	47.7	47.7
IDF1↑	74.9	58.6	57.7	70.4	60.4	0	60	59.8
MOTA↑	72.8	54.3	44.1	58.9	46.9	-0.6	41.7	43.3
IDs↓	6186	2790	23	44	140	0	8	152

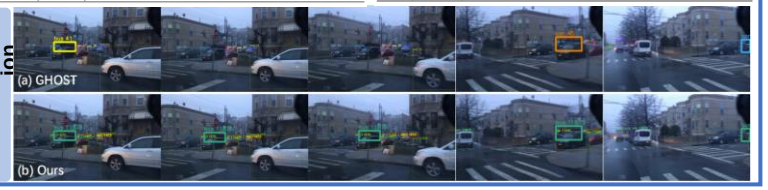
Attribute Maps for Datasets



Ablation study

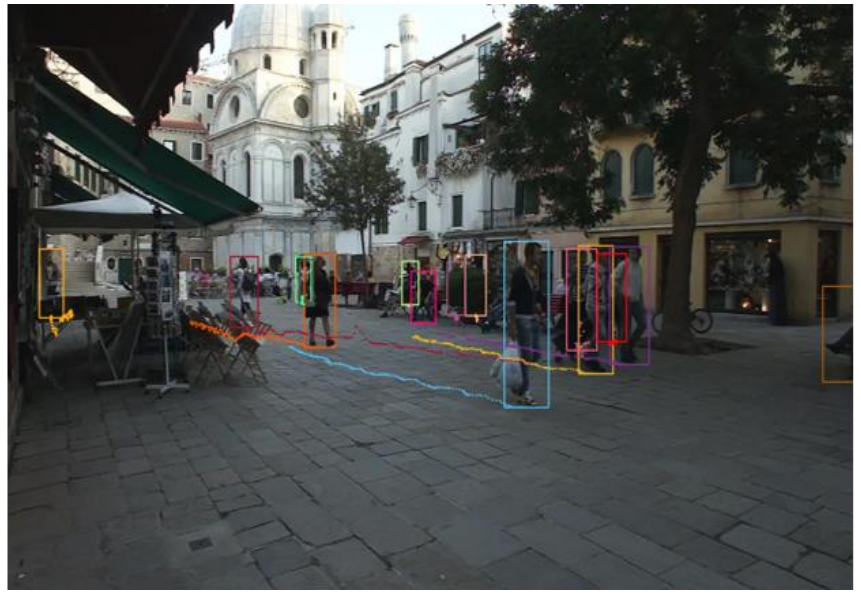
Setting	MRHRA		mHOTA↑	mIDF1↑	mMOTA↑	HOTA↑	IDF1↑	MOTA↑	IDs↓
#1	✓	✓	47.1	56.1	46.1	63.4	72.5	68.3	8503
#2	✓	✓	46.2	54.5	43.6	62.4	71.3	66.0	9070
#3	✓	✓	42.9	49.2	37.5	57.8	63.9	59.2	11584
#1		✓	46.9	55.7	45.6	63.1	72.2	67.9	9447
#2		✓	45.3	53.3	42.2	61.7	70.1	65.0	10673
#3		✓	41.7	47.8	36.3	55.8	61.2	56.7	14015
#1	✓		46.7	55.5	45.6	62.8	71.8	67.9	9070
#1	✓	✓ SR=1	46.9	55.7	46.0	63.1	72.1	68.2	9173
		SR=4	47.1	56.1	46.1	63.4	72.5	68.3	8503
		SR=7	46.9	55.7	46.0	63.5	72.7	68.3	8454

Visualization



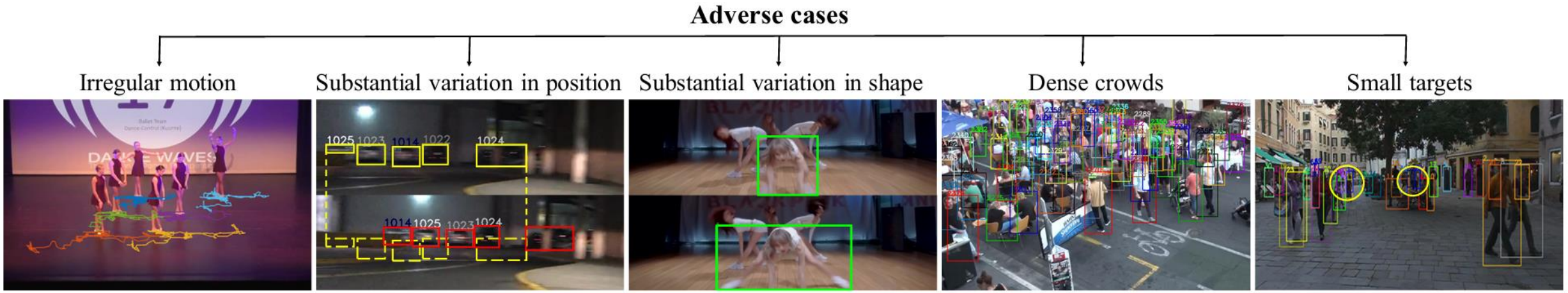
Background

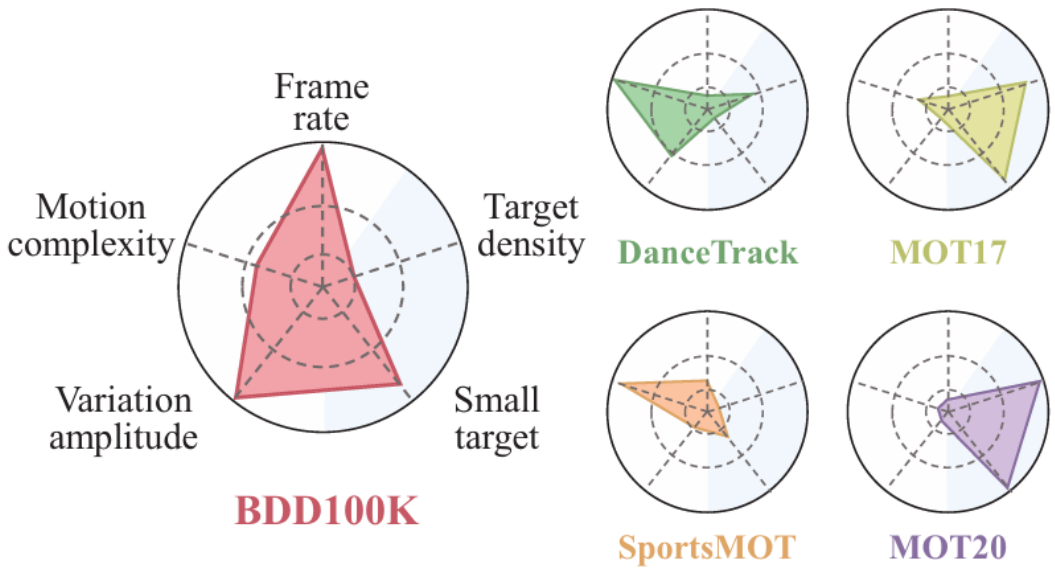
- （a）Locate targets through bounding boxes.
- （b）Recognize their identities throughout a whole video.



Motivation

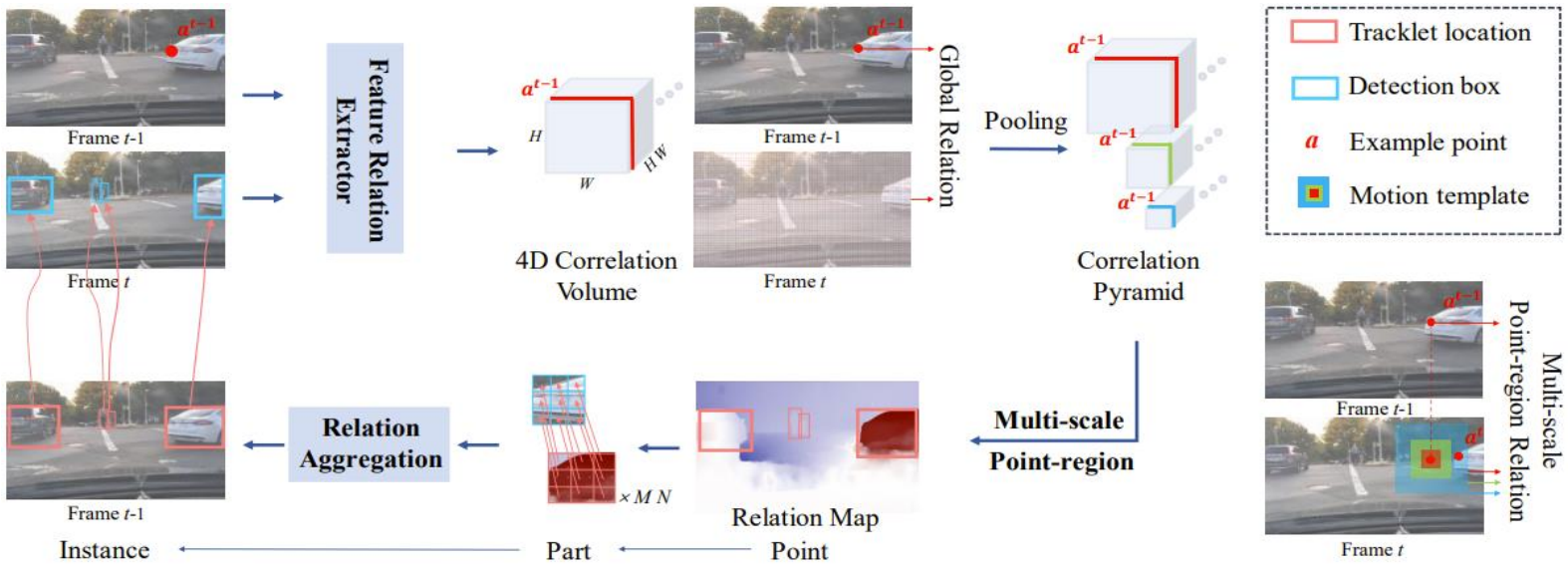
Existing trackers struggle to accommodate all aspects or necessitate hypothesis and experimentation to customize the association information (motion and/or appearance) for a given scenario, leading to narrowly tailored solutions with limited generalizability.



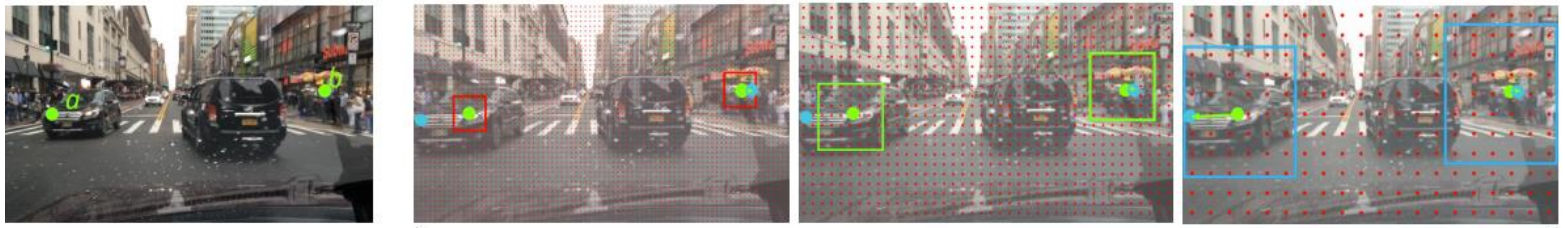


- **Motion Complexity** reflects the irregularity and unpredictability of target motion within the scenario. The more irregular and unpredictable the motion, the greater its complexity.
- **Variation Amplitude** reflects the target's variability, encompassing both shape and position variations.
- **Target Density** reflects the density of the crowds in the scenario, implicitly reflecting the degree of occlusion within the crowds.
- **Small Target** represents the average amount of small targets in the scenario.
- **Frame Rate** is the number of frames captured in one second of the input video stream.

Overview of MotionTrack



- ***Step 1: We use Feature Relation Extractor to construct global dense relations with frame t for each point in frame t − 1 by a 4D correlation volume.***
- ***Step 2: We transform the global relations into Multi-scale Point-region Relations, and form a relation map for frame t − 1.***
- ***Step 3: We perform Hierarchical Relational Aggregation according to the point-part-instance hierarchy to associate the tracklets and detections.***



Comparison with SOTA

	Venue	mHOTA↑	mIDF1↑	mMOTA↑	HOTA↑	IDF1↑	MOTA↑	IDs↓	MT↑	ML↓
validation										
QDTrack [38]	CVPR'21	-	50.8	36.6	-	71.5	63.5	**6262**	9481	3034
Unicorn [56]	ECCV'22	-	54.0	41.2	-	71.3	66.6	10876	10296	2505
MOTR [59]	ECCV'22	-	44.8	32.3	-	65.8	56.2	-	-	-
TETer [27]	ECCV'22	-	53.3	39.1	-	-	-	-	-	-
ByteTrack [62]	ECCV'22	45.3	54.8	45.2	61.3	70.4	**69.1**	9140	9626	3005
MOTRv2 [63]	CVPR'23	-	**56.5**	43.6	-	**72.7**	65.6	-	-	-
GHOST [46]	CVPR'23	45.7	55.6	44.9	61.7	70.9	68.1	-	-	-
GeneralTrack(Ours)		**46.9**	56.2	**46.4**	**63.1**	**72.7**	68.8	8496	**11830**	**2035**
test										
DeepBlueAI	-	-	38.7	31.6	-	56.0	56.9	25186	10296	12266
madamada	-	-	43.0	33.6	-	55.7	59.8	42901	16774	5004
QDTrack [58]	CVPR'21	41.9	52.4	35.7	60.5	72.5	64.6	**10790**	17353	5167
ByteTrack [62]	ECCV'22	-	55.8	**40.1**	-	71.3	**69.6**	15466	18057	5107
GHOST [46]	CVPR'23	46.8	**57.0**	39.5	62.2	72.0	68.9	-	-	-
GeneralTrack(Ours)		**47.9**	56.9	39.9	**63.7**	**73.6**	69.1	14489	**21281**	**3715**

	Venue	HOTA↑	MOTA↑	IDF1↑	AssA↑	DetA↑
Transformer based:						
MOTR [59]	ECCV'22	54.2	79.7	51.5	40.2	73.5
Hybird based:						
MOTRv2 [63]	CVPR'23	69.9	91.9	71.7	59.0	83.0
CNN based:						
ByteTrack [62]	ECCV'22	47.7	89.6	53.9	32.1	71.0
FineTrack [41]	CVPR'23	52.7	89.9	**59.8**	38.5	72.4
OC-SORT [6]	CVPR'23	55.1	**92.2**	54.9	40.4	80.4
GHOST [46]	CVPR'23	56.7	91.3	57.7	39.8	81.1
GeneralTrack (Ours)	-	**59.2**	91.8	59.7	**42.8**	**82.0**

	Venue	HOTA↑	MOTA↑	IDF1↑	AssA↑	DetA↑	IDs↓
MOT17							
MOTR [59]	ECCV'22	57.8	73.4	68.6	55.7	60.3	2439
ByteTrack [62]	ECCV'22	63.1	80.3	77.3	62.0	64.5	2196
OC-SORT [6]	CVPR'23	63.2	78.0	77.5	**63.2**	63.2	1950
MOTRv2 [63]	CVPR'23	62.0	78.6	75.0	60.6	63.8	-
GHOST [46]	CVPR'23	62.8	78.7	77.1	-	-	2325
GeneralTrack(Ours)	-	**64.0**	**80.6**	**78.3**	63.1	**65.1**	**1563**
MOT20							
ByteTrack [62]	ECCV'22	61.3	**77.8**	75.2	59.6	63.4	1223
OC-SORT [6]	CVPR'23	**62.1**	75.5	**75.9**	**62.0**	-	**913**
MOTRv2 [63]	CVPR'23	60.3	76.2	72.2	58.1	62.9	-
GHOST [46]	CVPR'23	61.2	73.7	75.2	-	-	1264
GeneralTrack(Ours)	-	61.4	77.2	74.0	59.5	**63.7**	1627

	Venue	HOTA↑	MOTA↑	IDF1↑	AssA↑	DetA↑
GTR [66]	CVPR'22	54.5	67.9	55.8	45.9	64.8
ByteTrack [62]	ECCV'22	64.1	95.9	71.4	52.3	78.5
OC-SORT [6]	CVPR'23	73.7	96.5	74.0	61.5	88.5
MixSort-Byte* [8]	ICCV'23	65.7	96.2	74.1	54.8	78.8
MixSort-OC* [8]	ICCV'23	**74.1**	96.5	74.4	**62.0**	88.5
GeneralTrack(Ours)	-	**74.1**	**96.8**	**76.4**	61.7	**89.0**

Ablation Study

Setting	MR	HRA	mHOTA↑	mIDF1↑	mMOTA↑	HOTA↑	IDF1↑	MOTA↑	IDs↓
#1	✓	✓	47.1	56.1	46.1	63.4	72.5	68.3	8503
#2	✓	✓	46.2	54.5	43.6	62.4	71.3	66.0	9070
#3	✓	✓	42.9	49.2	37.5	57.8	63.9	59.2	11584
#1		✓	46.9	55.7	45.6	63.1	72.2	67.9	9447
#2		✓	45.3	53.3	42.2	61.7	70.1	65.0	10673
#3		✓	41.7	47.8	36.3	55.8	61.2	56.7	14015
#1	✓		46.7	55.5	45.6	62.8	71.8	67.9	9070
#1	✓	✓							
	SR=1		46.9	55.7	46.0	63.1	72.1	68.2	9173
	SR=4		47.1	56.1	46.1	63.4	72.5	68.3	8503
	SR=7		46.9	55.7	46.0	63.5	72.7	68.3	8454

Class	HOTA↑	IDF1↑	MOTA↑	IDs↓
Pedestrian	50.3(+0.1)	60.7(+0.1)	55.6(+0.2)	2236(↓ 1.2%)
Rider	43.7(**+3.6**)	57.9(**+2.8**)	46.3(**+2.6**)	52(↓**44.2%**)
Car	66.2(+0.0)	75.4(+0.1)	73.1(-0.1)	6018(↓ 1.6%)
Bus	60.0(**+2.1**)	69.1(**+1.8**)	56.5(**+1.9**)	70(↓**35.7%**)
Truck	54.2(+0.8)	61.7(**-1.1**)	48.7(**-1.5**)	219(↓**12.3%**)
Train	0.0 (+0.0)	0.0 (+0.0)	-0.6 (+0.0)	0.0(↓ 0.0%)
Motorcycle	46.6(**+1.0**)	58.6(+0.7)	39.2(**+3.2**)	11(↓**27.3%**)
Bycicle	47.8(+0.2)	60.1(+0.3)	43.1(+0.5)	144(↓ 0.7%)
Detect_average	63.3(+0.1)	72.5(+0.0)	68.4(-0.1)	8750(↓ 2.8%)
Class_average	46.1(**+1.0**)	55.5(+0.6)	45.2(+0.9)	8750(↓ 2.8%)

Comparison for Domain generalization

Class	Car	Peds	Rider	Bus	Truck	Train	Motocy	Bycicle
Setting	**Source & Target**							
HOTA↑	66.2	50.4	47.3	62.1	55.0	0	47.6	48.0
IDF1↑	75.7	60.8	60.7	70.9	60.6	0	59.3	60.4
MOTA↑	73.0	55.8	48.9	58.4	47.2	-0.6	42.4	43.6
IDs↓	5917	2209	29	45	192	0	8	143
Setting	**Source**	**Target**						
HOTA↑	65.8	48.9	45.6	61.8	54.6	0	47.7	47.7
IDF1↑	74.9	58.6	57.7	70.4	60.4	0	60	59.8
MOTA↑	72.8	54.3	44.1	58.9	46.9	-0.6	41.7	43.3
IDs↓	6186	2790	23	44	140	0	8	152

Domain generalization for data with different classes.

Training (**Source**)	Inference (**Target**)	HOTA↑	MOTA↑	IDF1↑	AssA↑	DetA↑
SportsMOT	SportsMOT	75.0	95.6	77.9	63.6	88.4
BDD100K	SportsMOT	73.8	95.7	76.7	61.6	88.4
DanceTrack	DanceTrack	56.9	90.1	57.5	41.1	79.1
BDD100K	DanceTrack	54.9	89.2	55.3	38.4	78.7

Domain generalization for data in different datasets.

西安交通大学
XI'AN JIAOTONG UNIVERSITY

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

Thank you for listening !

qinzheng@stu.xjtu.edu.cn