Towards Generalizable Multi-Object Tracking -Supplementary Material-

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

1. Details of Tracking Scenario Attributes.

There are countless application scenarios in the world, each presenting unique characteristics. Designing an effective tracker requires the identification of factors with a significant impact on tracking, while disregarding those with minimal influence. To delve into the nature of these scenarios, a more concrete study and analysis are essential. In the following sections, we will provide definitions and quantitative calculations for each attribute.

1.1. Measurement Metric and Results

- **Motion Complexity.** This metric reflects the irregularity and unpredictability of target motion within the scenario. In our assessments, we decompose motion into direction and velocity. For motion velocity, we calculate the variance of successive velocity magnitudes for each target. For the direction of motion, we transform the continuous direction of the target into the polar coordinate form and calculate the direction mean and variance in the polar coordinate system. The final weighted sum of the two components is the motion complexity.
- **Variation Amplitude.** This metric reflects the magnitude of the target's variation, which consists of two components: shape variation and absolute position variation. For the former, we obtain the variance of the target's successive aspect ratios. For the latter, we calculate the magnitude of the target's movement relative to its own size. The final weighted sum of the two components is the variation amplitude.
- **Target Density.** This metric reflects the density of the crowd inside the scene, implicitly reflecting the degree of occlusion between the crowds. For a frame, we calculate the distance between each target in it and measure it by the average body size of the targets. Then we consider people to be occluded by each other when the distance between them is less than half of their body size. More generally, after averaging, this attribute represents the amount of occlusion per capita.
- **Small Target.** This metric represents the average content of small targets in the dataset. We use the target area to filter small targets with a certain threshold and count the average number of small targets in the scene.
- **Frame Rate.** This metric is the number of frames captured in one second of the input video stream. The larger the frame rate, the more information changes within the scene and the more difficult it is to tracking.

Scenario Attribute	Motion Complexity	Variation Amplitude	Target Density	Frame Rate	Small Target
BDD100K	1.76	1.80	0.90	5	7.11
SportsMOT	3.10	0.28	0.48	25	2.06
MOT17	1.19	0.03	2.77	30	7.51
MOT20	0.57	0.02	3.30	30	8.39
DanceTrack	3.44	1.34	1.75	30	0.00

Table 1. Scores on tracking scenario attributes on five datasets.

Scenario Attribute	Motion Complexity	Variation Amplitude	Target Density	Frame Rate	Small Target
BDD100K	0.41	1.00	0.15	1.00	0.85
SportsMOT	0.88	0.06	0.00	0.20	0.25
MOT17	0.22	0.00	0.81	0.00	0.90
MOT20	0.00	0.00	1.00	0.00	1.00
DanceTrack	1.00	0.55	0.45	0.00	0.00

Table 2. Normalization of detailed scores on tracking scenario attributes on five datasets.

Based on the definitions above, we measured these attributes on five datasets. The results are shown in Table 1. We normalized each attribute as shown in Table 2. We provide a qualitative comparison of motion complexity, displayed in Figure 1.

*Corresponding author.

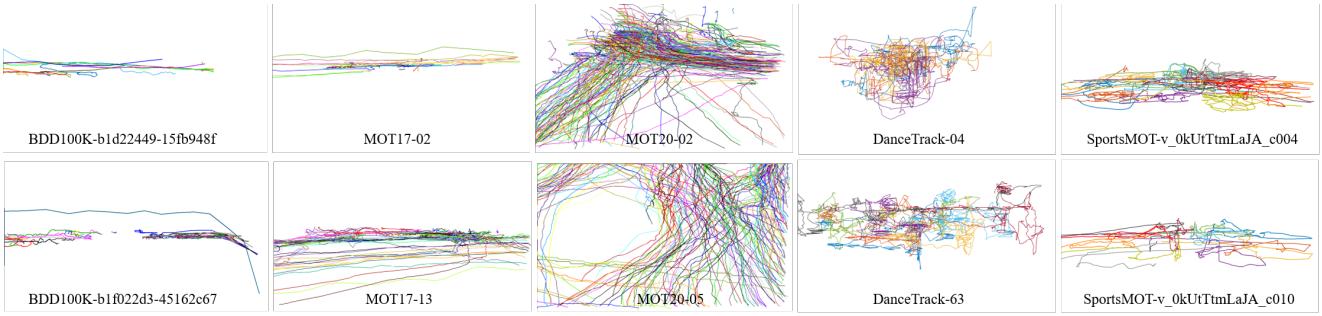

Figure 1. Qualitative comparison of motion complexity. Different colors represent the trajectories of different targets. The trajectories in both autonomous driving dataset (BDD100K) and pedestrian dataset (MOT17, MOT20) are linear and more predictable than the dancing and sports datasets (DanceTrack, SportsMOT).

	BDD100K	SportsMOT	DanceTrack	MOT17	MOT20
BytrTrack [28]	-	21.1	22.4	21.3	15.3
MOTRV2 [29]	11.2	-	6.8	6.4	6.5
GHOST [20]	11.1	-	2.7	1.2	0.6
Ours	18.7	13.5	14.6	15.6	7.6
Ours (Accel)	28.3	19.7	19.5	18.5	12.2

Table 3. Comparison of **FPS** on multiple datasets. 'Accel' represents the accelerated version. Note that the green color represents the inference speed on the basis of the detection result files and the black color represents the speed of the complete tracking process.

1.2. Candidate Attribute.

Appearance similarity. This attribute is used to describe the similarity of the appearance of the targets within the scene. Overly similar appearances, such as the target wearing the same dance outfit in DanceTrack, the same jersey in SportsMOT, *etc.*, can interfere with the use of the appearance-dominated method by reducing distinguishability between targets.

The reason we did not choose it as the major attribute is that it has a relatively small impact, compared to the damage that other attributes do to motion and appearance. For example, in DanceTrack [21], irregular motion can be fatal to the motion-based method. But even if the appearance is similar, we can still rely on the appearance-based method to track targets well.

2. Inference Speed.

As shown in Table 3, we give the inference speeds of our GeneralTrack and several commonly used SOTAs on multiple datasets. Note that all results are tested under 1 NVIDIA GeForce RTX 3090 Ti GPU. Among all prior trackers, BytrTrack is mostly the fastest, and our accelerated version can achieve inference speeds similar to it, but our performance is better on almost all datasets.

Accelerated Version. Because our Feature Relation Extractor constructs global relationships, reducing the image size can accelerate the inference speed with little impact on performance. So we can resize the input image to a smaller size if there is a need for acceleration.

3. Architecture and Training Details.

The weights-sharing convolutional neural network in Feature Relation Extractor consists of 6 residual blocks (2 at 1/2 resolution, 2 at 1/4 resolution, and 2 at 1/8 resolution). For training, we employ the AdamW optimizer [13] and limit the gradients to the interval [-1, 1]. The main purpose of dense flow and correspondence tasks is to construct a pixel-wise dense relationship of an image pair, and our task is to construct the pixel-wise relationship and transform them into instance-wise associations from fine to coarse. Therefore, we use KITTI [6] for pre-training in the settings of optical flow for enhancing the capability of Feature Relation Extractor (the weights-sharing convolutional neural network).

Background Mask. To focus more on foreground targets, we mask the background area to reduce its disturbance in the calculation of the correlation pyramid. Based on the tracklet bank $\mathbb{T}$ and prior detection $\mathcal{D}^t$, respectively, we generate binarized matrices, where each binarized element represents whether it belongs to a foreground target or not. Background mask is only used on DanceTrack in the benchmark results.

4. Elaborate Version of TbD in Related Works.

Tracking-by-Detection. The dominant paradigm in the field has long been tracking-by-detection [2, 3, 5, 8, 16, 17, 22, 24, 26, 28], which divides tracking into two steps: (i) frame-wise object detection, (ii) data association to link the detections and form trajectories. The core of the data association is to construct inter-frame relation (Affinity matrix) between tracklets and detections, and then complete the matching with the Hungarian algorithm [12]. The affinity matrix for matching is often driven by motion informa-

tion [2, 7, 17, 19] or appearance information [5, 18, 24, 25]. Motion-based trackers exploit the fact that object displacements tend to be small given two neighboring frames. This allows them to leverage spatial proximity for matching with tools such as Kalman filters [2, 9] or its variant version [7, 10, 27]. Some recent works [17, 19] use data-driven motion models for more accurate motion prediction and lead to more robust tracking. Motion trackers work well for pedestrian tracking scenarios with regular motion, *i.e.*, MOT17 and MOT20. But when the frame rate goes low, the movement amplitude gets larger, and the motion becomes more complex, motion wears out, and it's needed for appearance. Appearance relies on extracting discriminative features to construct instance-level relations. DeepSORT [24] firstly adopts a stand-alone Re-ID model to extract appearance features from the detection boxes. Follow-up efforts [4, 11, 15, 18, 20, 23, 25] use a variety of approaches to come up with better appearance models, such as domain adaptation [20], contrastive learning [16], *etc.*. These appearances rely on distinguishable overall voxel information and can handle the above occasions where motions cannot resolve. However, it has limitations when it encounters occlusion caused by dense crowds or the targets are too small to extract effective features.

In facing these issues, previous methods worked towards a better balance between motion and appearance. Some works [1, 5, 20, 24, 28] choose whether to emphasize motion or appearance more based on a very strong prior and multiple experiment attempts. TrackFlow [14] addresses these issues by building on an elegant probabilistic formulation that requires additional virtual datasets for training. In contrast, we propose a new tracking method that achieves generalization while avoiding the need to balance motion and appearance.

5. Detailed Analysis on Benchmarks.

BDD100K. Our GeneralTrack outperforms the state-of-the-art methods in most key metrics, *i.e.*, ranks first for metrics mTETA, mHOTA, mIDF1, mMOTA, HOTA, IDF1 and ranks second for MOTA in the validation set. On the test set, GeneralTrack achieves the best performance under most of the key metrics, with the rest of the metrics ranked second and very close to the best. Note that we use the same detection results as GHOST and ByteTrack, compared to which brings a big boost (+1.2 mHOTA, +1.4 HOTA, +0.6 mIDF1, + 1.8 IDF1, +1.5 mMOTA, +0.7 MOTA) in the validation set and (+1.1 mHOTA, +1.5 HOTA, + 1.6 IDF1) in the test set on GHOST; (+1.6 mHOTA, +1.8 HOTA, +1.4 mIDF1, +2.3 IDF1) in the validation set and (+1.1 IDF1, +2.3 IDF1) in the test set on ByteTrack. While Bytetrack selects appearance and GHOST weight summation of motion and appearance, in comparison, our approach outperforms such hand-designed algorithms by a large margin, demonstrating the generalizability of our approach to the multi-class tracking task with a low frame rate.

SportsMOT. GeneralTrack ranks first in all key metrics (HOTA, MOTA, IDF1). While using the same detections, we gain significant improvement (+8.4 HOTA, + 2.3 IDF1) on MixSort-Byte and (+ 2.0 IDF1) on MixSort-OC. Otherwise, MixSort-Byte and MixSort-OC train the association component on both the training set and the validation set; in contrast, we train only on the training set. Even so, we are still surpassing them and the improvements in metrics prove the superiority of our association capabilities even under very severe motion complexity.

DanceTrack. When being generalized to the dancing dataset, our method outperforms all CNN-based trackers. Note that all these CNN-based methods share the same detection, our GeneralTrack ranks first in HOTA and is 2.3 higher than the second place. Similarly, our AssA and DetA are 2.4 and 0.9 higher than the second place, respectively. Although our method is inferior to MOTRv2, it uses both YOLOX and MOTR with more than two hundred times training resource usage than ours. And our method outperforms it on several other datasets. These results indicate that our method is robust to large variation amplitudes of the target in addition to handling complex motions.

MOT17 and MOT20. Both datasets are pedestrian tracking datasets with regular motion patterns and both have dense crowd distributions and smaller targets. Our method ranks first in all key metrics HOTA, MOTA, IDF1 and DetA, IDs on MOT17 and ranks second in HOTA and MOTA on MOT20. On MOT17, GeneralTrack improves over the best previous methods, *e.g.*, gaining the improvement (+1.2 HOTA, +1.9 MOTA, + 1.2 IDF1) on GHOST and (+0.9 HOTA, +0.3 MOTA, + 1.0 IDF1) on ByteTrack. For MOT20, our method performs more stably under three key metrics compared to other trackers. These results show that our method can generalize well to scenarios where crowded and small targets exist.

6. More Discussion on Domain Generalization.

On domain generalization experiments for cross-class on BDD100K, domain generalization of GHOST [20] focuses on the ReID model while our method addresses the association model. We provide the performance of both training with one class and then tracking in the entire dataset (GeneralTrack:mHOTA 46.5, mIDF1 55.2, mMOTA 45.5 ; GHOST: mHOTA 45.7, mIDF1 55.6, mMOTA 44.9). It is worth noting that we only train on one class (car) on BDD100K, whereas GHOST train one class (people) with data outside of BDD100K.

Inuput Size	DS	mHOTA	mIDF1	mMOTA	HOTA	IDF1	MOTA	IDs
720×1280	2	47.1	56.1	46.1	63.4	72.5	68.3	8503
720×1280	4	46.6	55.3	45.3	62.6	71.5	67.7	10283
360×640	2	46.2	54.8	44.8	62.3	71.0	67.4	9781

7. Resolution and Downsampling Scale.

The ablation study below shows that our method is insensitive to both the frame resolution and the downsampling scale (DS).

References

[1] Nir Aharon, Roy Orfaig, and Ben-Zion Bobrovsky. Bot-sort: Robust associations multi-pedestrian tracking. *arXiv preprinbytetrack arXiv:2206.14651*, 2022. 3

[2] Alex Bewley, Zongyuan Ge, Lionel Ott, Fabio Ramos, and Ben Upcroft. Simple online and realtime tracking. In *ICIP*, pages 3464–3468, 2016. 2, 3

[3] Jinkun Cao, Jiangmiao Pang, Xinshuo Weng, Rawal Khirodkar, and Kris Kitani. Observation-centric sort: Rethinking sort for robust multi-object tracking. In *CVPR*, pages 9686–9696, 2023. 2

[4] Yonghao Dong, Le Wang, Sanping Zhou, Gang Hua, and Changyin Sun. Recurrent aligned network for generalized pedestrian trajectory prediction. *arXiv preprint arXiv:2403.05810*, 2024. 3

[5] Yunhao Du, Zhicheng Zhao, Yang Song, Yanyun Zhao, Fei Su, Tao Gong, and Hongying Meng. StrongSORT: Make deepsort great again. *IEEE T-MM*, 2023. 2, 3

[6] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *IJRR*, 32 (11):1231–1237, 2013. 2

[7] Shoudong Han, Piao Huang, Hongwei Wang, En Yu, Donghaisheng Liu, and Xiaofeng Pan. MAT: Motion-aware multi-object tracking. *Neurocomputing*, 476:75–86, 2022. 3

[8] Jiawei He, Zehao Huang, Naiyan Wang, and Zhaoxiang Zhang. Learnable graph matching: Incorporating graph partitioning with deep feature learning for multiple object tracking. In *CVPR*, pages 5299–5309, 2021. 2

[9] Rudolph Emil Kalman. A new approach to linear filtering and prediction problems. 1960. 3

[10] Tarasha Khurana, Achal Dave, and Deva Ramanan. Detecting invisible people. In *ICCV*, pages 3174–3184, 2021. 3

[11] Chanho Kim, Li Fuxin, Mazen Alotaibi, and James M Rehg. Discriminative appearance modeling with multi-track pooling for real-time multi-object tracking. In *CVPR*, 2021. 3

[12] Harold W Kuhn. The hungarian method for the assignment problem. *Naval research logistics quarterly*, pages 83–97, 1955. 2

[13] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 2

[14] Gianluca Mancusi, Aniello Panariello, Angelo Porrello, Matteo Fabbri, Simone Calderara, and Rita Cucchiara. DARTHTrackFlow: Multi-object tracking with normalizing flows. In *ICCV*, pages 9531–9543, 2023. 3

[15] Jiangmiao Pang, Linlu Qiu, Xia Li, Haofeng Chen, Qi Li, Trevor Darrell, and Fisher Yu. Quasi-dense similarity learning for multiple object tracking. In *CVPR*, pages 164–173, 2021. 3

[16] Jiangmiao Pang, Linlu Qiu, Xia Li, Haofeng Chen, Qi Li, Trevor Darrell, and Fisher Yu. Quasi-dense similarity learning for multiple object tracking. In *CVPR*, pages 164–173, 2021. 2, 3

[17] Zheng Qin, Sanping Zhou, Le Wang, Jinghai Duan, Gang Hua, and Wei Tang. MotionTrack: Learning robust short-term and long-term motions for multi-object tracking. In *CVPR*, pages 17939–17948, 2023. 2, 3

[18] Hao Ren, Shoudong Han, Huilin Ding, Ziwen Zhang, Hongwei Wang, and Faquan Wang. Focus on details: Online multi-object tracking with diverse fine-grained representation. In *CVPR*, pages 11289–11298, 2023. 3

[19] Fatemeh Saleh, Sadegh Aliakbarian, Hamid Rezatofighi, Mathieu Salzmann, and Stephen Gould. Probabilistic tracklet scoring and inpainting for multiple object tracking. In *CVPR*, pages 14329–14339, 2021. 3

[20] Jenny Seidenschwarz, Guillem Brasó, Víctor Castro Serrano, Ismail Elezi, and Laura Leal-Taixé. Simple cues lead to a strong multi-object tracker. In *CVPR*, pages 13813–13823, 2023. 2, 3

[21] Peize Sun, Jinkun Cao, Yi Jiang, Zehuan Yuan, Song Bai, Kris Kitani, and Ping Luo. DanceTrack: Multi-object tracking in uniform appearance and diverse motion. In *CVPR*, pages 20993–21002, 2022. 2

[22] Pavel Tokmakov, Jie Li, Wolfram Burgard, and Adrien Gaidon. Learning to track with object permanence. In *ICCV*, pages 10840–10849, 2021. 2

[23] Qiang Wang, Yun Zheng, Pan Pan, and Yinghui Xu. Multiple object tracking with correlation learning. In *CVPR*, pages 3876–3886, 2021. 3

[24] Nicolai Wojke, Alex Bewley, and Dietrich Paulus. Simple online and realtime tracking with a deep association metric. In *ICIP*, pages 3645–3649, 2017. 2, 3

[25] En Yu, Zhuoling Li, and Shoudong Han. Towards discriminative representation: Multi-view trajectory contrastive learning for online multi-object tracking. In *CVPR*, pages 8834–8843, 2022. 3

[26] Jimuyang Zhang, Sanping Zhou, Xin Chang, Fangbin Wan, Jinjun Wang, Yang Wu, and Dong Huang. Multiple object tracking by flowing and fusing. *arXiv preprint arXiv:2001.11180*, 2020. 2

[27] Yifu Zhang, Chunyu Wang, Xinggang Wang, Wenjun Zeng, and Wenyu Liu. FairMOT: On the fairness of detection and re-identification in multiple object tracking. *IJCV*, 129: 3069–3087, 2021. 3

[28] Yifu Zhang, Peize Sun, Yi Jiang, Dongdong Yu, Fucheng Weng, Zehuan Yuan, Ping Luo, Wenyu Liu, and Xinggang Wang. ByteTrack: Multi-object tracking by associating every detection box. In *ECCV*, pages 1–21, 2022. 2, 3

[29] Yuang Zhang, Tiancai Wang, and Xiangyu Zhang. Motrv2: Bootstrapping end-to-end multi-object tracking by pre-trained object detectors. In *CVPR*, pages 22056–22065, 2023. 2