DanceTrack: Multi-Object Tracking in Uniform Appearance and Diverse Motion

Peize Sun[1*], Jinkun Cao[2*], Yi Jiang[3], Zehuan Yuan[3], Song Bai[3], Kris Kitani[2], Ping Luo[1]
[1]The University of Hong Kong [2]Carnegie Mellon University [3]ByteDance Inc.

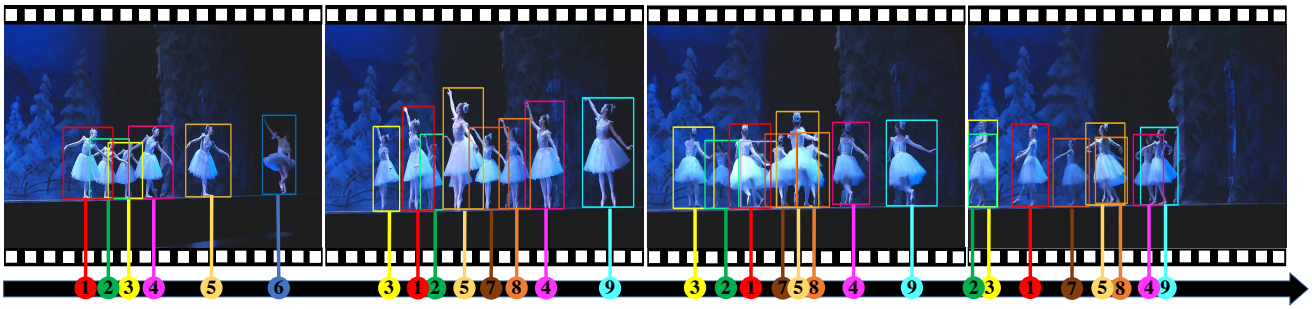


Figure 1 – Sample images from a video in DanceTrack: 1st, 66th, 307th and 327th frame in DanceTrack0027 video. . The emphasized properties of this dataset are (1) *uniform appearance*: humans are in highly similar and almost undistinguished appearance. (2) *diverse motion*: they are in complicated motion and interaction pattern. The numbers below show their identifications which experience frequent relative position switches and occlusions. We expect the combination of uniform appearance and complicated motion pattern makes DanceTrack a platform to encourage more comprehensive and intelligent multi-object tracking algorithms.

Abstract

A typical pipeline for multi-object tracking (MOT) is to use a detector for object localization, and following re-identification (re-ID) for object association. This pipeline is partially motivated by recent progress in both object detection and re-ID, and partially motivated by biases in existing tracking datasets, where most objects tend to have distinguishing appearance and re-ID models are sufficient for establishing associations. In response to such bias, we would like to re-emphasize that methods for multi-object tracking should also work when object appearance is not sufficiently discriminative. To this end, we propose a large-scale dataset for multi-human tracking, where humans have similar appearance, diverse motion and extreme articulation. As the dataset contains mostly group dancing videos, we name it "DanceTrack". We expect DanceTrack to provide a better platform to develop more MOT algorithms that rely less on visual discrimination and depend more on motion analysis. We benchmark several state-of-the-art trackers on our dataset and observe a significant performance drop on DanceTrack when compared against existing benchmarks. The dataset, project code and competition is released at: https://github.com/DanceTrack.

* indicates equal contribution.

1. Introduction

Object tracking has been long studied and can be beneficial to applications such as autonomous driving, video analysis and robot planning [1,4,25,35]. Multi-object tracking aims to localize and associate objects of interest along time. Interestingly, we observe that recent developments in multi-object tracking rely heavily on a paradigm of detection followed by re-ID, where mostly appearance cues are used to associate objects. This trend in algorithmic development makes existing solutions fail catastrophically in situations where objects share very similar appearance, *e.g.*, group dancing where performers wear uniform clothes. It inspires us to propose more comprehensive solutions by taking other cues into modeling, such as object motion patterns and temporal dynamics.

As with many other areas of computer vision, the development of multi-object tracking is influenced by benchmark datasets. Based on specified datasets [9, 13, 22, 36], data-driven methods are sometimes argued to be biased to certain data distributions. In this work, we recognize the limitations of existing multi-object tracking datasets lie on that many objects have distinct appearance and the motion pattern of objects are very regular or even linear. Motivated by these dataset properties, most recently developed multi-object tracking methods [23, 32, 33, 39] highly rely

on appearance matching to associate detected objects while taking little other cues into consideration. The dominant paradigm will fail in situations out of the biased distribution. This phenomenon is not what we expect if we aim to build more general and intelligent tracking algorithms.

To provide a new platform for more comprehensive multi-object tracking studies, we propose a new dataset in this paper. Because it mostly contains group dancing videos, we name it "DanceTrack". The dataset contains over 100K image frames (almost $10\times$ more than MOT17 datatset [22]). As shown in Figure 1, the emphasized properties of this dataset are (1) **uniform appearance**: people in videos wear very similar or even the same clothes, making their visual features hard to be distinguished by re-ID model and (2) **diverse motion**: people usually have very large-range motion and complex body gesture variation, proposing higher requirements for motion modeling. The second property also brings occlusion and crossover as a side-effect that human body has a large ratio of overlap with each other and their relative position exchanges frequently.

With the proposed dataset, we build a new benchmark including existing popular multi-object tracking methods. The results prove that current state-of-the-art algorithms [23, 27, 33, 37–40] fail to make satisfactory performance when they simply use appearance matching or linear motion models to associate objects across frames. Considering the cases focused on in this dataset happen frequently in our real life, we believe it shows the limitations of existing multi-object tracking algorithms on practical applications. To provide potential guidelines for further research, we analyze a range of choices in associating objects and achieve some beneficial conclusions: (1) fine-grained representations of objects, e.g., segmentation and pose, exhibit better ability than coarse bounding box; (2) depth information shows positive influence on associating objects, though we are solving a 2D tracking task; (3) motion modeling of temporal dynamics is important.

To conclude, the key contributions of our work to the object tracking community are as follows:

1. We build a new large-scale multi-object tracking dataset, DanceTrack, covering the scenarios where tracking suffers from low distinguishability of object appearance and diverse non-linear motion patterns.

2. We benchmark baseline methods on this newly built dataset with various evaluation metrics, showing the limitation of existing multi-object tracking algorithms.

3. We provide comprehensive analysis to discover more cues for developing multi-object trackers that are more robust in complicated real-life situations.

2. Related Works

Multi-object tracking datasets. Many multi-object tracking datasets have been proposed for different scenarios. Similar to our proposed dataset, many existing datasets focus on human tracking. PETS2009 [11] dataset is one of the earliest in this area. The more recent MOT15 [17], MOT17 [22] and MOT20 [9] datasets are all popular in this community. These datasets are limited in the aspects of undistinguished appearance and diverse motion. For example, MOT17 contains only a handful of videos and scenarios. Even MOT20 increases the density of objects and emphasizes the occlusion among them, the movements of objects are very regular and they still have distinguishable appearances. Association by pure appearance matching [23] could easily make success on these datasets and we will show that given the perfect detector, the tracking problem on these datasets can be solved by a very naive association strategy, in Section 4.2.

Besides, many other datasets are proposed for diverse objectives, e.g., WILDTRACK [6] for multi-camera tracking, Youtube-VIS [34] for video instance segmentation and tracking. With the increasing attraction of autonomous driving, some datasets are specifically built where the objects of interest are vehicles and pedestrians. KITTI [13] is one of the earliest large-scale multi-object tracking datasets for driving scenarios. More recently, BDD100K [36], Waymo [28] and KITTI360 [18] are made available to the public, still focusing on autonomous driving scenarios but providing much larger scale data than KITTI. With the limitation of lanes and traffic rules, the motion patterns of objects in these datasets are even more regular than those focusing on only moving people. There are many datasets focusing on more diverse object categories than persons and vehicles. The ImageNet-Vid [10] provides trajectory annotations for 30 object categories in over 1000 videos and TAO [8] annotates even 833 object categories to study object tracking on long-tailed distribution.

Tracking by matching appearance. In the recent development of multi-object tracking, appearance similarity serves as the dominant cue in many popular methods. For example, JDE [31] and FairMOT [39] learn object localization and appearance embedding using a shared backbone for better appearance representation. QDTrack [23] designs a contrastive training paradigm and dense localization for object detection and uses highly sensitive appearance comparison to match objects across frames. More recently, with the new focus of applying transformers [30] in vision tasks, TransTrack [27], TrackFormer [21] and MOTR [37] make attempts to leverage the attention mechanism in tracking objects in videos. In these works, the features of previous tracklets are passed to the following frames as the query to associate the same objects across frames. The appearance

information contained in the query is critical to keep tracklet consistency.

Although the rise of deep-learning model brings much more powerful visual representations than ever before, we still witness the failure of appearance matching in many real-world situations and expect to improve the tracking performance by taking other cues into account.

Motion analysis in object tracking. The displacement of objects-of-interest provides important cues for object tracking. Tracking objects by estimating their motions has inspired a line of researches. These tracking algorithms mainly follow the tracking-by-detection paradigm. Sequential analysis tools such as Particle filter [14, 15] and Kalman filter [16] are found efficient in such applications, for example, SORT [3] is developed on the Kalman filter motion model. Even though motion analysis has been used in many object tracking methods [31, 38, 39], all these methods can only handle simple linear motion pattern and provide limited help in more complicated situations. Furthermore, as deep networks bring the revolutionary ability to extract high-quality visual features, DeepSORT [32] tries to combine deep visual features and motion models to gain performance gain. Since then, motion-based object tracker has shown weak competitiveness and many focuses are towards appearance cues.

However, we argue that a more comprehensive and intelligent tracking algorithm should pay more attention to motion analysis since appearance is not always reliable.

3. DanceTrack

3.1. Dataset Construction

Dataset design. We focus on the scenarios where objects have similar or even the same appearance and diverse motion patterns, including frequent crossover, occlusion and body deformation. The first property makes tracking by purely comparing object appearance invalid because the extracted visual features are no longer distinguishable for different objects. The second property further requires more informative clues rather than appearance in tracking, such as motion analysis and temporal dynamics.

We argue that "crowd" by simply increasing the density of objects is not what we expect. For example, MOT20 [9] contains videos where groups of pedestrians are very crowded. But as the pedestrian movement is very regular, the relative position and occlusion area keep almost consistent, such "crowd" is not an obstacle for appearance matching. Therefore, we focus on situations where multiple objects are moving in a "relatively" large range, where the occluded areas are dynamically changing, and they are even in crossover. Such cases are common in real world but naive linear motion models can not handle them anymore.

Dataset	MOT17 [22]	MOT20 [9]	DanceTrack
Videos	14	8	**100**
Avg. tracks	96	**432**	9
Total tracks	1342	**3456**	990
Avg. len. (s)	35.4	**66.8**	52.9
Total len. (s)	463	535	**5292**
FPS	**30**	25	20
Total images	11,235	13,410	**105,855**

Table 1 – The comparison of dataset meta-information between DanceTrack and its closest benchmark for multi-human tracking, MOT17 and MOT20. DanceTrack contains much more videos and images than MOT datasets.

Video collection. To achieve the design goals described above, we collected videos including mostly group dancing from the Internet. As shown in Figure 2, the dancers usually wear very similar or even the same clothes. They make a large-range motion, diverse gestures and frequent crossover. These properties greatly satisfy our motivation. We collect the videos from different search engines with keywords like "street dance", "hip-pop dance", "cheerleading dance", "rhythmic gymnastics" and so on. The collection is only for publicly available videos and under the permit of fair use of video resources.

Annotation. We use a commercial tool to annotate the collected videos. The annotated labels include bounding boxes and identifications. For a partly-occluded object, a full-body box is annotated. For a fully-occluded object, we do not annotate it; when it re-appears in the future frame, its identification is kept as the same as in the previous frame when it is visible. To facilitate the annotation process, our tool can automatically propagate the annotated boxes from the previous frame to the current frame, and the annotator only needs to refine the boxes in the current frame. To build a high-quality dataset, the annotations have been checked by another group of people and errors are reported back to the annotators for re-annotation.

3.2. Dataset Statistic

We provide some analytical information of DanceTrack dataset and compare it with existing multi-object tracking datasets. The statistical information helps to understand the uniqueness of the proposed dataset.

Dataset split. We collect 100 videos in DanceTrack dataset, by default using 40 videos as training set, 25 as validation set and 35 as test set. For splitting, we keep the distribution of subsets close in terms of average length, average bounding box number, scenes and the motion diversity. We make the annotation of training set and validation set public while keeping the testing set annotation private for competition use. Some basic information of DanceTrack is

Figure 2 – Sampled scenes from DanceTrack dataset. DanceTrack contains multiple genres of dance, including classical dance, street dance, pop dance, large group dance and sports. The scenes in DanceTrack are diverse: (a) outdoor scenes; (b) low-lighting and distant camera scenes; (c) large group of dancing people; (d) gymnastics scene where the motion is usually even more diverse and people have more aggressive deformation.

shown in Table 1. Compared with MOT datasets, DanceTrack has much larger volume (10x more images and 10x more videos). MOT20 focuses on crowded scenes, so it has more tracks but the appearance of objects is very distinguishable and their motion is regular. As a consequence, the association on MOT20 still requires little motion estimation when good detection results are provided.

Scene diversity. DanceTrack contains diverse scenes. Samples from all 100 videos are provided in Figure 2. One shared property for all videos is that the instances of people in a video usually have very similar appearance. This is designed on purpose to avoid the shortcut of tracking by pure appearance matching. DanceTrack contains multiple genres of dance, such as street dance, pop dance, classical dance (ballet, tango, etc.) and large group dancing. It also contains some sports scenarios such as gymnastics, Chinese Kung Fu and cheerleader dancing. Figure 2(a) shows outdoor scenes though most included videos are indoor. Figure 2(b) shows some especially hard cases, such as low lighting and distant camera. Figure 2(c) shows a large group of people dancing, including at most 40 people. Figure 2(d) shows gymnastics where people show extremely diverse body gestures, frequent pose variation and complicated motion pattern.

Appearance similarity. We make quantitative analysis about how appearance-only matching is not reliable on DanceTrack by measuring the appearance similarity among objects. We use a pre-trained re-ID model [24] to extract the appearance features $F(B_i^t)$ of object B_i on a frame t, and then compute the sum of cosine distance of the re-ID features among objects in the video as

$$V = \frac{1}{T}\sum_{t=1}^{T}\frac{1}{N_t^2}\sum_{i}^{N_t}\sum_{j\neq i}^{N_t}(1-\cos < F(B_i^t), F(B_j^t) >), \tag{1}$$

where T is the number of frames in the video sequence, N_t is the number of objects on the frame t and $<\cdot>$ is the angle between two vectors.

We compare the object appearance similarity in DanceTrack to that in MOT17 dataset, as shown in Figure 3(a), each bin represents one video sequence. It is obvious that the cosine distance of re-ID features of DanceTrack is lower than that of MOT17, in other words, the appearance similarity among co-existing objects is higher. This quantitative analysis shows the challenge of DanceTrack to current popular appearance matching for association.

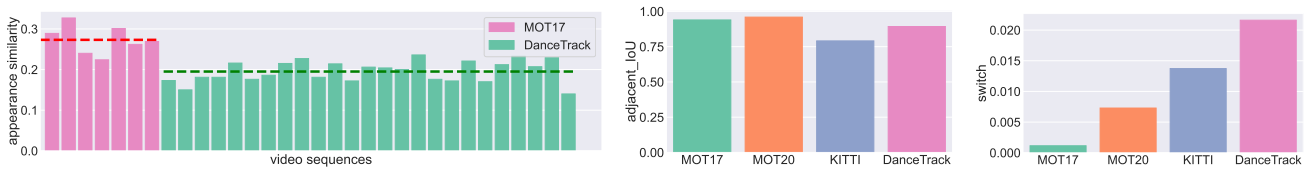
Motion pattern. We introduce two metrics to analyze the motion pattern in DanceTrack dataset and compare that to other multi-object tracking datasets.

IoU on adjacent frames: a natural measurement of object movement range is its bounding-box IoU (Intersection-over-Union) on two adjacent frames. A low IoU indicates fast-moving objects or the low frame rate of videos. Given a video with N objects and T frames, the averaged IoU on adjacent frames for this video is

$$U = \frac{1}{N(T-1)}\sum_{i}^{N}\sum_{t=1}^{T-1} IoU(B_i^t, B_i^{t+1}). \tag{2}$$

Frequency of Relative Position Switch: a metric to measure the diversity of objects' motion in a global view is the frequency for two objects to switch their relative position. This could happen between leftward and rightward or between upward and downward. On the contrary, movement with consistent velocity tends to cause a lower chance of relative position switch. Given a video, the average frequency of relative position switch is defined as

$$S = \frac{\sum_i^N \sum_{j\neq i}^N \sum_{t=1}^{T-1} sw(B_i^t, B_j^t, B_i^{t+1}, B_j^{t+1})}{2N(T-1)(N-1)}, \tag{3}$$



(a) Cosine distance of re-ID feature

(b) IoU on adjacent frames

(c) Frequency of relative position switch

Figure 3 – (a) Cosine distance of re-ID features. The dashed lines are for the average cosine distance similarity for the two datasets. The cosine distance of re-ID features of DanceTrack is lower than that of MOT17, in other words, the appearance similarity between different objects is higher. (b) IoU on adjacent frames. Compared to MOT17 and MOT20, DanceTrack has a similar score. It means that the frame rate and object motion speed are still reasonable in DanceTrack. (c) Frequency of relative position switch. This metric measures the frequency of crossover and is highly related to the occlusion between objects. DanceTrack has much more frequent relative position switches than other pedestrian tracking datasets, such as MOT17 and MOT20. Even compared to the driving dataset KITTI, where the moving camera naturally causes many relative position switches, DanceTrack still has a higher frequency.

where sw is an indicator function, where $sw(\cdot)$=1 if the two objects swap their left-right relative position or top-down relative position on the adjacent frames, $sw(\cdot)$=0 if there is no swap. We measure their relative position by comparing their bounding box center locations. And considering that such crossover causes potential difficulty only when the objects have overlap, we only take the objects with overlap into the calculation.

From the results shown in Figure 3(b), we could find that DanceTrack and MOT datasets have close average IoU on adjacent frames. This indicates that DanceTrack does not have unreasonably fast object movement.

On the other hand, from Figure 3(c) we could find that DanceTrack has much more frequent relative position switches than other datasets such as KITTI, MOT17 and MOT20. The frequent relative position switches are caused by highly non-linear motion pattern and result in frequent crossover and inter-object occlusion. This result shows that the challenge of motion diversity in DanceTrack.

3.3. Evaluation Metrics

For a long time, multi-object tracking community used Multi-Object Tracking Accuracy (MOTA) as the main metric for evaluation. However, recently, the community realizes that MOTA focuses too much on detection quality instead of association quality. Thus, Higher Order Tracking Accuracy (HOTA) [20] is proposed to correct this historical bias. Up to now, HOTA has been used for the main metrics to evaluate tracking quality on multiple popular benchmarks such as BDD100K [36] and KITTI [13]. We follow this setting for evaluation metrics of DanceTrack. In our protocol, the main metric is HOTA. We also use AssA and IDF1 score to measure association performance and DetA and MOTA for detection quality. For the detailed definitions of these metrics, we refer to [2, 20, 26]. To make it convenient to run for fine-grained analysis, the evaluation tools also provide previously widely-used statistics, such as False Positive (FP), False Negative (FN) and ID switch (IDs).

3.4. Limitation

We discuss some limitations of the proposed dataset. First, given the mentioned motivation and the proposed dataset, we do not provide an algorithm that highly outperforms previous multi-object tracking algorithms but keep this as an open question for future study. Second, for the cases we emphasize in this work, the annotation of human pose or segmentation mask should be important for more fine-grained study. But limited by time and resources, we only provide the annotation of bounding boxes in this version.

4. Experiments

4.1. Experiment Setup

Dataset configurations We compare DanceTrack with its closest dataset, MOT17. For MOT17, because the test server is not available easily, we follow the train-val splitting provided in CenterTrack [41] to evaluate on the validation subset, unless in Section 4.3. For DanceTrack, we follow the default splitting described in the previous section.

Model configuration Unless specified otherwise, we inherit the default training settings of the investigated algorithms provided in the original papers or the officially released codebases.

4.2. Oracle Analysis

To decompose the analysis over object localization and association, we perform oracle analysis here. We use the ground truth bounding boxes with different association algorithms to achieve the upper-bound performance. This analysis can help us to understand what is the true bottleneck of tracking on different datasets.

We compare IoU matching, motion modeling and appearance matching for the association. IoU matching is simply performed by calculating the IoU of objects' bound-

Appearance	IoU	Motion	MOT17					DanceTrack (Proposed Dataset)				
			HOTA	DetA	AssA	MOTA	IDF1	HOTA	DetA	AssA	MOTA	IDF1
	✓		**98.1**	98.9	**97.3**	98.0	97.8	**72.8**	**98.9**	53.6	98.7	63.5
	✓	✓	96.4	97.1	95.8	**99.7**	98.1	69.4	87.9	**54.8**	**99.4**	**71.3**
✓	✓	✓	95.0	94.7	95.4	99.3	**98.8**	59.7	82.5	43.2	97.2	60.5
✓			93.3	**99.0**	87.9	98.9	90.9	68.0	97.7	47.4	97.9	58.7

Table 2 – Oracle analysis of different association models on MOT17 and DanceTrack validation set, respectively. The detection boxes are ground-truth boxes. The result comparison shows the evident increased difficulty of performing multi-object tracking on DanceTrack than MOT17 dataset.

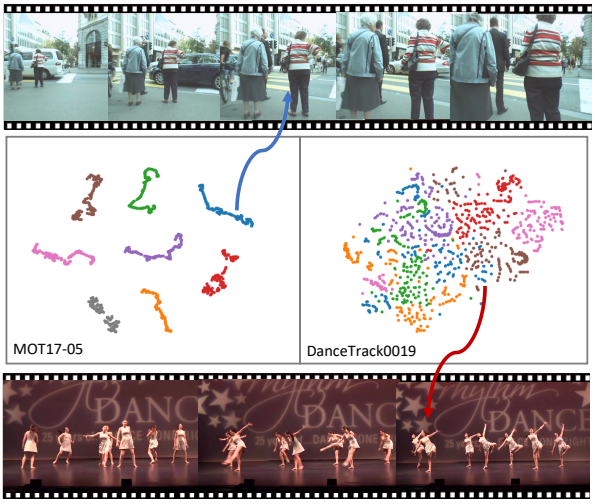


Figure 4 – Visualization of re-ID feature from sampled video in MOT17 and DanceTrack dataset using t-SNE [29]. The same object is coded by the same color. For better visualization, we only select first 200 frames in each video sequence.

ing boxes in adjacent frames. We use a pre-trained Re-ID model [24] for appearance matching and a Kalman Filter [16] for motion modeling under linear motion assumption. We have experiments on MOT17 and DanceTrack respectively. The results are shown in Table 2.

From the results, the performance is almost perfect in terms of all metrics on MOT17. And it is interesting that using only IoU matching achieves the best performance, which proves that MOT17 contains objects with simple and regular motion patterns and the bottleneck does not lie in association in most cases.

On the other hand, using only IoU matching on DanceTrack gives a much lower performance than on MOT17. Given DetA and MOTA scores are already close to 100, the bottleneck is obviously in the association part. All association metric scores in all cases experience a dramatic drop compared with that on MOT17. Besides, the best performance lies in only IoU matching, even combining a linear motion model or additional appearance information does not help. When using appearance similarity, all metrics are worse than not using any appearance cue. This is because the objects in DanceTrack videos usually have indistinguishable appearance so simply using appearance matching makes negative effects in some cases. In Figure 4, we visualize the appearance feature of objects extracted from DanceTrack and MOT17 videos respectively. We can observe that the appearance features of different objects are very distinguishable in the feature space on MOT17 while highly entangled on DanceTrack. This qualitatively provides evidence for the high similar appearance of objects in the proposed DanceTrack dataset.

Given the results shown in the analysis with oracle object localization, we can reach a clear conclusion that existing datasets have a heavy bias that it focuses more on the detection quality only and the involved simple trajectory patterns limit the study in this area. On the contrary, DanceTrack is proposing a much higher requirement to develop multi-object trackers with improvement in association ability. Considering the scenarios included in DanceTrack are what we experience in real life, we believe it is meaningful to provide such a platform.

4.3. Benchmark Results

We benchmark the current state-of-the-art multi-object tracking algorithms on MOT17 and DanceTrack. The evaluation is in the "private" setting that the algorithm performs both detection and association. The benchmark results are reported in Table 3.

For tracking quality measured by HOTA, IDF1 and AssA, all algorithms show a significant performance gap from MOT17 to DanceTrack. For all investigated methods, their performance on DanceTrack is far from satisfactory. Notably, the detection quality metrics, MOTA and DetA, of all algorithms are in fact higher on DanceTrack than on MOT17. This suggests that detection is not the bottleneck to have good tracking performance on DanceTrack and further highlights the drop of association performance. The benchmark results prove that DanceTrack raises the challenge to make robust association in the cases of the uniform appearance and the diverse motion of objects.

Methods	MOT17					DanceTrack (Proposed Dataset)				
	HOTA	DetA	AssA	MOTA	IDF1	HOTA	DetA	AssA	MOTA	IDF1
CenterTrack [40]	52.2	53.8	51.0	67.8	64.7	41.8	**78.1**	22.6	86.8	35.7
FairMOT [39]	59.3	60.9	58.0	73.7	72.3	39.7	66.7	23.8	82.2	40.8
QDTrack [23]	53.9	55.6	52.7	68.7	66.3	45.7	72.1	29.2	83.0	44.8
TransTrack [27]	54.1	61.6	47.9	75.2	63.5	45.5	75.9	27.5	88.4	45.2
TraDes [33]	52.7	55.2	50.8	69.1	63.9	43.3	74.5	25.4	86.2	41.2
MOTR [37]	57.2	58.9	55.8	71.9	68.4	**54.2**	73.5	**40.2**	79.7	51.5
ByteTrack [38]	**63.1**	**64.5**	**62.0**	**80.3**	**77.3**	47.7	71.0	32.1	**89.6**	**53.9**

Table 3 – Tracking performance of investigated algorithms on MOT17 and DanceTrack **test set**. The result comparison shows the evident increased difficulty of performing multi-object tracking on DanceTrack than MOT17 dataset.

Association	HOTA	DetA	AssA	MOTA	IDF1
IoU	44.7	**79.6**	25.3	87.3	36.8
SORT [3]	**47.8**	74.0	31.0	**88.2**	48.3
DeepSORT [32]	45.8	70.9	29.7	87.1	46.8
MOTDT [7]	39.2	68.8	22.5	84.3	39.6
BYTE [38]	47.1	70.5	**31.5**	**88.2**	**51.9**

Table 4 – Comparison of different association algorithms on DanceTrack validation set. The detection results are output by YOLOX [12] detector, trained on DanceTrack training set.

Motion	HOTA	DetA	AssA	MOTA	IDF1
None(IoU)	44.7	**79.6**	25.3	87.3	36.8
Kalman filter [3]	47.8	74.0	31.0	88.2	48.3
LSTM [5]	**51.6**	78.2	**34.2**	**89.2**	**50.8**

Table 5 – Comparison of different motion models on DanceTrack validation set. The detection results are output by YOLOX [12] detector, trained on DanceTrack training set.

4.4. Association Strategy

The methods in the previous section entangle the detection and tracking modules. To have an independent study on association algorithms, we use the most recently developed YOLOX [12] detector for object detection on DanceTrack and conduct different association algorithms following that. The results are shown in Table 4.

SORT [3] uses Kalman Filter to model the object motion and DeepSORT [32] adds appearance matching. Compared to SORT, DeepSORT shows no performance boost but worse performance instead, suggesting the negative gain due to appearance matching. On the other hand, MOTDT [7] uses the tracking result to help detect bounding boxes. But in fact, detection performance can be really good on DanceTrack dataset and the exact bottleneck is the association part, so MOTDT shows even worse performance on both detection quality and association quality with its design. Lastly, BYTE [38] uses a high-tolerance strategy to select detection results into the association stage. The design aims to decrease tracklet fragmentation in tracking. With such a strategy, BYTE shows the best association performance in terms of IDF1 and AssA metrics. This also reveals that DanceTrack is not a strict challenge for object detectors, the true challenge is in the object association part.

We further use different motion models to introduce temporal dynamics in the tracking process to facilitate better association, as shown in Table 5. Obviously, both Kalman filter [3] and LSTM [5] outperform naive IoU association (without temporal dynamics) by a large margin, indicating the great potential of motion models in tracking objects, especially when appearance cues are not reliable. With the relatively slow progress of object model motion in the field of multi-object tracking, we expect to see more researches.

4.5. Analysis of More Modalities

Considering high scores of MOTA and DetA on DanceTrack, the limited performance on DanceTrack is an exact failure of trackers instead of detectors. To boost performance, a straightforward strategy is to add more cues other than frame-wise bounding box. Since DanceTrack contains bounding boxes and identities annotations only, we propose to use joint-training technology with other datasets to enable the model output more modalities.

Does fine-grained representation help ? We investigate the influence of adding segmentation mask into the model. From Table 6, we observe a performance boost by using the segmentation mask. First, the introduction of more fine-grained annotation benefits the model by multi-task learning. Second, for crowded and occluded situations, mask is a more reliable information than bounding box to associate objects. Besides mask, adding pose information in training better boosts the model performance on DanceTrack, and using the output pose in association further helps to achieve better tracking results. When most areas of a human body are occluded, bounding box usually can not provide reliable output while the pose estimation model focusing on certain human body key-points usually shows higher robustness.

Does depth information help ? We use additional depth information to help tracking on DanceTrack. The results are shown in Table 6. In contrast to the COCO segmentation mask and human pose, depth information learned from KITTI dataset does not increase the performance on Dance-

Figure 5 – Visualization of adding more information beyond bounding box on DanceTrack. Tracks are coded by color. The 1st, 2nd and 3rd column are frame20, 120 and 200 of DanceTrack0007 video.

Data	Ass.	HOTA	DetA	AssA	MOTA	IDF1
DanceTrack	box	36.9	63.6	21.6	78.8	39.2
+ COCOmask [19]	box	38.1 (**+1.2**)	64.5 (**+0.9**)	22.6 (**+1.0**)	80.6 (**+1.8**)	40.3 (**+1.1**)
+ COCOmask	+ mask	39.2 (**+1.1**)	64.9 (**+0.4**)	23.9 (**+1.3**)	80.7 (**+0.1**)	41.6 (**+0.3**)
DanceTrack	box	36.9	63.6	21.6	78.8	39.2
+ COCOpose [19]	box	40.6 (**+3.7**)	65.5 (**+1.9**)	25.3 (**+3.7**)	82.9 (**+4.1**)	42.9 (**+3.7**)
+ COCOpose	+ pose	41.0 (**+0.4**)	65.9 (**+0.4**)	25.6 (**+0.3**)	83.1 (**+0.3**)	43.9 (**+1.0**)
DanceTrack	box	36.9	63.6	21.6	78.8	39.2
+ KITTI [13]	box	34.4 (- **2.5**)	57.8 (- **5.8**)	20.7 (- **0.9**)	72.9 (- **5.9**)	38.5 (- **0.7**)
+ KITTI	+ depth	35.1 (**+0.7**)	57.3 (- **0.5**)	21.6 (**+0.9**)	72.8 (- **0.1**)	40.2 (**+1.7**)

Table 6 – Ablation study on adding more information beyond bounding box on DanceTrack validation set. All experiments are based on CenterNet [41] model and BYTE [38] association. (a) Segmentation mask improves the tracking performance on DanceTrack. (b) Pose information boosts the tracking performance with an even larger gap than segmentation mask. (c) Though adding depth information into association shows a slightly positive influence, the results still blame the domain shift between KITTI and DanceTrack.

Track. We explain that COCO segmentation and pose estimation datasets contain human as the main category, while KITTI mainly contains vehicle instances. Thus, the object and scene prior in DanceTrack and KITTI change and this domain shift degenerates the model. Nevertheless, depth information indeed helps association performance if we regard the baseline as the model trained on joint-dataset of DanceTrack and KITTI. However, limited by the available resources of depth-annotated data, this is the best we could try for now. We expect more study on the influence of depth information to associate objects with uniform appearance and diverse motion.

5. Conclusion

In this paper, we propose a new multi-object tracking dataset called DanceTrack. The objects have uniform appearance and diverse motion pattern in DanceTrack, preventing being taken short-cuts by Re-ID algorithms. The motivation behind it is to reveal the bias in existing datasets that tend to emphasize detection quality and matching appearance only. This makes other cues to associate objects underrepresented. We believe that the ability to analyze the complex motion pattern is necessary for building a more comprehensive and intelligent tracker. DanceTrack provides such a platform to encourage future works.

Acknowledgement We would like to thank the annotator teams and coordinators to build DanceTrack dataset. We appreciate Xinshuo Weng, Yifu Zhang for valuable discussion and suggestions. We would also like to thank Vivek Roy, Pedro Morgado, Shuyang Sun for their proof reading and suggestions on paper writing. This work was sponsored in part by NSF NRI Award IIS2024173. Ping Luo is supported by the General Research Fund of HK No.27208720 and 17212120.

References

[1] Philipp Bergmann, Tim Meinhardt, and Laura Leal-Taixe. Tracking without bells and whistles. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 941–951, 2019. 1

[2] Keni Bernardin and Rainer Stiefelhagen. Evaluating multiple object tracking performance: the clear mot metrics. *EURASIP Journal on Image and Video Processing*, 2008:1–10, 2008. 5

[3] Alex Bewley, Zongyuan Ge, Lionel Ott, Fabio Ramos, and Ben Upcroft. Simple online and realtime tracking. In *IEEE international conference on image processing*, pages 3464–3468, 2016. 3, 7

[4] Jinkun Cao, Xin Wang, Trevor Darrell, and Fisher Yu. Instance-aware predictive navigation in multi-agent environments. In *IEEE International Conference on Robotics and Automation*, pages 5096–5102, 2021. 1

[5] Mohamed Chaabane, Peter Zhang, Ross Beveridge, and Stephen O'Hara. Deft: Detection embeddings for tracking. *arXiv preprint arXiv:2102.02267*, 2021. 7

[6] Tatjana Chavdarova, Pierre Baqué, Stéphane Bouquet, Andrii Maksai, Cijo Jose, Timur Bagautdinov, Louis Lettry, Pascal Fua, Luc Van Gool, and François Fleuret. Wildtrack: A multi-camera hd dataset for dense unscripted pedestrian detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5030–5039, 2018. 2

[7] Long Chen, Haizhou Ai, Zijie Zhuang, and Chong Shang. Real-time multiple people tracking with deeply learned candidate selection and person re-identification. In *IEEE international conference on multimedia and expo*, pages 1–6, 2018. 7

[8] Achal Dave, Tarasha Khurana, Pavel Tokmakov, Cordelia Schmid, and Deva Ramanan. Tao: A large-scale benchmark for tracking any object. In *European Conference on Computer Vision*, pages 436–454. Springer, 2020. 2

[9] Patrick Dendorfer, Hamid Rezatofighi, Anton Milan, Javen Shi, Daniel Cremers, Ian Reid, Stefan Roth, Konrad Schindler, and Laura Leal-Taixé. Mot20: A benchmark for multi object tracking in crowded scenes. *arXiv preprint arXiv:2003.09003*, 2020. 1, 2, 3

[10] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. 2

[11] James Ferryman and Ali Shahrokni. Pets2009: Dataset and challenge. In *IEEE international workshop on performance evaluation of tracking and surveillance*, pages 1–6, 2009. 2

[12] Zheng Ge, Songtao Liu, Feng Wang, Zeming Li, and Jian Sun. Yolox: Exceeding yolo series in 2021. *arXiv preprint arXiv:2107.08430*, 2021. 7

[13] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3354–3361, 2012. 1, 2, 5, 8

[14] Fredrik Gustafsson. Particle filter theory and practice with positioning applications. *IEEE Aerospace and Electronic Systems Magazine*, 25(7):53–82, 2010. 3

[15] Rooji Jinan and Tara Raveendran. Particle filters for multiple target tracking. *Procedia Technology*, 24:980–987, 2016. 3

[16] Rudolph Emil Kalman. A new approach to linear filtering and prediction problems. 1960. 3, 6

[17] Laura Leal-Taixé, Anton Milan, Ian Reid, Stefan Roth, and Konrad Schindler. Motchallenge 2015: Towards a benchmark for multi-target tracking. *arXiv preprint arXiv:1504.01942*, 2015. 2

[18] Yiyi Liao, Jun Xie, and Andreas Geiger. Kitti-360: A novel dataset and benchmarks for urban scene understanding in 2d and 3d. *arXiv preprint arXiv:2109.13410*, 2021. 2

[19] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*, pages 740–755. Springer, 2014. 8

[20] Jonathon Luiten, Aljosa Osep, Patrick Dendorfer, Philip Torr, Andreas Geiger, Laura Leal-Taixé, and Bastian Leibe. Hota: A higher order metric for evaluating multi-object tracking. *International Journal of Computer Vision*, 129(2):548–578, 2021. 5

[21] Tim Meinhardt, Alexander Kirillov, Laura Leal-Taixe, and Christoph Feichtenhofer. Trackformer: Multi-object tracking with transformers. *arXiv preprint arXiv:2101.02702*, 2021. 2

[22] Anton Milan, Laura Leal-Taixé, Ian Reid, Stefan Roth, and Konrad Schindler. Mot16: A benchmark for multi-object tracking. *arXiv preprint arXiv:1603.00831*, 2016. 1, 2, 3

[23] Jiangmiao Pang, Linlu Qiu, Xia Li, Haofeng Chen, Qi Li, Trevor Darrell, and Fisher Yu. Quasi-dense similarity learning for multiple object tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 164–173, 2021. 1, 2, 7

[24] Ziqiang Pei. Deepsort pytorch. https://github.com/ZQPei/deep_sort_pytorch, 2019. 4, 6

[25] Akshay Rangesh and Mohan Manubhai Trivedi. No blind spots: Full-surround multi-object tracking for autonomous vehicles using cameras and lidars. *IEEE Transactions on Intelligent Vehicles*, 4(4):588–599, 2019. 1

[26] Ergys Ristani, Francesco Solera, Roger Zou, Rita Cucchiara, and Carlo Tomasi. Performance measures and a data set for multi-target, multi-camera tracking. In *European Conference on Computer Vision*, pages 17–35. Springer, 2016. 5

[27] Peize Sun, Yi Jiang, Rufeng Zhang, Enze Xie, Jinkun Cao, Xinting Hu, Tao Kong, Zehuan Yuan, Changhu Wang, and Ping Luo. Transtrack: Multiple-object tracking with transformer. *arXiv preprint arXiv:2012.15460*, 2020. 2, 7

[28] Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, et al. Scalability in perception for autonomous driving: Waymo open dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2446–2454, 2020. 2

[29] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008. 6

[30] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017. 2

[31] Zhongdao Wang, Liang Zheng, Yixuan Liu, Yali Li, and Shengjin Wang. Towards real-time multi-object tracking. In *European Conference on Computer Vision*, pages 107–122. Springer, 2020. 2, 3

[32] Nicolai Wojke, Alex Bewley, and Dietrich Paulus. Simple online and realtime tracking with a deep association metric. In *IEEE international conference on image processing*, pages 3645–3649, 2017. 1, 3, 7

[33] Jialian Wu, Jiale Cao, Liangchen Song, Yu Wang, Ming Yang, and Junsong Yuan. Track to detect and segment: An online multi-object tracker. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12352–12361, 2021. 1, 2, 7

[34] Ning Xu, Linjie Yang, Yuchen Fan, Dingcheng Yue, Yuchen Liang, Jianchao Yang, and Thomas Huang. Youtube-vos: A large-scale video object segmentation benchmark. *arXiv preprint arXiv:1809.03327*, 2018. 2

[35] Alper Yilmaz, Omar Javed, and Mubarak Shah. Object tracking: A survey. *Acm computing surveys (CSUR)*, 38(4):13–es, 2006. 1

[36] Fisher Yu, Wenqi Xian, Yingying Chen, Fangchen Liu, Mike Liao, Vashisht Madhavan, and Trevor Darrell. Bdd100k: A diverse driving video database with scalable annotation tooling. *arXiv preprint arXiv:1805.04687*, 2(5):6, 2018. 1, 2, 5

[37] Fangao Zeng, Bin Dong, Tiancai Wang, Cheng Chen, Xiangyu Zhang, and Yichen Wei. Motr: End-to-end multiple-object tracking with transformer. *arXiv preprint arXiv:2105.03247*, 2021. 2, 7

[38] Yifu Zhang, Peize Sun, Yi Jiang, Dongdong Yu, Zehuan Yuan, Ping Luo, Wenyu Liu, and Xinggang Wang. Bytetrack: Multi-object tracking by associating every detection box. *arXiv preprint arXiv:2110.06864*, 2021. 2, 3, 7, 8

[39] Yifu Zhang, Chunyu Wang, Xinggang Wang, Wenjun Zeng, and Wenyu Liu. Fairmot: On the fairness of detection and re-identification in multiple object tracking. *International Journal of Computer Vision*, 129(11):3069–3087, 2021. 1, 2, 3, 7

[40] Xingyi Zhou, Vladlen Koltun, and Philipp Krähenbühl. Tracking objects as points. In *European Conference on Computer Vision*, pages 474–490. Springer, 2020. 2, 7

[41] Xingyi Zhou, Dequan Wang, and Philipp Krähenbühl. Objects as points. *arXiv preprint arXiv:1904.07850*, 2019. 5, 8