2019 International Conference on Robotics and Automation (ICRA)
Palais des congres de Montreal, Montreal, Canada, May 20-24, 2019

BLVD: Building A Large-scale 5D Semantics Benchmark for Autonomous Driving

Jianru Xue[1], Jianwu Fang[1,2], Tao Li[1], Bohua Zhang[1], Pu Zhang[1], Zhen Ye[1] and Jian Dou[1]

***Abstract*— In autonomous driving community, numerous benchmarks have been established to assist the tasks of 3D/2D object detection, stereo vision, semantic/instance segmentation. However, the more meaningful dynamic evolution of the surrounding objects of ego-vehicle is rarely exploited, and lacks a large-scale dataset platform. To address this, we introduce `BLVD`, a large-scale 5D semantics benchmark which does not concentrate on the static detection or semantic/instance segmentation tasks tackled adequately before. Instead, BLVD aims to provide a platform for the tasks of dynamic 4D (3D+temporal) tracking, 5D (4D+interactive) interactive event recognition and intention prediction. This benchmark will boost the deeper understanding of traffic scenes than ever before. We totally yield $249,129$ 3D annotations, $4,902$ independent individuals for tracking with the length of overall $214,922$ points, $6,004$ valid fragments for 5D interactive event recognition, and $4,900$ individuals for 5D intention prediction. These tasks are contained in four kinds of scenarios depending on the object density (low and high) and light conditions (daytime and nighttime). The benchmark can be downloaded from our project site `https://github.com/VCCIV/BLVD/`.**

I. INTRODUCTION

Developing a safer, agiler and more dexterous autonomous vehicle with excellent ability of traffic scene understanding is the focus of much recent research in modern computer vision and robotic applications. Facing this urgent demand, many benchmarks have been constructed [1] and the research progress is heavily linked with them. Two classic and attractive ones are KITTI Vision Benchmark Suite [2] and Cityscapes [3], where KITTI was designed to test the functions of 2D/3D object detection, depth recovery, road segmentation, scene flow and optical flow, and Cityscapes were built to evaluate the semantic/instance segmentation performance, which derived the Citypersons [4] for video-level person detection.

Although these benchmarks have pushed the autonomous driving research forward largely, they are all collected from day and sunny time and with limited scale. Some institutions aim to build larger and more challenging benchmarks. For example, Berkeley Deep Drive [5] collected 100K color videos under various environments, different cities, diverse road conditions, where each video labeled 4 frames for every 10 seconds with instance label. Apollo scape [6] built a benchmark with 147K images with fine instance annotation. Although these benchmarks are becoming larger and larger, many of them still concentrate on the static images or frames without continuous labeling, and do not label the large-scale 3D data [5], [7], [8]. In addition, the representation based on these vision datasets only can get static scene layout [9] and sparse moving pattern [10] of the observed scene. However, the more meaningful 3D dynamic evolution of the surrounding objects of ego-vehicle when driving was rarely exploited, and lacks a large-scale platform with unifying metrics for performance evaluation [11], [12], [13], [14].

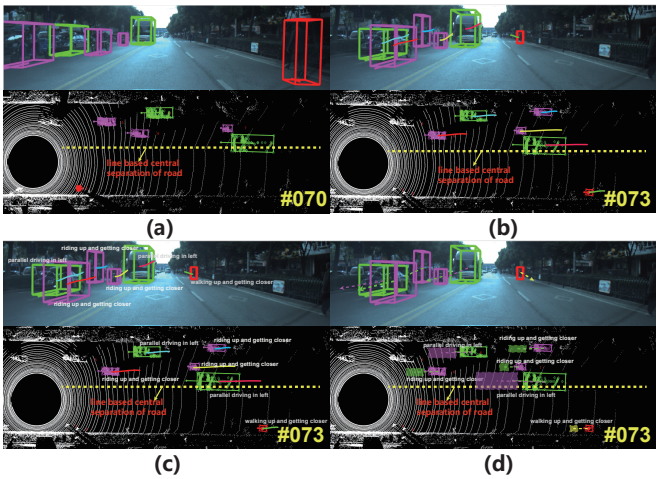

Fig. 1. The task flow of BLVD with 5D semantic annotations. (a) denotes a frame with static 3D bounding boxes, (b) is another frame where the 4D trajectories of objects are shown, (c) specifies the frame where the 5D interactive event type is assigned in each trajectory, and (d) demonstrates the 5D intention prediction with the prediction of location, geometry structure of 3D bounding boxes, orientation and interactive event state.

In this paper, we **b**uild a **l**arge-scale **5D** semantics benchmark (BLVD), specifically tailored on the tasks of 4D (3D+temporal) tracking, 5D (4D+interactive) interactive event recognition and intention prediction in autonomous driving. We defined three kinds of participants, including *vehicles*, *pedestrians* and *riders*, where riders contains cyclists and motorbikes which always demonstrate ruleless moving. The benchmark is constructed by a self-driving platform providing multiple kinds of sensors for surrounding perception, including a Velodyne HDL-64E LIDAR scanner, a GPS/inertial system, two multi-view cameras with high resolution. It is worth noting that all the sensors are registered and synchronized automatically. Different from many other datasets, this benchmark is collected under different driving scenarios (urban and highway), various light conditions (day-

*This work was supported the National Key R&D Program Project of China (No. 2016YFB1001004), National Natural Science Foundation of China (No. 61751308, 61773311 and 61603057), and China Postdoctoral Science Foundation (No. 2017M613152).

[1]The authors are with the Institute of Artificial Intelligence and Robotics, Xian Jiaotong University, Xi'an, China. `jrxue@mail.xjtu.edu.cn`

[2]Jianwu Fang is also with the School of Electronic and Control Engineering, Chang'an University, Xi'an, China. `j.w.fangit@gmail.com`

time and nighttime), and with full 5D semantics annotation for 120K frames. BLVD benchmark exceeds the previous efforts to deeper traffic scene understanding, in terms of annotation size, light richness, and tasks. In particular, the large-scale 4D participant trajectories and 5D interaction between surrounding objects with ego-vehicle are defined in the vehicle-based coordinate system. One typical snapshot representing a task flow from static 3D annotations to 5D intention prediction of our benchmark is shown in Fig. 1.

II. BLVD BENCHMARK

A. Sensors and Acquisitions

Our dataset aims at deeper understanding for the dynamic traffic scene. Multiple kinds of sensors are equipped for surrounding perception, including a Velodyne HDL-64E LIDAR scanner (10Hz, 64 laser beams, range of 100m), a GPS/inertial system, two multi-view color cameras with high resolution (30Hz, resolution: 1920×500 pixels larger than 1242×375 of KITTI [2]). The sensors are mounted on the top of our vehicle, where the cameras are built-in a box with a windshield and capture the front view of road. Velodyne HDL-64E unit can provide accurate 3D information from moving platforms. The egomotion in the 3D laser measurements is compensated by our GPS/IMU system. Different from the data collection way from multiple cities (e.g., Cityscapes [4]), we gather the data in a same city (Changshu, Jiangsu province, China) under different light conditions (daytime and nighttime), diverse densities (low and high density), and distinct scenarios (highway and urban), where the dynamic participants are fully annotated in 3D mode. In this benchmark, we design a fast online calibration method which can efficiently register and synchronize camera and 3D LIDAR, and costs only one day for the accurate calibration of 120k frames. Consequently, we obtain 654 calibrated video clips with the frequency of 10Hz, which contain images and 3D point cloud simultaneously.

B. 5D Semantics Specifications

In the annotation process, we locate in vehicle-based coordinates system. Unlike most of benchmarks relying online crowd-sourcing labeling, we hired a set of annotators to assign 5D semantics for each individual, and frequently gathered the feedbacks of each annotator to form many consistent and reasonable labeling principles.

Our benchmark aims to assist the ego-vehicle to comprehend its driving scene and make effective decision. Our benchmark removed the perception range which is not helpful for driving decision of ego-vehicle, such as the zones outside a guardrail. This strategy is different from other detection benchmarks which labeled all participants in the image. Similar to KITTI, we annotated the participants that appear both in images and 3D point clouds within 50 meters in the front view. We developed a semi-automatic tool for fast annotation, where the corresponding image and its bird's eye view (BEV) of 3D point clouds were shown in the interface.

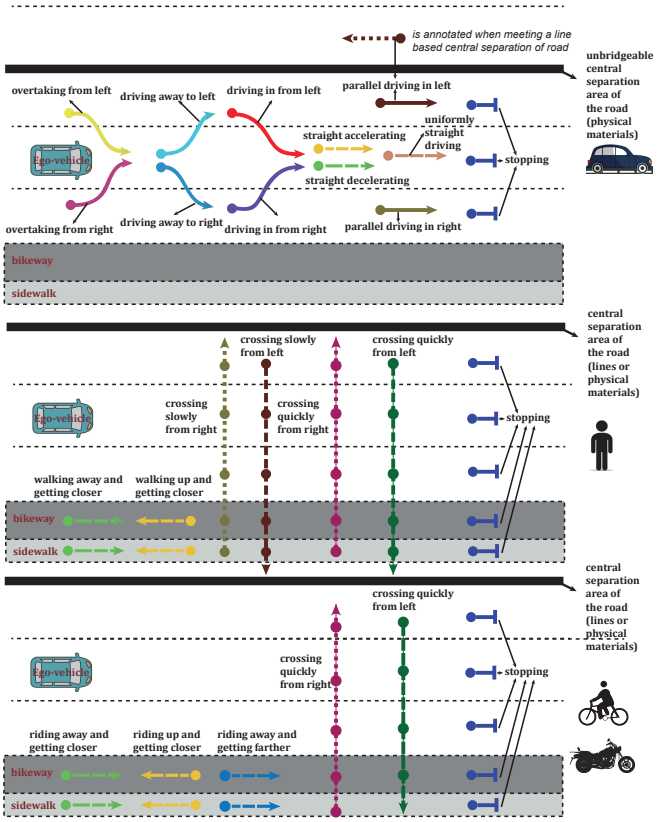

Fig. 2. The illustration of interactive events standing at vehicle-based coordinate system. From top to bottom, the event types of vehicles, pedestrians and riders are demonstrated. Note that, there is an extra event type of participants (specified as "*others*") for denoting the ambiguous interactive event.

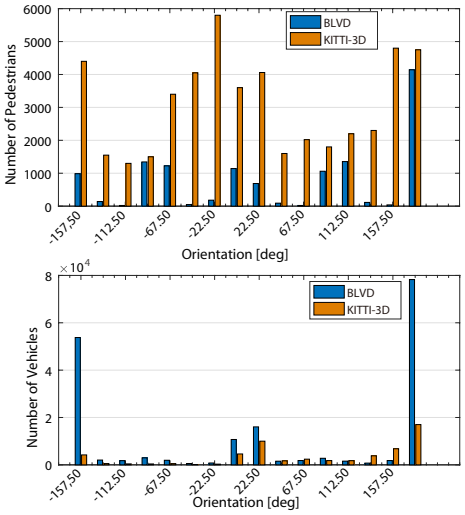

Fig. 3. The orientation distribution of BLVD and KITTI-3D [2].

3D bounding box: Each 3D bounding box is labeled by manually click four points (top-left point, top-right point, bottom-right point and an arbitrary point for determining orientation) on the BEV of 3D point clouds shown in screen. After these operations, the length, width and orientation of a 3D bounding box is automatically computed and stored. The height of the bounding box is automatically determined

TABLE I

TRAINING AND TESTING SET STATISTIC W.R.T., NI (NUMBER OF INDIVIDUALS) AND TTL (TOTAL TRAJECTORY LENGTH) UNDER FOUR KINDS OF SCENE CONDITIONS OF I (DAYTIME WITH LOW DENSITY), II (DAYTIME WITH HIGH DENSITY), III (NIGHTTIME WITH LOW DENSITY) AND IV (NIGHTTIME WITH HIGH DENSITY).

Classes	Data Splits	I		II		III		IV		Total	
		NI	TTL	NI	TTL	NI	TTL	NI	TTL	NI	TTL
Pedestrians	Training	28	819	61	2,890	30	849	41	4,154	160	8,172
	Testing	30	790	60	1,808	23	517	44	1,838	157	4,953
Vehicles	Training	433	16,114	546	24,795	545	23,423	215	17,209	1,739	81,541
	Testing	355	14,242	442	21,162	659	30,069	276	9,533	1,732	75,006
Riders	Training	121	4,790	250	10,197	113	3,272	79	5669	563	23,928
	Testing	122	4,004	205	8,349	116	3,242	108	5,187	551	20,782

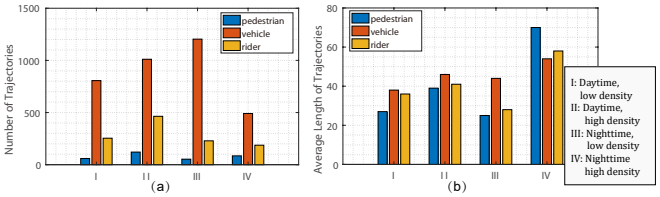

Fig. 4. The statistics of trajectories. (a) denotes the number of individuals under four different scene conditions and (b) is the average trajectory length.

by finding the minimum value enclosing most of the 3D points in the 3D box. Then, we project the 3D point cloud back to image plane and visualize it for a double check. All the labeled 3D annotations can be easily modified for guaranteing accuracy. The benchmark is fully annotated, which yields $214,922$ 3D annotations comprising $179,073$ vehicles, $12,599$ pedestrians, and $51,917$ riders. We compare BLVD with the ones of KITTI in terms of orientation distribution of vehicles and pedestrians, as shown in Fig. 3. The vehicle annotations in BLVD has more complexity than KITTI-3D. BLVD benchmark aims to provide a platform for deeper traffic scene understanding, we focused on the individuals which may interact with the ego-vehicle. Therefore, we did not care the pedestrians outside the useful perception range.

4D object IDs: When assigning the IDs for each individual, we adopt the principle that the first frame must be labeled carefully as much as possible, where the 3D bounding boxes, orientations, IDs and object classes are accurately initialized. Then, for subsequent frames, we only need to drag each box to its new location with a tiny modification of orientation and interactive event state. This strategy can largely boost the annotation efficiency. We assign the same ID for the object which has appeared before and re-appears again in the same video clip, and add new ID for newly observed objects. The IDs increase till the end of the video clip, and are re-assigned for a new clip from 1. We obtain $4,902$ valid individuals with the trajectory length of overall $214,922$ points.

5D interactive event: In driving, perceiving the interactive events of other participants to ego-vehicle is necessary for making a reasonable decision. We defined 13 kinds of events, 8 kinds of events and 7 kinds of events for vehicles, pedestrians and riders, respectively. For a clearer understanding, the types of events are demonstrated in Fig. 2. Note that, we assign these event types to each point of trajectories. Each kind of event here corresponds its reasonable road location that the participant may appear and is determined by multiple trajectory points. Additionally, this benchmark also labeled 8 kinds of events of ego-vehicle. They are denoted as: *straight accelerating*, *straight decelerating*, *turning right*, *turning left*, *uniformly straight driving*, *changing line to left*, *changing line to right*, and *stopping*. We obtain $6,004$ valid event fragments of surrounding participants.

5D intention: 5D intention prediction inherits the 4D trajectories. Different from the location based intention works [15], [16], we advocate a prediction of locations, event types, geometrical structures of 3D bounding boxes, and orientations.

C. Dataset Splits

We split the data as training and testing sets, involving a balanced distribution regarding following properties of equal shares: 1) light conditions (daytime and nighttime), 2) participants densities (low and high), 3) participant category (pedestrians, vehicles and riders), and 4) event types. The statistics of BLVD will be demonstrated in following section.

III. BENCHMARKING

There is no publicly available dataset like BLVD involving 4D tracking, 5D interactive event recognition, and 5D intention prediction simultaneously. We will analyze the statistics of BLVD and compare it with other benchmarks when focusing certain tasks. Additionally, we provide the metrics for the performance evaluation of each task.

A. 4D Tracking

The first task is to track multiple 3D individuals in one video clip. In the literatures, the works for multi-object tracking mainly are based on RGB videos [17], stereo sequential images [18] or LiDAR point sequence [19]. Recently, Frossard and Urtasun [20] addressed the object tracking by fusing RGB image and 3D point cloud on KITTI benchmark which is only with 40 sequences for 4D tracking. We contribute a larger one owning 654 video sequences. In our benchmark, we separate the trajectories as the ones of vehicles, pedestrians and riders.

Statistic analysis: A trajectory is valid when its length is larger than 10 frames whatever the participant is. The statistics of trajectories with respect to lighting conditions,

participant densities and classes are analyzed in Fig. 5. With the condition of participant density and light, we gathered 164, 283, 114, 93 video clips under daytime with low density, nighttime with low density, daytime with high density and nighttime with high density, respectively. The corresponding average numbers of individuals are 10.7, 5.3, 14.0 and 8.2, respectively. Because the number of video clips with low density is larger than the ones of high density, the videos of low density have more individuals than the ones of high density. To depress this imbalance, we make an equal distribution for data split. The statistics for training set and testing set are listed in Table. I.

Metrics: To assess the performance of tracking hypothesis, we rely on the metrics evaluating both accuracy and efficiency. The common known metrics for multiple object tracking are launched by MOT challenge[1]. For 3D object tracking, multiple object tracking accuracy (MOTA) [21], multiple object tracking precision (MOTP) [21], ID F1 Score (IDF1) [22], mostly tracked targets (MT), mostly lost targets (ML), total number of false positives (FP), total number of missed targets (FN), total number of identity switches (ID Sw) [23], total number of times of trajectory fragmenting (Frag) and processing speed (Hz) all can be used, where the most important metrics for evaluating accuracy are MOTA and MOTP. MOTA combines three error sources of false positives, missed targets and identity switches, and MOTP measures the misalignment between the annotated and the predicted bounding boxes. The detailed meaning of each metric can be referred to the official website of 3D tracking of MOT challenge.

B. 5D Interactive Event Recognition

For understanding the dynamic evolution of traffic scene, event recognition or reasoning is a core problem, because it can reflect the dynamic evolution process of scene with tractable reasoning strategy [24]. As far as we know, we are the first attempt to launch the 5D interactive event recognition platform. Accurate interactive event recognition can supply a powerful information for path planning and motion decision.

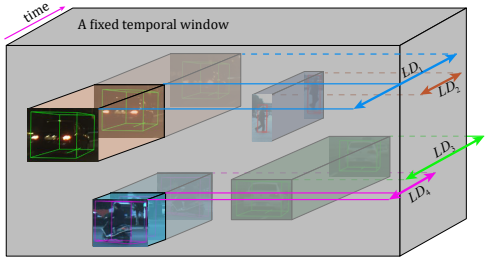

Fig. 5. The illustration for individuals with different length of duration. LD_1 and LD_4 links some frames before entering a fixed temporal window, LD_2 links the frames falling into the fixed temporal window, and LD_3 associates some frames exceeding the fixed temporal window.

Statistic analysis: As aforementioned, we totally obtain 4902 individuals with valid trajectories. We assigned interactive event type into each point of trajectories. Actually, each kind of interactive event has different temporal durations, as shown by the example in Fig. 5 with a reference of a fixed temporal window. In one trajectory, there may be multiple kinds of interactive events.

Apparently, it is difficult to recognize the interactive events of all the individuals within a fixed temporal window. Therefore, this benchmark paves this task as recognizing the *interactive event state* in each point lying on individuals' trajectories, and formulates it as a **sequence to sequence recognition** problem. In this problem, the recognition of an event state relies on multiple nearly observed trajectory points. Hence, the input sequence should be longer than the output one. The detailed illustration is demonstrated in Fig. 6. Under this problem setting, we fix the minimum number

[1]https://motchallenge.net/

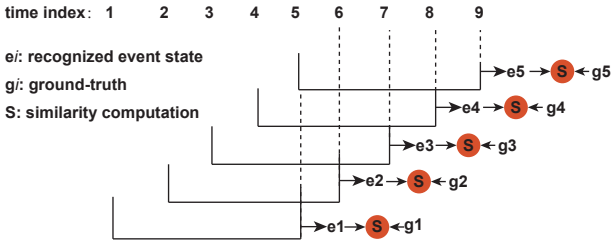

Fig. 6. A schematic example for recognizing an event fragment with 5 nodes of state chain.

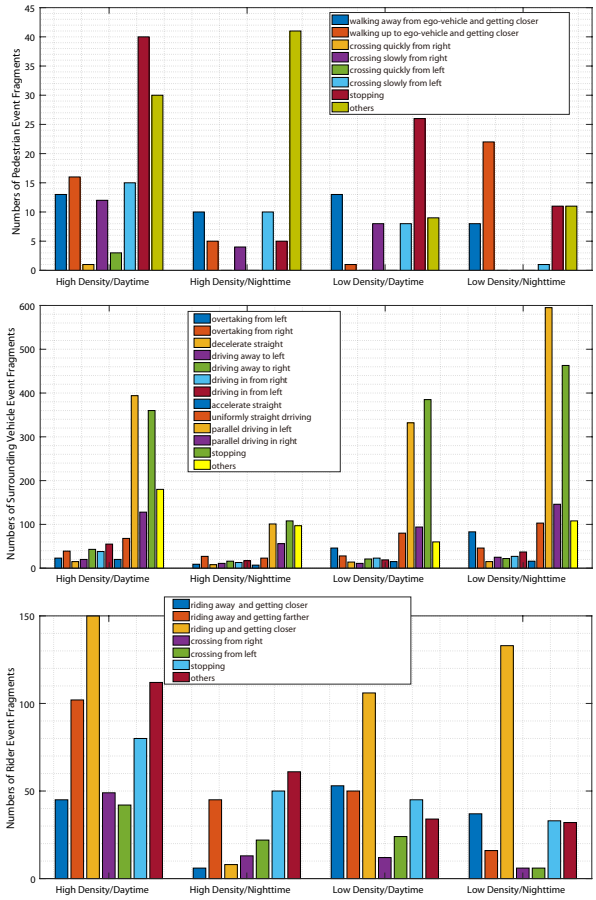

Fig. 7. Statistics of interactive event type of pedestrians, vehicles and riders in terms of light condition and participant density.

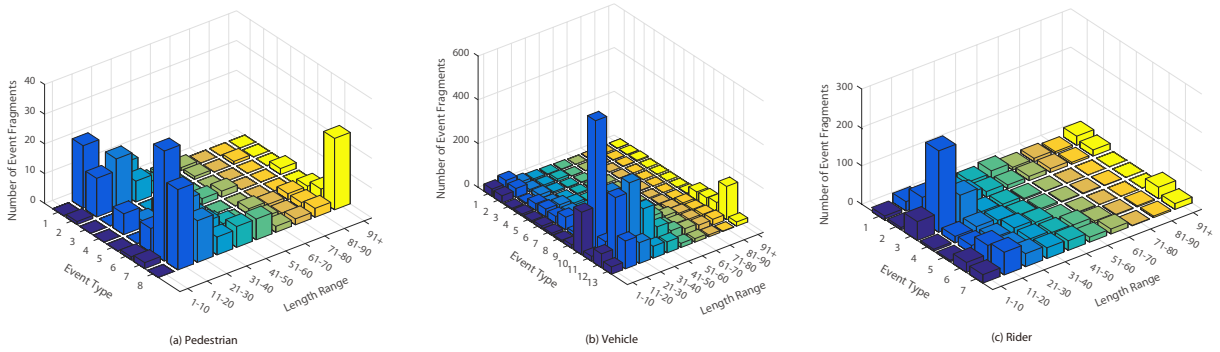

Fig. 8. Statistics of the number of event fragments with respect to event type and and length range of fragments. The event type indexes in axis are the same as ones in Table. II.

TABLE II

PERFORMANCE MATRIX FOR INTERACTIVE EVENT RECOGNITION OF VEHICLES, PEDESTRIANS AND RIDERS UNDER FOUR KINDS OF SCENE CONDITIONS OF I (DAYTIME WITH LOW DENSITY), II (DAYTIME WITH HIGH DENSITY), III (NIGHTTIME WITH LOW DENSITY) AND IV (NIGHTTIME WITH HIGH DENSITY).

Interactive Event Type	I	II	III	IV	Precision	Recall
vehicle overtaking from left	*TP/FP/FN*	*TP/FP/FN*	*TP/FP/FN*	*TP/FP/FN*	*P*	*R*
vehicle overtaking from right	*TP/FP/FN*	*TP/FP/FN*	*TP/FP/FN*	*TP/FP/FN*	*P*	*R*
vehicle driving away to left	*TP/FP/FN*	*TP/FP/FN*	*TP/FP/FN*	*TP/FP/FN*	*P*	*R*
vehicle driving away to right	*TP/FP/FN*	*TP/FP/FN*	*TP/FP/FN*	*TP/FP/FN*	*P*	*R*
vehicle driving in from left	*TP/FP/FN*	*TP/FP/FN*	*TP/FP/FN*	*TP/FP/FN*	*P*	*R*
vehicle driving in from right	*TP/FP/FN*	*TP/FP/FN*	*TP/FP/FN*	*TP/FP/FN*	*P*	*R*
vehicle parallel driving in left	*TP/FP/FN*	*TP/FP/FN*	*TP/FP/FN*	*TP/FP/FN*	*P*	*R*
vehicle parallel driving in right	*TP/FP/FN*	*TP/FP/FN*	*TP/FP/FN*	*TP/FP/FN*	*P*	*R*
vehicle straight accelerating	*TP/FP/FN*	*TP/FP/FN*	*TP/FP/FN*	*TP/FP/FN*	*P*	*R*
vehicle straight decelerating	*TP/FP/FN*	*TP/FP/FN*	*TP/FP/FN*	*TP/FP/FN*	*P*	*R*
vehicle uniformly straight driving	*TP/FP/FN*	*TP/FP/FN*	*TP/FP/FN*	*TP/FP/FN*	*P*	*R*
vehicle stopping	*TP/FP/FN*	*TP/FP/FN*	*TP/FP/FN*	*TP/FP/FN*	*P*	*R*
walking/riding away and getting closer	*TP/FP/FN*	*TP/FP/FN*	*TP/FP/FN*	*TP/FP/FN*	*P*	*R*
walking/riding up and getting closer	*TP/FP/FN*	*TP/FP/FN*	*TP/FP/FN*	*TP/FP/FN*	*P*	*R*
riding away and getting farther	*TP/FP/FN*	*TP/FP/FN*	*TP/FP/FN*	*TP/FP/FN*	*P*	*R*
pedestrian/rider crossing slowly from right	*TP/FP/FN*	*TP/FP/FN*	*TP/FP/FN*	*TP/FP/FN*	*P*	*R*
pedestrian/rider crossing slowly from left	*TP/FP/FN*	*TP/FP/FN*	*TP/FP/FN*	*TP/FP/FN*	*P*	*R*
pedestrian crossing quickly from right	*TP/FP/FN*	*TP/FP/FN*	*TP/FP/FN*	*TP/FP/FN*	*P*	*R*
pedestrian crossing quickly from left	*TP/FP/FN*	*TP/FP/FN*	*TP/FP/FN*	*TP/FP/FN*	*P*	*R*
pedestrian/rider stopping	*TP/FP/FN*	*TP/FP/FN*	*TP/FP/FN*	*TP/FP/FN*	*P*	*R*
others (vehicles/pedestrians/riders)	*TP/FP/FN*	*TP/FP/FN*	*TP/FP/FN*	*TP/FP/FN*	*P*	*R*
Precision	*P*	*P*	*P*	*P*		
Recall	*R*	*R*	*R*	*R*		

of points for recognizing the first point of output sequence as 5, and the minimum number of points in output sequence as 3. Consequently, the trajectories with the length larger than 8 are valid for this task, and yield $6,004$ fragments for interactive event recognition. The statistics of event types in terms of participant category, light condition and participant density are demonstrated in Fig. 7. Additionally, we also analyzed the number of event fragments in relative to the event type and and length range of each fragment. The analyzed statistics are shown in Fig. 8. In this benchmark, the splitting of event recognition data inherits the principle of the one of trajectories, i.e, the training set and testing set of interactive event are taken out from the training and testing trajectories, as analyzed in Table. I.

Metrics: We formulate the 5D interactive event recognition task as a sequence to sequence recognition problem. The performance evaluation depends on the precision $P = TP/(TP + FP)$ and recall $R = TP/(TP + FN)$ of the recognized event states, where TP is true positives, FP is false positives, and FN is false negatives. These metrics are also adopted by [25], [26] for trajectory-based event recognition. Specially, the performance evaluation of our benchmark will be carried out on different situations, including low-density/daytime, high-density/daytime, low-density/nighttime and high-density/nighttime, and can be clearly visualized by filling out Table. II.

C. 5D Intention Prediction

In autonomous driving, the purpose of aforementioned 4D object tracking and 5D interactive event recognition is to provide an accurate reasoning clue for future driving, i.e., to make a precise movement prediction of surrounding participants in the future. Thus, ego-vehicle can smoothly pass the observed scene and reach the destination as fast as possible. 5D intention prediction is another task launched for the first time by our benchmark. Previous intention works [15], [16] in computer vision concentrate on predicting the future time-step locations of the target. Commonly, they evaluate the performance by two metrics proposed in [15]: average displacement error (ADE) and final displacement error (FDE), where ADE denotes the Euclidean distance

TABLE III

THE RESULTS OF FIVE-FIVE PREDICTION, TEN-TEN PREDICTION AND TEN-FIVE PREDICTION. *NTrainT* IS THE NUMBER OF TRAINING TRAJECTORIES, AND *NTestT* DENOTES THE NUMBER OF TESTING TRAJECTORIES.

	Classes	NTrainT	NTestT	ADE/meters	FDE/meters
Five-Five	Pedestrians	160	157	0.34	0.49
	Vehicles	1,739	1,732	0.47	0.69
	Riders	563	557	0.44	0.66
Ten-Ten	Pedestrians	122	94	0.44	0.79
	Vehicles	1,220	1,197	0.6	1.12
	Riders	344	357	0.65	1.26
Ten-Five	Pedestrians	137	129	**0.27**	**0.39**
	Vehicles	1,450	1,456	**0.38**	**0.57**
	Riders	441	447	**0.38**	**0.56**

between the predicted trajectory and the actual trajectory averaged over all time-steps for all targets, and FDE specifies the average Euclidean distance between the predicted trajectory point and the actual trajectory point at the end of n time steps. ADE and FDE are computed as:

$$ADE = \frac{\sum_{i=1}^{N}\sum_{m=1}^{M}\sum_{t=t_{obs}+1}^{t_{obs}+t_{pred}}\sqrt{(x_t^i-\hat{x}_t^i)^2+(y_t^i-\hat{y}_t^i)^2}}{(N+M)t_{pred}},$$
$$FDE = \frac{\sum_{i=1}^{N}\sum_{m=1}^{M}\sqrt{(x_{t_{pred}}^i-\hat{x}_{t_{pred}}^i)^2+(y_{t_{pred}}^i-\hat{y}_{t_{pred}}^i)^2}}{(N+M)t_{pred}}, \tag{1}$$

where (x_i, y_i) and $(\hat{x}_i, \hat{y}_i)$ are the locations of the i^{th} observed point (ground-truth) and its predicted one, N is the number of trajectories, M is the number of batches after partitioning the trajectory into some equal fragments, t_{obs} and t_{pred} are the numbers of observed frames and the ones to be predicted, respectively.

Controlled experiments: Similarly, we have conducted three groups of of experiments on location based prediction: 1) predicting future 5 frames with observed 5 frames (five-five prediction), 2) predicting future 10 frames with observed 10 frames (ten-ten prediction), and 3) predicting future 5 frames with observed 10 frames (ten-tive prediction), where the location of each point is the center of 3D bounding box in 3D point cloud. Consequently, the trajectories with over 10 frames are valid for the first setting, and 20 frames for the second one. The controlled experiments follow the Long Short-Term Memory (LSTM) network [27], [15]. The experimental details are as follows.

We partition the trajectories as many batches with the total number of observed frames and the ones to be predicted, and then pass the batches to train a LSTM network. The training and testing sets inherit the statistics of Table. I. We first generate the vectors with fixed length (set as 32) by embedding the location of each annotation using a linear layer with relu nonlinearity. These embeddings are utilized as the input to the LSTM cell. The hidden state of LSTM is set as 64, and the output of LSTM network utilizes a linear transform of 64 to 2 (location coordinates). We trained the network in 10 epoch with the learning rate of 0.001. The experimental results are demonstrated in Table IV. We can see that longer prediction cause larger predicted error, and longer observation will generate more accurate prediction.

Metrics for 5D intention prediction: Actually, in our benchmark, we provide not only the location prediction task, the event state chain, geometrical structure of 3D bounding boxes, and orientations should all be considered in prediction. Therefore, beside ADE and FDE in [15]. The precision and recall evaluation for predicted event state (stated in the metrics of 5D interactive event recognition), average precision (AP) of predicted 3D bounding boxes and orientation are advocated. Following ADE and FDE, we evaluate the prediction of 3D boxes and orientation as:

$$ADE_{orientation} = \frac{\sum_{i=1}^{N}\sum_{m=1}^{M}\sum_{t=t_{obs}+1}^{t_{obs}+t_{pred}}\delta_t^i\left(1+\cos(\theta_t^i-\hat{\theta}_t^i)\right)/2}{(N+M)t_{pred}},$$
$$ADE_{3Dbox} = \frac{\sum_{i=1}^{N}\sum_{m=1}^{M}\sum_{t=t_{obs}+1}^{t_{obs}+t_{pred}}s(V_t^i,\hat{V}_t^i)}{(N+M)t_{pred}}, \tag{2}$$

and

$$FDE_{orientation} = \frac{\sum_{i=1}^{N}\sum_{m=1}^{M}\delta_t^i\left(1+\cos(\theta_{t_{pred}}^i-\hat{\theta}_{t_{pred}}^i)\right)/2}{(N+M)t_{pred}},$$
$$FDE_{3Dbox} = \frac{\sum_{i=1}^{N}\sum_{m=1}^{M}s(V_{t_{pred}}^i,\hat{V}_{t_{pred}}^i)}{(N+M)t_{pred}}, \tag{3}$$

where θ and $\hat{\theta}$ are the observed orientations of targets and their predicted ones, and $s(V, \hat{V})$ computes the overlapping rate of two 3D bounding boxes V and the predicted $\hat{V}$ by PASCAL VOC criterion [28]. δ penalizes the mis-predictions. If the predicted 3D bounding box overlaps the ground truth at least 50%, $\delta = 1$ and 0 vice versa. In these metrics, ADE_{3Dbox} and FDE_{3Dbox} enjoy a larger value for better performance, and others pursuit smaller value. Therefore, our benchmark provide a platform for more challenging intention prediction task.

IV. CONCLUSIONS

In this paper, we built a large-scale 5D semantics benchmark for autonomous driving which was captured under a wide range of interesting scenarios, and calibrated, synchronized and rectified efficiently and accurately. Different from the previously static detection/segmentation tasks, we focused on the deeper understanding of traffic scenes. Specifically, the tasks of 4D tracking, 5D interactive event recognition, and 5D intention prediction were launched in this benchmark. With the careful annotation, the benchmark yielded $249,129$ 3D annotations, $4,902$ independent instances for tracking with the length of overall $214,922$ points, $6,004$ 3D annotations for 5D interactive event recognition, and $4,900$ individuals for 5D intention prediction. These annotations were gathered under different light conditions (daytime and nighttime), diverse density of participants (low density and high density) and distinct driving scenarios (highway and urban). We believe that this benchmark will be highly useful in robotics and computer vision fields. In the future, we will embrace the 3D detection task, and make the task flow as an integrated chain, where each task can promote following ones. In addition, we will balance the annotations for a better utilization.

REFERENCES

[1] H. Yin and C. Berger, "When to use what data set for your self-driving car algorithm: An overview of publicly available driving datasets," in *IEEE International Conference on Intelligent Transportation Systems*, 2017, pp. 1–8.

[2] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets robotics: The kitti dataset," *International Journal of Robotics Research*, vol. 32, no. 11, pp. 1231–1237, 2013.

[3] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 3213–3223.

[4] S. Zhang, R. Benenson, and B. Schiele, "Citypersons: A diverse dataset for pedestrian detection," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 4457–4465.

[5] F. Yu, W. Xian, Y. Chen, F. Liu, M. Liao, V. Madhavan, and T. Darrell, "BDD100K: A diverse driving video database with scalable annotation tooling," *CoRR*, vol. abs/1805.04687, 2018.

[6] X. Huang, X. Cheng, Q. Geng, B. Cao, D. Zhou, P. Wang, Y. Lin, and R. Yang, "The apolloscape dataset for autonomous driving," *CoRR*, vol. abs/1803.06184, 2018.

[7] G. Neuhold, T. Ollmann, S. R. Bulò, and P. Kontschieder, "The mapillary vistas dataset for semantic understanding of street scenes," in *IEEE International Conference on Computer Vision*, 2017, pp. 5000–5009.

[8] G. Ros, L. Sellart, J. Materzynska, D. Vázquez, and A. M. López, "The SYNTHIA dataset: A large collection of synthetic images for semantic segmentation of urban scenes," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 3234–3243.

[9] C. Landsiedel and D. Wollherr, "Road geometry estimation for urban semantic maps using open data," *Advanced Robotics*, vol. 31, no. 5, pp. 282–290, 2017.

[10] A. Geiger, M. Lauer, C. Wojek, C. Stiller, and R. Urtasun, "3D traffic scene understanding from movable platforms." *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 5, pp. 1012–1025, 2014.

[11] J. R. Xue, J. W. Fang, and P. Zhang, "A survey of scene understanding by event reasoning in autonomous driving," *International Journal of Automation and Computing*, vol. 15, no. 3, pp. 1–18, 2018.

[12] W. Yao, Q. Zeng, Y. Lin, D. Xu, H. Zhao, F. Guillemard, S. Geronimi, and F. Aioun, "On-road vehicle trajectory collection and scene-based lane change analysis part ii," *IEEE Transactions on Intelligent Transportation Systems*, vol. 18, no. 1, pp. 206–220, 2017.

[13] S. Ernst, J. Rieken, and M. Maurer, "Behaviour recognition of traffic participants by using manoeuvre primitives for automated vehicles in urban traffic," in *Proc. IEEE Conf. Intelligent Transportation Systems*, 2016.

[14] F. Schneemann and P. Heinemann, "Context-based detection of pedestrian crossing intention for autonomous driving in urban environments," in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2016, pp. 2243–2248.

[15] A. Alahi, K. Goel, V. Ramanathan, A. Robicquet, F. F. Li, and S. Savarese, "Social lstm: Human trajectory prediction in crowded spaces," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 961–971.

[16] A. Vemula, K. Mülling, and J. Oh, "Social attention: Modeling attention in human crowds," *CoRR*, vol. abs/1710.04689, 2017.

[17] S. Sharma, J. A. Ansari, J. K. Murthy, and K. M. Krishna, "Beyond pixels: Leveraging geometry and shape cues for online multi-object tracking," *CoRR*, vol. abs/1802.09298, 2018.

[18] V. Romero-Cano, G. Agamennoni, and J. Nieto, "A variational approach to simultaneous multi-object tracking and classification," *The International Journal of Robotics Research*, vol. 35, no. 6, pp. 654–671, 2016.

[19] H. Wang, B. Wang, B. Liu, X. Meng, and G. Yang, "Pedestrian recognition and tracking using 3d lidar for autonomous vehicle," *Robotics and Autonomous Systems*, vol. 88, pp. 71–78, 2017.

[20] D. Frossard and R. Urtasun, "End-to-end learning of multi-sensor 3d tracking by detection," *CoRR*, vol. abs/1806.11534, 2018.

[21] K. Bernardin and R. Stiefelhagen, "Evaluating multiple object tracking performance: the clear mot metrics," *Eurasip Journal on Image and Video Processing*, vol. 2008, no. 1, pp. 246–309, 2008.

[22] E. Ristani, F. Solera, R. Zou, R. Cucchiara, and C. Tomasi, "Performance measures and a data set for multi-target, multi-camera tracking," in *European Conference on Computer Vision*, 2016, pp. 17–35.

[23] Y. Li, C. Huang, and R. Nevatia, "Learning to associate: Hybridboosted multi-target tracker for crowded scene," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 2953–2960.

[24] D. L. Waltz, "Understanding scene descriptions as event simulations," in *Proc. Meeting on Association for Computational Linguistics*, 1980, pp. 7–11.

[25] M. S. Kristoffersen, J. V. Dueholm, R. K. Satzoda, M. M. Trivedi, A. M?gelmose, and T. B. Moeslund, "Towards semantic understanding of surrounding vehicular maneuvers: A panoramic vision-based framework for real-world highway studies," in *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2016, pp. 1584–1591.

[26] J. V. Dueholm, M. S. Kristoffersen, R. K. Satzoda, T. B. Moeslund, and M. M. Trivedi, "Trajectories and maneuvers of surrounding vehicles with panoramic camera arrays," *IEEE Transactions on Intelligent Vehicles*, vol. PP, no. 99, pp. 1–1, 2016.

[27] S. Hochreiter and J. Schmidhuber, "Long short-term memory." *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[28] M. Everingham, L. V. Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (VOC) challenge," *International Journal of Computer Vision*, vol. 88, no. 2, pp. 303–338, 2010.