

The H3D Dataset for Full-Surround 3D Multi-Object Detection and Tracking in Crowded Urban Scenes

Abhishek Patil, Srikanth Malla, Haiming Gang and Yi-Ting Chen

Abstract—3D multi-object detection and tracking are crucial for traffic scene understanding. However, the community pays less attention to these areas due to the lack of a standardized benchmark dataset to advance the field. Moreover, existing datasets (e.g., KITTI [1]) do not provide sufficient data and labels to tackle challenging scenes where highly interactive and occluded traffic participants are present. To address the issues, we present the Honda Research Institute 3D Dataset (H3D), a large-scale full-surround 3D multi-object detection and tracking dataset collected using a 3D LiDAR scanner. H3D comprises of 160 crowded and highly interactive traffic scenes with a total of 1 million labeled instances in 27,721 frames. With unique dataset size, rich annotations, and complex scenes, H3D is gathered to stimulate research on full-surround 3D multi-object detection and tracking. To effectively and efficiently annotate a large-scale 3D point cloud dataset, we propose a labeling methodology to speed up the overall annotation cycle. A standardized benchmark is created to evaluate full-surround 3D multi-object detection and tracking algorithms. 3D object detection and tracking algorithms are trained and tested on H3D. Finally, sources of errors are discussed for the development of future algorithms.

I. INTRODUCTION

Multi-object detection and tracking are two essential tasks for traffic scene understanding. The field has been significantly boosted by recent advances of deep learning algorithms [2], [3], [4], [5], [6] and an increasing number of datasets [7], [8], [9], [10], [11]. While tremendous progress has been made in 2D traffic scene understanding, it still suffers from the fundamental limitations in the sensing capability and lack of 3D information. Recently, with the emerging technology of 3D range scanners, the range sensor directly measure 3D distances by illuminating the environment with pulsed laser light. It enables a wide range of robotic applications in the 3D world. While 3D scene understanding is important for these applications, relatively small efforts [1], [12], [13], [14] have been attempted in comparison to its 2D counterpart.

The Oxford RobotCar dataset [12] was proposed to address the challenges of robust localization and mapping under significantly different weather and lighting conditions. Recently, Jeong et al. [14] introduced a complex urban LiDAR dataset collected in metropolitan areas, large building complexes, and underground parking lots. However, these datasets mainly focus on Simultaneous Localization and Mapping (SLAM).

The authors are with Honda Research Institute, 375 Ravendale Dr, Suite B, Mountain View, CA, 94043, USA. {apatil, smalla, hgang, ychen}@honda-ri.com



Fig. 1: The Honda Research Institute 3D Dataset (H3D) for full-surround 3D multi-object detection and tracking in crowded urban scenes.

Among the existing attempts, KITTI dataset [1] enables various scene understanding tasks including 3D object detection and tracking. Specifically, it comprises of more than 200k manually labeled 3D objects captured in cluttered scenes. However, KITTI dataset is insufficient to advance the future development of 3D multi-object detection and tracking for the following reasons. First, the 3D object annotations are only labeled in the frontal view that limits the applications required full-surround reasoning. Second, KITTI dataset has relatively simple scene complexity without extensive data from crowded urban scenes, e.g., metropolitan areas where highly interacting and occluding traffic participants are present. Third, the richness of existing labels in KITTI dataset is inadequate for deep learning algorithms to learn diverse appearances from data. Fourth, KITTI dataset does not have a standardized evaluation for full-surround multi-object detection and tracking in 3D.

To address the aforementioned issues, H3D is designed and collected with the explicit goal of stimulating research on full-surround 3D multi-object detection and tracking in crowded urban scenes. The H3D is gathered from HDD

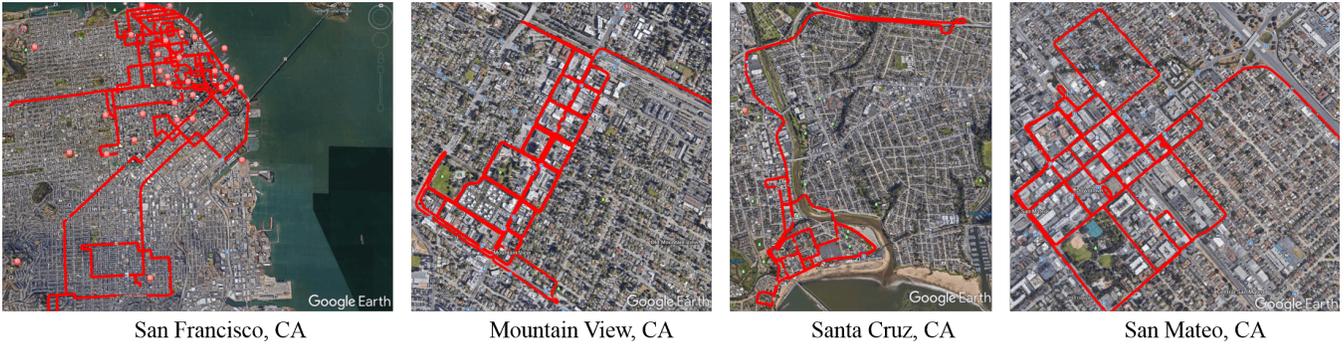


Fig. 2: Geographical distribution of H3D.

dataset¹ [15], a large scale naturalistic driving dataset collected in San Francisco Bay Area. Diverse, rich, and complex traffic scenes are selected in four major urban scenes as shown in Fig. 2 to develop and evaluate 3D multi-object detection and tracking algorithms. To annotate a large-scale dataset, we establish an effective and efficient labeling process to speed up the overall annotation cycle. The details will be discussed in Sec. III-C.2.

The contributions are summarized as follows. First, H3D is the first dataset for full-surround 3D multi-object detection and tracking in crowded urban scenes comprising of 1,071,302 3D bounding box labels of 8 common traffic participants. Second, a labeling methodology is introduced to annotate large-scale 3D bounding boxes. Third, a standardized benchmark of full-surround 3D multi-object detection and tracking is established for future algorithm developments. The dataset is available at <http://usa.honda-ri.com/H3D>.

II. TRAFFIC SCENE DATASETS

An increasing number of 2D scene understanding datasets [9], [10], [11] are proposed in recent years. In particular, these datasets aim to stimulate research on semantic segmentation for traffic scenes by providing high quality labels and scalable dataset generation methodologies. The Cityscapes dataset provides 5000 images with high quality pixel-level annotations and additional 20,000 images with coarse annotations for methods that leverage large volumes of weakly-labeled data [9]. The Mapillary dataset [10] increase the size of pixel-level annotations to 20,000 images and the diversity of images by selecting images from all around the world. While the two datasets provide high quality and high volume 2D annotations, they lack information from 3D to enable research on 3D object detection and tracking.

With a comparison to 2D scene understanding, relatively small efforts [16], [13], [17] have been made in 3D due to the costs for installing a high quality 3D range scanner and difficulties in labeling annotations in point cloud on a large-scale. Semantic3D.Net [13] and Oakland dataset [16] are two point cloud datasets that provide semantic labels for point cloud classification. The Ford Campus LiDAR Dataset [17]

consists of point cloud data collected in urban environments from multiple LiDAR devices. However, 3D bounding boxes and tracks of objects are not available to enable research on 3D detection and tracking of traffic participants.

It is non-trivial to manually label large-scale datasets. Huang et al. [11] annotate large-scale semantic segmentation by projecting labeled semantic labels on survey-grade dense 3D points. In the proposed labeling methodology, we leverage a similar idea by applying LiDAR SLAM to register multiple LiDAR scans to form a dense point cloud. In this case, static objects will only have to be labeled once instead of a frame-by-frame annotation. This methodology significantly improves the overall labeling cycle. More details will be discussed in Sec. III-C.2.

III. H3D DATASET

An outline of the steps involved in dataset generation is shown in Fig. 3:

- Calibration between GPS/IMU (ADMA sensor) and LiDAR (Velodyne HDL-64E) is obtained using hand-eye calibration method [18] which is a well-known approach to find the relationship between two given trajectories from different coordinate system. Data from all five sensors (3 cameras, LiDAR and GPS/IMU) is time-synchronized with GPS time-stamps.
- Undistortion in point cloud data is performed to remove motion artifacts for superior annotation.
- Point cloud registration is done to get ego-vehicle odometry estimates and point cloud data in each scenario is transformed to a fixed set of *World* coordinates.
- Annotation of objects in point clouds is done in the *World* coordinates by a group of annotators.
- The labeled data (bounding boxes and point cloud) is then converted back to *Velodyne* coordinates and changed to raw point cloud data.

A. Sensor Setup

The vehicle is equipped with the following sensors (as shown in Fig. 4):

- three color PointGrey Grasshopper3 video cameras (30HZ frame rate, 1920×1200 resolution and 90° field-of-view (FOV) for left and right, 80° FOV for center)

¹<https://usa.honda-ri.com/HDD>

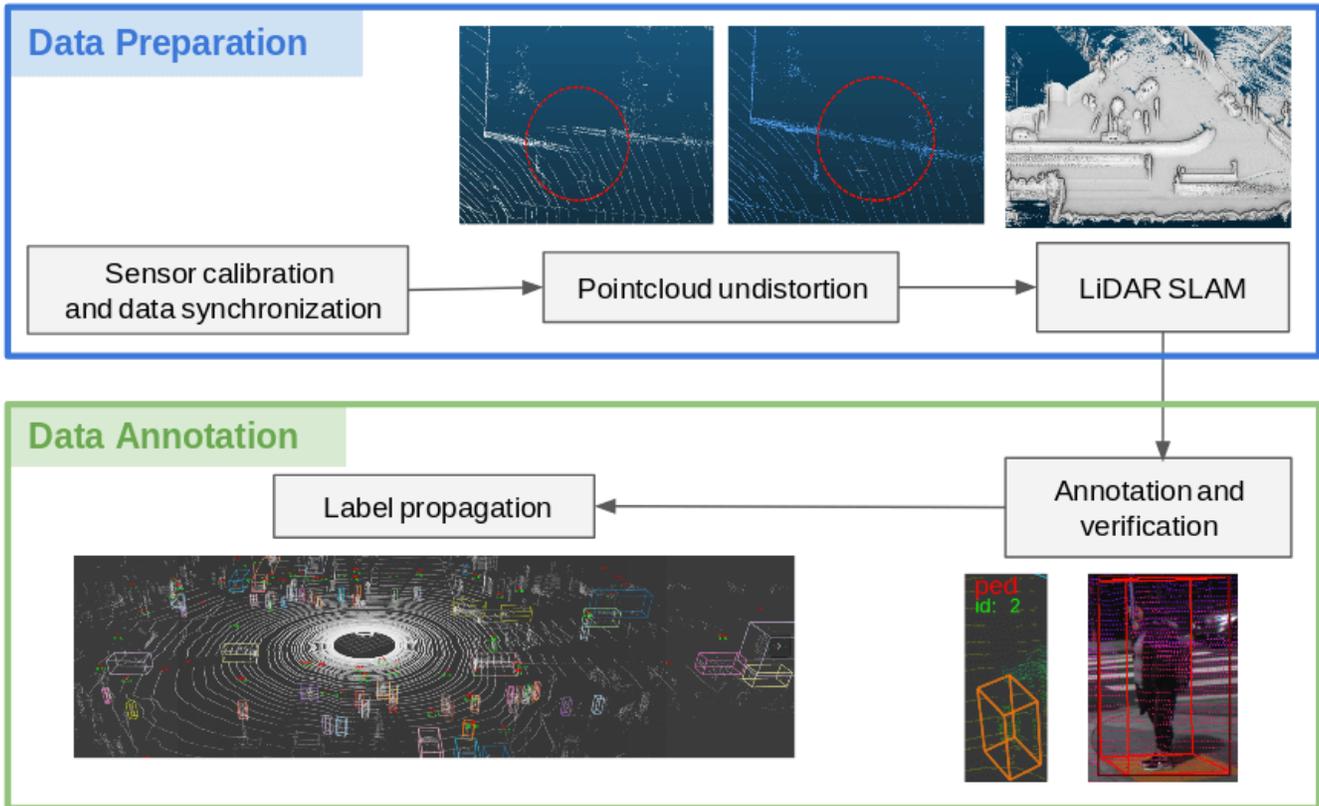


Fig. 3: Data labelling procedure

- a Velodyne HDL-64E S2 3D LiDAR (10 HZ spin-rate, 64 laser beams, range: 100m, vertical FOV 26.9°)
- a GeneSys Eletronik GmbH Automotive Dynamic Motion Analyzer (ADMA) with DGPS output gyros, accelerometers and GPS (frequency: 100 HZ)

Sensor data is recorded using a Ubuntu 14.04 machine with two eight-core Intel i5-6600K 3.5 GHz Quad-Core processors, 16 GB DDR3 memory, and a RAID 0 array of four 2TB SSDs.

B. Data Collection

Data is collected in 4 urban areas in the San Francisco Bay Area from April to September 2017 using an instrumented vehicle shown in Fig. 4(a). The routes for data collection are overlaid on images from Google Earth as highlighted in Fig. 2. Sensor data is synchronized using Robot Operating System (ROS)² via a customized hardware setup.

C. Data Labelling Procedure

In this section, we describe details of the 3D objects and tracklets labeling procedure for H3D.

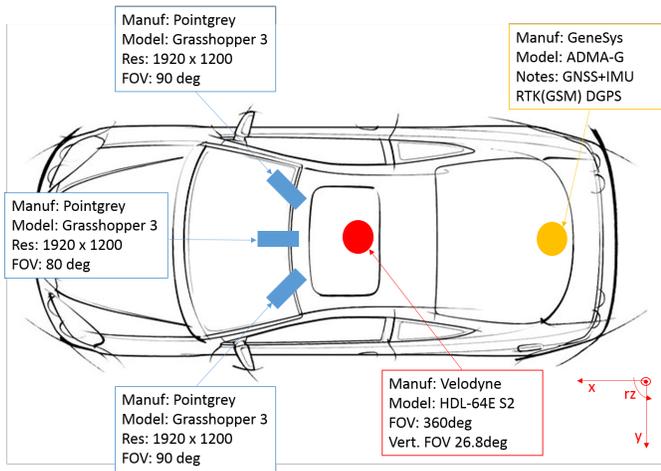
1) *Data Preparation:* Cameras and LiDAR are hardware-timestamped using the GPS time-stamps and other sensor data is synchronized via ROS. To prepare point cloud for annotation, an undistortion process is necessary because a raw point cloud is distorted due to a spinning LiDAR.

The process of undistortion is described as follows. The motion distortion is corrected using high-frequency fused GPS data obtained from the GPS/IMU sensor using linear interpolation method mentioned in [19].

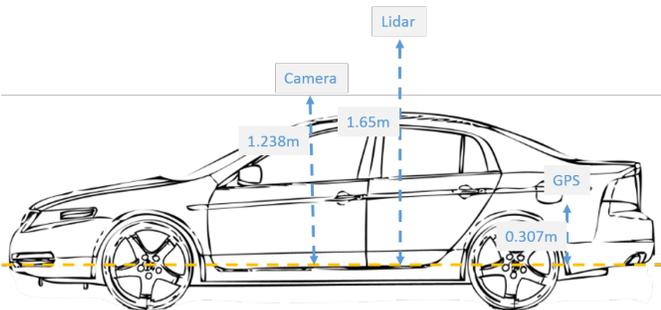
Normal Distributive Transform (NDT) [20] method is used for point cloud registration as shown in Fig. 3. With each sequence being independent, point cloud is registered with respect to the initial frame (*World*) of that particular sequence. Such a registration process is needed for odometry estimation as GPS data is unreliable in urban areas with enclosed spaces and hence the transformation of point cloud to *World* frame cannot be achieved accurately. Transforming point cloud data to *World* coordinates simplifies the data annotation process as data association between static objects can be easily achieved given the correspondence between point cloud data in various frames.

2) *Data Annotation:* The registered point cloud data allows annotators to determine corresponding objects easily. Additionally, the three cameras are utilized to assist the annotation process in order to determine object categories. We registered a sequence of point clouds at 2Hz from the odometry computed using NDT. Bounding boxes and track IDs are annotated on the registered point cloud. Doing so, the static objects in the registered frames can be annotated in one shot and this significantly reduces the labeling efforts. Moreover, the registered point cloud provides easier association of objects across frames. The human-labeled annotations

²<http://www.ros.org/>



(a) sensor layout of the vehicle



(b) side view of the vehicle

Fig. 4: Vehicle sensor setup

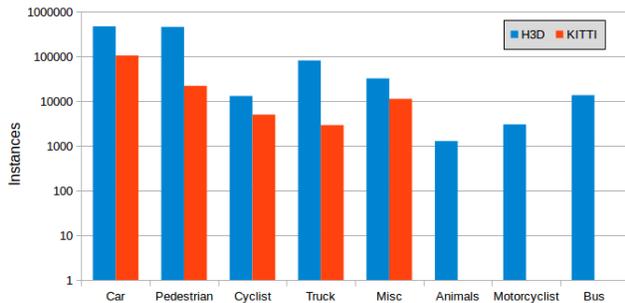


Fig. 5: Distribution of classes in H3D and KITTI

are then propagated to 10Hz using a linear interpolation technique, assuming a constant velocity model between each frame. The labeled data (bounding boxes with track IDs) is transformed back to the *Velodyne* coordinates using odometry estimates.

The quality of final labeled data is verified frame-by-frame by projecting the labeled bounding boxes onto corresponding images via methods similar to [21], [22] and by visually inspecting the labeled data in BEV as shown in Fig. 3.

D. Statistics

- **Complexity:** A comparison of density of common traffic participants averaged across 21 labeled scenarios

	train	validation	test
Scenarios	50	30	80
Frames	8873	5170	13678
Car	157174	84646	228738
Pedestrian	147985	65549	242697

TABLE I: Instances for the train, validation, and test split

	BEV	3D
car	76.50	68.31
pedestrian	50.88	50.39

TABLE II: mAP scores with 0.5 IoU for Car and 0.25 IoU for Pedestrian

in KITTI and 160 labeled scenarios in H3D is done to show the complexity of H3D dataset. For a fair comparison, number of annotations in H3D’s 360° scene is assumed to be 4 times that of number of annotations in KITTI which are in frontal view of the scene. We observed that density of traffic participants in H3D is 15 times higher than that in KITTI.

- **Volume:** The total number of bounding box annotations and the various classes annotated are shown in Fig. 5. Also, it can be seen from Table I that the proportion of cars and pedestrians is consistent among training/validation/test datasets.

IV. 3D DETECTION

H3D is currently the only dataset that enables full 360-degree object detection in point cloud. This paper evaluates VoxelNet [23] on H3D to obtain baseline values and assess the complexity of the dataset.

A similar training procedure is adapted to that from the original literature (VoxelNet) with following modifications. Points within 40 meters radius of ego-vehicle are considered for car detection and points within 25.6 meters radius are considered for pedestrian detection. The models for both car and pedestrian detection are trained using an ADAM optimizer. A learning rate of 0.01 is used for the first 40 epochs, then decreased to 0.001 for the next 20 epochs and further decreased to 0.0001 for the last 20 epochs (total 80 epochs). A batch size of 12 is used during training.

For evaluation, a similar protocol as KITTI is adapted. The IoU threshold for class car is set to 0.5 and that for class pedestrian is set to 0.25. Car and truck classes are combined when evaluating car detection performance. The results are summarized in (Table II) and shown in Fig. 7. The following challenges are encountered in 3D detection as highlighted in Fig. 8. The yaw estimation is not good where number of points for the particular object is less. Pedestrian detection fails to perform due to occlusion in crowded scenes.

V. 3D MULTI-OBJECT TRACKING

3D objects are tracked using an Unscented Kalman Filter (UKF) via following four steps - prediction, data-association,

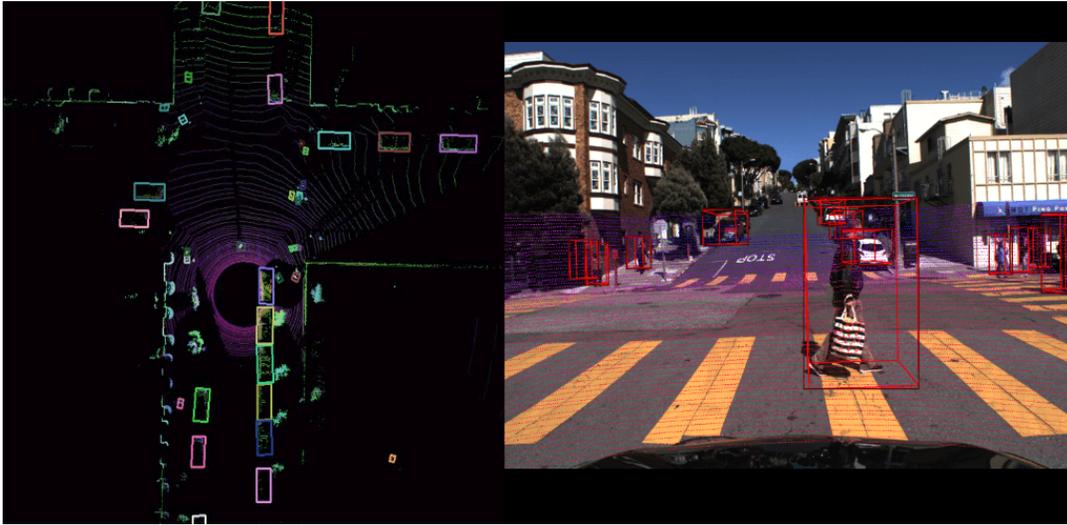
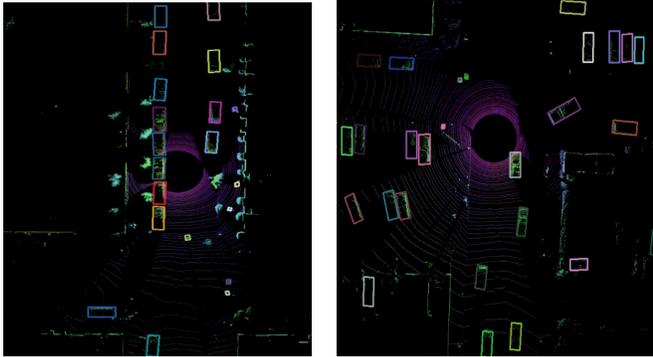


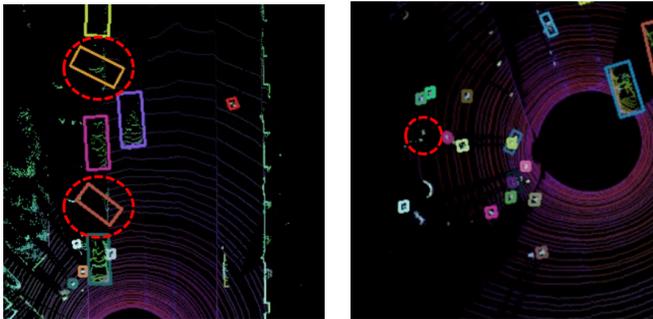
Fig. 6: Data annotation verification by projecting annotation onto image and bird's eye view (BEV), with color coded track ID in BEV



(a) all pedestrians and cars are detected with correct orientation

(b) open area detections output

Fig. 7: Successful detection cases in BEV



(a) wrong orientation for static vehicles in red dotted circle

(b) missing pedestrians in red dotted circle (because of occlusions)

Fig. 8: Failure detection cases in BEV

update and track-management. Data association of objects is done via euclidean distance between centroids of objects.

Parameters used for tracking are summarized as follows.

	MOTA	MOTP	MT	ML
car	76.2	73.1	56.7	25.1
pedestrian	36.8	63.0	14.1	43.4

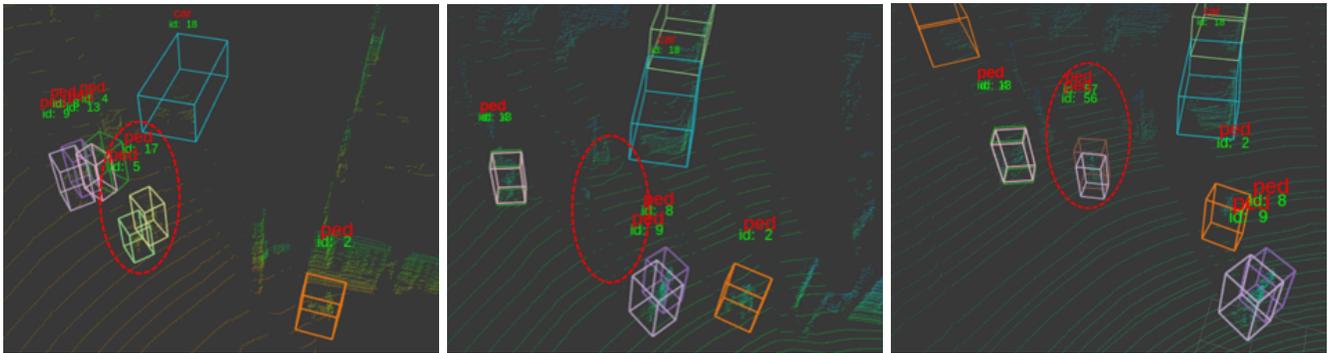
TABLE III: MOT scores with 0.5 3D IoU for Car and Pedestrian

The state vector comprises of 5 variables, namely, 'x' and 'y' position of objects (in m), their velocities (in m/s), their orientation (in rad) and their angular velocities (in rad/s). The euclidean distance threshold is set to 2 meters for data association for both car and pedestrian classes. An occlusion factor of 2, where occlusion factor is multiplied by the vertical area of object to determine if it becomes highly occluded. Lastly, an aging factor of 2 is used such that an object is kept in the history of tracks for at most 2 frames.

The evaluation protocol from KITTI is adapted for tracking [1] with 0.5 3D IoU for both car and pedestrian classes. In the tracking algorithm evaluation, CLEAR MOT metrics are used [24] which include Multi-Object Tracking Precision (MOTP), Multi-Object Tracking Accuracy (MOTA) and Mostly Tracked (MT), Mostly Lost (ML) as mentioned in [25]. The results for tracking are summarized in Table III. The analysis of tracking results shows that output is highly affected by quality of detections. The tracking algorithm is also evaluated with ground-truth locations of objects. The results indicate a considerable increase in accuracy with 0.99 MOTA, 1.00 MOTP, 1.00 MT and 0.00 ML for cars; 0.83 MOTA, 1.00 MOTP, 0.77 MT and 0.11 ML for pedestrians. Tracking is also affected when occlusions are present as shown in Fig. 9 for Pedestrians (track ID=15,17) with red dotted circle.

VI. CONCLUSION

This paper demonstrates the uniqueness and importance of H3D for research on full-surround 3D multi-object



(a) frame number=1, starting frame in sequence (b) frame number=22, pedestrians disappeared due to occlusion (c) frame number=34, new track IDs assigned to pedestrians

Fig. 9: Failure cases in long-term tracking via UKF for pedestrians; the red dotted circles in three different frames over time; (a) original pedestrian tracks; (b) highlight missing tracks due to occlusion; (c) change in data association

detection and tracking in crowded urban scenes. Labelling methodology of H3D allows annotation of 3D objects and their track IDs on a large-scale efficiently. A standard benchmark for future 3D point cloud detection and tracking algorithms development is established in the paper. Given the significantly raised attention for 3D scene understanding, we hope that H3D can push the performance envelope.

Acknowledgement: We are grateful to our colleagues Behzad Dariush, Kalyani Polagani, Kenji Nakai, Athma Narayanan, and Wei Zhan for their valuable input.

REFERENCES

- [1] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? The KITTI Vision Benchmark Suite," in *CVPR*, 2012.
- [2] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *NIPS*, 2015.
- [3] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*, 2016.
- [4] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *CVPR*, 2017.
- [5] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *CVPR*, 2017.
- [6] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *ICCV*, 2017.
- [7] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *CVPR*, 2009.
- [8] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollr, and C. L. Zitnick, "Microsoft COCO: Common objects in context," in *ECCV*, 2009.
- [9] 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 *CVPR*, 2016.
- [10] G. Neuhold, T. Ollmann, S. R. Bulo, and P. Kotschieder, "The Mapillary Vistas dataset for semantic understanding of street scenes," in *ICCV*, 2017.
- [11] X. Huang, X. Cheng, Q. Geng, B. Cao, D. Zhou, P. Wang, Y. Lin, and R. Yang, "The ApolloScape dataset for autonomous driving," in *CVPR*, 2018.
- [12] W. Maddern, G. Pascoe, C. Linegar, and P. Newman, "1 Year, 1000km: The Oxford RobotCar Dataset," *The International Journal of Robotics Research*, vol. 36, no. 1, pp. 3–15, 2017.
- [13] T. Hackel, N. Savinov, L. Ladicky, J. D. Wegner, K. Schindler, and M. Pollefeys, "Semantic3d.net: A new large-scale point cloud classification benchmark," in *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 2017.
- [14] J. Jeong, Y. Cho, Y.-S. Shin, H. Roh, and A. Kim, "Complex urban lidar data set," in *ICRA*, 2018.
- [15] V. Ramanishka, Y.-T. Chen, T. Misu, and K. Saenko, "Toward driving scene understanding: A dataset for learning driver behavior and casual reasoning," in *CVPR*, 2018.
- [16] D. Munoz, A. Bagnell, N. Vandapel, and M. Hebert, "Contextual classification with functional max-margin Markov networks," in *CVPR*, 2009.
- [17] G. Pandey, J. McBride, and R. Eustice, "Ford campus vision and lidar data set," *The International Journal of Robotics Research*, vol. 30, no. 13, pp. 1543–1552, 2011.
- [18] F. Dornaika and R. Horaud, "Simultaneous robot-world and hand-eye calibration," *IEEE transactions on Robotics and Automation*, vol. 14, no. 4, pp. 617–622, 1998.
- [19] J. Zhang and S. Singh, "Loam: Lidar odometry and mapping in real-time," in *RSS*, 2014.
- [20] P. Biber and W. Straßer, "The normal distributions transform: A new approach to laser scan matching," in *IROS*, 2003.
- [21] Q. Zhang and R. Pless, "Extrinsic calibration of a camera and laser range finder (improves camera calibration)," in *IROS*, 2004.
- [22] F. Vasconcelos, J. P. Barreto, and U. Nunes, "A minimal solution for the extrinsic calibration of a camera and a laser-rangefinder," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 11, pp. 2097–2107, 2012.
- [23] Y. Zhou and O. Tuzel, "Voxelnet: End-to-end learning for point cloud based 3d object detection," 2018.
- [24] K. Bernardin and R. Stiefelhagen, "Evaluating multiple object tracking performance: the CLEAR MOT metrics," *Journal on Image and Video Processing*, vol. 2008, p. 1, 2008.
- [25] Y. Li, C. Huang, and R. Nevatia, "Learning to associate: Hybrid-boosted multi-target tracker for crowded scene," in *CVPR*, 2009.