2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) | 978-1-7281-6212-6/20/$31.00 ©2020 IEEE | DOI: 10.1109/IROS45743.2020.9341020

the performance of surrounding vehicle pose estimation.

The remainder of this paper is organized as follows. Section II summarizes existing works for autonomous driving perception. Section III analyzes the subsystems of a typical perception system to help us better understand their purposes and principles. Details of our proposed method is introduced in section IV. Section V shows the experiments and results, while section VI concludes the paper.

II. Related Work

It is crucial to perceive the environment and extract useful information. This mainly includes vehicle detection, tracking, localization and mapping.

A. Vehicle Detection

Vehicle detection refers to estimating the position, heading and size of surrounding vehicles. There are three main subclasses: (1) Image-based vehicle detection generates bounding boxes on front-view camera image [26], [27]; (2) Semantic segmentation assigns each pixel of the image with a class label. Pixels belonging to same objects have the same label [14], [20]; and (3) 3D object detection obtains the 3D poses (or bird-eye poses) of surrounding vehicles, which is usually achieved with the help of lidar point cloud [10], [30].

B. Vehicle Tracking

Direct prediction from the vehicle detection system is often insufficient, more accurate vehicle state needs to be estimated given historical detection. Typical vehicle tracking methods have two phases: (1) data association to connect objects between frames [15], [23]; and (2) filtering methods such as Kalman Filters and Particle Filters to smooth the vehicle dynamics [12], [25].

C. Localization and Mapping

Localization is the task of estimating ego vehicle pose relative to a reference frame in a map, which can be either a raw point cloud map or an annotated semantic map, depending on the algorithm we choose. There are two main approaches: (1) Simultaneous localization and mapping (SLAM) makes map online and localize the ego vehicle in the map at the same time [4]; and (2) A priori map-based localization that estimate the ego vehicle pose by finding the best match to a detailed a priori map [19].

III. Preliminary

In this section, we will analyze the typical subsystems in an autonomous driving perception system, including detection, tracking, localization and mapping. They are reformulated into graphical models. This helps us better understand their purposes and relationships to fuse them into a single end-to-end framework.

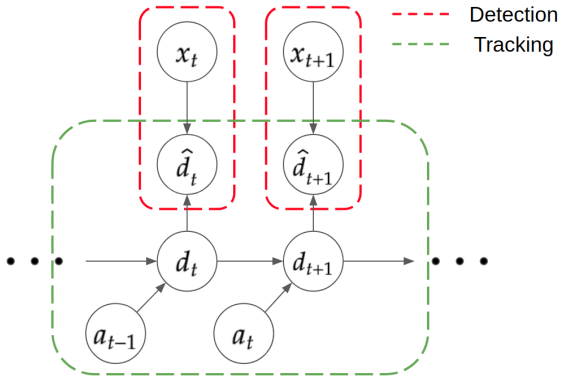

Fig. 2: Graphical model of typical detection and tracking systems. The red block represents the detection system. The green block represents the tracking system.

A. Typical Detection and Tracking Systems

The goal of the detection and tracking subsystems is to accurately estimate the states of surrounding vehicles. This includes their positions and heading angles relative to the ego vehicle, as well as their length and width. Current detection and tracking systems are divided into two subsystems. First, the detection subsystem predicts the vehicles' poses based on single frame sensor inputs. Then, the tracking subsystem smooths the results from the detection system based on its historical outputs.

These processes can be interpreted as a graphical model, as shown in Fig 2. The red block represents the detection subsystem, which contains a detection model $p\left(\hat{d}_t | x_t\right)$ that maps the sensor inputs x_t to an estimation of the surrounding vehicles' poses $\hat{d}_t$. This part is usually performed by fitting a deep neural network model using supervised learning techniques. The green block represents the tracking subsystem, which estimates the true vehicles' poses d_t given historical outputs of the detection system $\hat{d}_{1:t}$ and sometimes the historical ego vehicle actions $a_{1:t}$. This estimation problem can be formulated as a filter problem $p\left(d_t | \hat{d}_{1:t}; a_{1:t}\right)$ which is usually solved by Kalman filters with hard-coded transition model $p\left(d_{t+1} | d_t; a_t\right)$ and observation model $p\left(\hat{d}_t | d_t\right)$.

B. Typical Localization and Mapping Systems

The localization and mapping system needs to accurately estimate the pose of the ego vehicle in the coordinate of a global map. Then a local semantic map indicating road geometry, topology and traffics will be obtained, which are used for downstream planning and control tasks. The estimated global ego vehicle pose is also used for routing.

These processes can be interpreted as a graphical model, as shown in Fig 3. The global ego vehicle pose l_t is estimated based on historical sensor inputs and actions $p\left(l_t | x_{1:t}; a_{1:t}\right)$. The global map m can be edited offline or constructed online, depending on the methodology we use. After estimating the ego vehicle pose, a local semantic map S_t^{loc} is obtained by locating the ego vehicle on the global map, which requires

Fig. 9: Example results. The first row shows sensor inputs and ground truth, second row shows reconstructed outputs. Left to right: camera image, lidar image, local semantic roadmap, surrounding vehicle bounding boxes. Note that only camera and lidar inputs are given for the reconstruction, the ground truth semantic roadmap and bounding boxes are displayed here only for comparison.

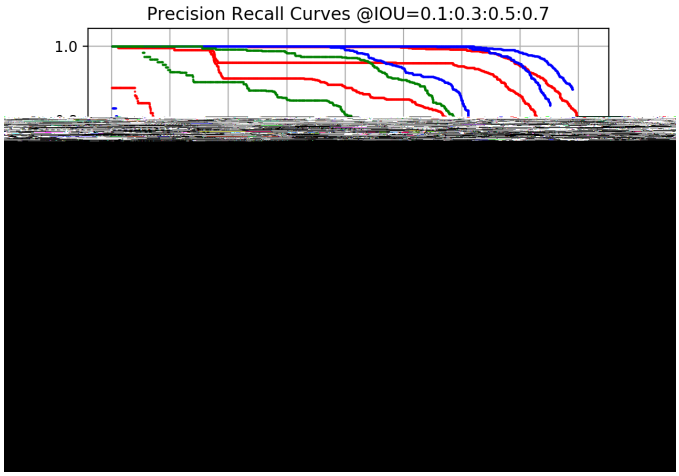

Fig. 10: Precision-Recall Curves for surrounding vehicles bounding boxes prediction.

TABLE I: Average precision for surrounding vehicles bounding boxes prediction.

	$AP_{0:1}$	$AP_{0:3}$	$AP_{0:5}$	$AP_{0:7}$
Inputs and roadmap	**79.4%**	72.0%	56.5%	16.8%
No inputs reconstruction	78.4	**74.4%**	**61.0%**	**22.1%**
No roadmap reconstruction	57.4%	45.3%	7.8%	0.3 %

This shows that fusing the information of map might improve the performance of detection.

For ego vehicle global state, we calculate its average prediction error. There are two values we care about, the location error (in meter) and heading error (in rad). Table.II shows the evaluation results. We can see that we get an average global location error of 7.1 meters, and heading error of 0.17 rad. This is obtained purely from the raw camera and lidar sensor inputs, with no GPS or stored global map used.

TABLE II: Average error of ego vehicle global pose estimation.

	Location (m)	Heading (rad)
Inputs and roadmap	11.4	0.33
No inputs reconstruction	8.6	**0.17**
No roadmap reconstruction	**7.1**	0.42

VI. Conclusion and Future Works

In this paper, we proposed and implemented an end-to-end autonomous driving perception system based on sequential latent representation learning. The system was able to replace the functionalities of the typical detection, tracking, localization and mapping subsytems with minimum human engineering efforts and without online stored maps. The method was evaluated in a realistic autonomous driving simulator, taking camera RGB image and lidar point cloud as sensor inputs. Evaluation results showed the learned model can obtain accurate surrounding vehicles' poses, local semantic roadmaps, and global ego vehicle pose based purely on historical raw sensor inputs.

Although the learned model performs reasonably well, it has a large space for improvement. The neural network architectures used in this paper are only the very basic ones, such as shallow convolutional layers. The size of input images is only 128 128, making it very hard to detect objects that are small or far away. In the future, we will deploy more advanced network architectures, and train the model on images with higher definition. We will also test this system with a downstream autonomous driving decision making system.

REFERENCES

[1] C. Badue, R. Guidolini, R. V. Carneiro, P. Azevedo, V. B. Cardoso, A. Forechi, L. Jesus, R. Berriel, T. Paixão, F. Mutz, et al. Self-driving cars: A survey. *arXiv preprint arXiv:1901.04407*, 2019.

[2] M. Bansal, A. Krizhevsky, and A. Ogale. Chauffeurnet: Learning to drive by imitating the best and synthesizing the worst. *arXiv preprint arXiv:1812.03079*, 2018.

[3] R. Bauer, D. Delling, P. Sanders, D. Schieferdecker, D. Schultes, and D. Wagner. Combining hierarchical and goal-directed speed-up techniques for dijkstra's algorithm. *Journal of Experimental Algorithmics (JEA)*, 15:2–1, 2010.

[4] G. Bresson, Z. Alsayed, L. Yu, and S. Glaser. Simultaneous localization and mapping: A survey of current trends in autonomous driving. *IEEE Transactions on Intelligent Vehicles*, 2(3):194–220, 2017.

[5] J. Chen, S. E. Li, and M. Tomizuka. Interpretable end-to-end urban autonomous driving with latent deep reinforcement learning. *arXiv preprint arXiv:2001.08726*, 2020.

[6] J. Chen, C. Tang, L. Xin, S. E. Li, and M. Tomizuka. Continuous decision making for on-road autonomous driving under uncertain and interactive environments. In *2018 IEEE Intelligent Vehicles Symposium (IV)*, pages 1651–1658. IEEE, 2018.

[7] J. Chen, B. Yuan, and M. Tomizuka. Deep imitation learning for autonomous driving in generic urban scenarios with enhanced safety. *arXiv preprint arXiv:1903.00640*, 2019.

[8] J. Chen, B. Yuan, and M. Tomizuka. Model-free deep reinforcement learning for urban autonomous driving. In *2019 IEEE Intelligent Transportation Systems Conference (ITSC)*, pages 2765–2771. IEEE, 2019.

[9] J. Chen, W. Zhan, and M. Tomizuka. Autonomous driving motion planning with constrained iterative lqr. *IEEE Transactions on Intelligent Vehicles*, 4(2):244–254, 2019.

[10] X. Chen, H. Ma, J. Wan, B. Li, and T. Xia. Multi-view 3d object detection network for autonomous driving. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1907–1915, 2017.

[11] A. Dosovitskiy, G. Ros, F. Codevilla, A. Lopez, and V. Koltun. Carla: An open urban driving simulator. *arXiv preprint arXiv:1711.03938*, 2017.

[12] A. Ess, K. Schindler, B. Leibe, and L. Van Gool. Object detection and tracking for autonomous navigation in dynamic environments. *The International Journal of Robotics Research*, 29(14):1707–1725, 2010.

[13] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2):303–338, 2010.

[14] K. He, G. Gkioxari, P. Dollár, and R. Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017.

[15] S. Hwang, N. Kim, Y. Choi, S. Lee, and I. S. Kweon. Fast multiple objects detection and tracking fusing color camera and 3d lidar for intelligent vehicles. In *2016 13th International Conference on Ubiquitous Robots and Ambient Intelligence (URAI)*, pages 234–239. IEEE, 2016.

[16] D. P. Kingma and M. Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[17] R. G. Krishnan, U. Shalit, and D. Sontag. Deep kalman filters.(2015). *arXiv preprint arXiv:1511.05121*, 2015.

[18] A. X. Lee, A. Nagabandi, P. Abbeel, and S. Levine. Stochastic latent actor-critic: Deep reinforcement learning with a latent variable model. *arXiv preprint arXiv:1907.00953*, 2019.

[19] J. Levinson, M. Montemerlo, and S. Thrun. Map-based precision vehicle localization in urban environments. In *Robotics: science and systems*, volume 4, page 1. Citeseer, 2007.

[20] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015.

[21] W. Luo, J. Xing, A. Milan, X. Zhang, W. Liu, X. Zhao, and T.-K. Kim. Multiple object tracking: A literature review. *arXiv preprint arXiv:1409.7618*, 2014.

[22] M. Montemerlo, J. Becker, S. Bhat, H. Dahlkamp, D. Dolgov, S. Ettinger, D. Haehnel, T. Hilden, G. Hoffmann, B. Huhnke, et al. Junior: The stanford entry in the urban challenge. *Journal of field Robotics*, 25(9):569–597, 2008.

[23] T.-N. Nguyen, B. Michaelis, A. Al-Hamadi, M. Tornow, and M.-M. Meinecke. Stereo-camera-based urban environment perception using occupancy grid and object tracking. *IEEE Transactions on Intelligent Transportation Systems*, 13(1):154–165, 2011.

[24] B. Paden, M. Čáp, S. Z. Yong, D. Yershov, and E. Frazzoli. A survey of motion planning and control techniques for self-driving urban vehicles. *IEEE Transactions on intelligent vehicles*, 1(1):33–55, 2016.

[25] A. Petrovskaya and S. Thrun. Model based vehicle detection and tracking for autonomous urban driving. *Autonomous Robots*, 26(2-3):123–139, 2009.

[26] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016.

[27] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015.

[28] C. Tang, J. Chen, and M. Tomizuka. Adaptive probabilistic vehicle trajectory prediction through physically feasible bayesian recurrent neural network. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 3846–3852. IEEE, 2019.

[29] C. Urmson, J. Anhalt, D. Bagnell, C. Baker, R. Bittner, M. Clark, J. Dolan, D. Duggins, T. Galatali, C. Geyer, et al. Autonomous driving in urban environments: Boss and the urban challenge. *Journal of Field Robotics*, 25(8):425–466, 2008.

[30] Y. Yan, Y. Mao, and B. Li. Second: Sparsely embedded convolutional detection. *Sensors*, 18(10):3337, 2018.

[31] B. Yang, W. Luo, and R. Urtasun. Pixor: Real-time 3d object detection from point clouds. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 7652–7660, 2018.

[32] E. Yurtsever, J. Lambert, A. Carballo, and K. Takeda. A survey of autonomous driving: common practices and emerging technologies. *arXiv preprint arXiv:1906.05113*, 2019.

[33] J. Ziegler, P. Bender, M. Schreiber, H. Lategahn, T. Strauss, C. Stiller, T. Dang, U. Franke, N. Appenrodt, C. G. Keller, et al. Making bertha drive—an autonomous journey on a historic route. *IEEE Intelligent transportation systems magazine*, 6(2):8–20, 2014.