3D Car Tracking using Fused Data in Traffic Scenes for Autonomous Vehicle

Can Chen, Luca Zanotti Fragonara and Antonios Tsourdos
SATM, Cranfield University, College Road, Cranfield, Bedford, U.K.

Keywords: 3D Bounding Box, Car Detection, Multiple Object Tracking, Autonomous Vehicle.

Abstract: Car tracking in a traffic environment is a crucial task for the autonomous vehicle. Through tracking, a self-driving car is capable of predicting each car's motion and trajectory in the traffic scene, which is one of the key components for traffic scene understanding. Currently, 2D vision-based object tracking is still the most popular method, however, multiple sensory data (e.g. cameras, Lidar, Radar) can provide more information (geometric and color features) about surroundings and show significant advantages for tracking. We present a 3D car tracking method that combines more data from different sensors (cameras, Lidar, GPS/IMU) to track static and dynamic cars in a 3D bounding box. Fed by the images and 3D point cloud, a 3D car detector and the spatial transform module are firstly applied to estimate current location, dimensions, and orientation of each surrounding car in each frame in the 3D world coordinate system, followed by a 3D Kalman filter to predict the location, dimensions, orientation and velocity for each corresponding car in the next time. The predictions from Kalman filtering are used for re-identifying previously detected cars in the next frame using the Hungarian algorithm. We conduct experiments on the KITTI benchmark to evaluate tracking performance and the effectiveness of our method.

1 INTRODUCTION

Multiple object tracking (MOT) is one of the key roles in computer vision problems. The aim of the MOT is to predict trajectories and maintain identities individually in the next frame, given initial positions and identities of multiple objects from sequential frames. The majority of current tracking methods are detection-based tracking on RGB images, and wide applications involve many scenarios, for example, pedestrian surveillance in public (Chen et al., 2018), (Breitenstein et al., 2011), sport players tracking in the video (Lu et al., 2013), (Xing et al., 2011), (Nillius et al., 2006). However, multiple object tracking in autonomous driving scenario needs to involve more information, such as accurate depth information, for safe navigation and environment perception, as pure visual tracking presents insufficient information for estimation of surrounding objects' 3D locations and motions, which is important to traffic scene understanding for the autonomous vehicle (Kocić et al., 2018).

In order to sense surrounding objects in autonomous system, many kinds of sensors are utilized to obtain information of whereabouts, such as lidar, kinds of cameras, etc. Visual sensors (e.g. monocular camera, stereo camera, RGB-D camera) can indicate a large number of color information from environment. Some application examples include object detection in image domain (Girshick, 2015) (Redmon et al., 2016), traffic sign recognition (Luo et al., 2018) and other applications. While lidar is able to provide precise 3D laser scan to generate point clouds for 3D representation of the environment. Although the disadvantage of highly cost, it can provide precise distance information and is normally used to create high-resolution maps with applications of SLAM for navigation (Durrant-Whyte and Bailey, 2006) and object detection for 3D detection in point clouds domain (Zhou and Tuzel, 2017) in autonomous system. Considering that camera takes advantage of obtain color information, while lidar sensor has high accuracy of distance information without color data, the data fusion problem combining lidar with camera has become increasingly popular recently in applications of 3D object (cars, pedestrians, cyclists) detection with orientation estimation, 3D object tracking for traffic scene understanding. Meanwhile information from GPS/IMU will support ego-car to estimate self pose and attitude for tracking objects in the world space.

Chen, C., Fragonara, L. and Tsourdos, A.
3D Car Tracking using Fused Data in Traffic Scenes for Autonomous Vehicle.
DOI: 10.5220/0007674203120318
In *Proceedings of the 5th International Conference on Vehicle Technology and Intelligent Transport Systems (VEHITS 2019)*, pages 312-318
ISBN: 978-989-758-374-2
Copyright © 2019 by SCITEPRESS – Science and Technology Publications, Lda. All rights reserved

Driven by the applications in autonomous driving scenario and advantages of kinds of sensors for perception, many solutions are proposed for 3D object detection based on multiple sensory data in robotics and autonomous scenario for environment perception. (Chen et al., 2017), (Ku et al., 2017) show state of art 3D detection performance based on camera images and lidar point clouds data. However few works are presented for 3D objects tracking in traffic scene due to the open issues of understanding of complex traffic environment, precise data fusion and 3D object detection from multiple sensors, motion estimation for each traffic participator.

2 RELATED WORK

According to different criteria, multiple object tracking problem can be categorized to different sets(Luo et al., 2014). For example, depending on target initialization, tracking methods are grouped in detection-based and detection-free tracking. Detection-based tracking applies trained object detector to each frame to obtain the location and appearance of each object in advance, then tracking model links locations to trajectories for the same object from the sequence (Berclaz et al., 2011), (Breitenstein et al., 2009), (Ess et al., 2009). Detection-free tracking needs to initialize the positions of the respective targets manually in the first frame, then the tracking model will track the given targets in the sequence according to targets appearance (Hu et al., 2012), (Zhang and van der Maaten, 2013). MOT can also be classified in online tracking and offline tracking. The difference is that online tracking does not leverage observations from future sequence (Hong et al., 2015), (Choi et al., 2017), (Hu et al., 2012), while offline tracking utilizes a batch of frames or the whole sequence as input before tracking (Yang et al., 2011), (Brendel et al., 2011).

For instance, in the 2D image MOT domain, (Xiang et al., 2015) considers the MOT problem as a Markov Decision Process (MDP) that leverages reinforcement learning approach to learn a policy to realize the data association. (Choi, 2015) introduced an aggregated local flow descriptor for similarity measurement, the descriptor encodes the relative motion pattern in the 2D bounding box from each frame for data association, which will be combined with object dynamics, appearance similarity and regularization of long-term trajectories to achieve robust tracking.

In the 3D image with depth information MOT domain, stereo cameras (Geiger et al., 2010) or time-of-flight cameras (RGB-D) (Zollhöfer et al., 2018) are involved to exploit sparse depth or per-pixel depth information. (Osep et al., 2017) utilizes the stereo camera to detect cars and pedestrians in the street and estimate visual odometry (VO) in advance, then VO is used for the ego-car's pose estimation that is important to transform objects to world space system whilst ego-car is also moving. At last, a 2D-3D Kalman filter combines objects in a 2D image bounding box and a 3D point cloud bounding box to obtain robust tracking performance.

In the lidar-sensor-based 3D MOT domain, the tracking process is normally in detection-free mode (Dequaire et al., 2016), (Dequaire et al., 2018), (Dewan et al., 2016), (Kaestner et al., 2012), as object detection based on pure point clouds data needs lots of computation to extract information from 100K point clouds in traffic scenario, which leads to very slow inference speed. As a result, (Dequaire et al., 2016), (Dequaire et al., 2018) introduced end-to-end neural networks to learn object links from raw lidar data, realizing a tracking pipeline without detection process. In addition, (Dequaire et al., 2016) demonstrated the end-to-end tracking approach on a moving platform in a busy traffic environment to track both static and dynamic objects through occlusion, while (Dequaire et al., 2018) proposed an end-to-end tracking framework in the totally unsupervised manner.

In the multi-sensors-based 3D tracking domain, (Asvadi et al., 2016) leveraged multiple sensors (2D image camera, 3D point cloud lidar) to estimate the location of cars, pedestrians, and cyclists, then two mean-shift (Cheng, 1995) methods are applied to 2D images and 3D point clouds respectively for location estimation of each object. At last, a constant acceleration Kalman filter is used for tracking both objects' location in the 2D image and 3D point clouds. However, the tracking method is applied as single-object tracking rather than MOT, as the initial position of the object need to be given in advance before tracking.

Some researcher leverages semantic segmentation processing for object detection and tracking (Ošep et al., 2016). They proposed a two-stage segmentation and scale-stable clustering to detect and track more generic objects in the complex street scenes instead of limited categories of objects.

3 METHOD OVERVIEW

In our work, depicted in Figure 1, we use images, point clouds and GPS/IMU data from KITTI benchmark (Geiger et al., 2012), and take advantage of a 3D object detection model AVOD (Ku et al., 2017) for accurate 3D car detection, followed by a spatial transformer module used to map 3D cars from the camera

coordinate system to the world coordinate system. At last, we introduce a 3D Kalman filter to estimate 3D car trajectories.

3.1 3D Car Detection based on Images and Point Clouds Data

The proposed 3D car tracking method is a detection-based and online tracking algorithm, which deeply relies on the performance of the 3D car detector. We compared previous state-of-the-art models for 3D car detection in Table 1, and use Aggregate View Object Detection (AVOD) (Ku et al., 2017) as our 3D car detector. Although VoxelNet achieved higher accuracy than AVOD, the forward time takes much longer than AVOD. For purpose of high-resolution feature maps, the AVOD encodes images and BEV (Bird Eye View) of point cloud to a modified VGG-16 network (Simonyan and Zisserman, 2014) to generate full-sized feature maps, which are used for estimating the 3D proposal generation and orientation vector. The detector model is generated by minimizing a multiple loss that includes a regression task for the 3D bounding box and orientation vector prediction, and a classification task for the object category.

In our tracking application, the output of the AVOD network is used as an observation model. The detection results for each frame will provide orientation ry^{3D}_{img}, length l^{3D}_{img}, width w^{3D}_{img}, height h^{3D}_{img}, centerpoint of the bottom $[x^{3D}_{img}, y^{3D}_{img}, z^{3D}_{img}]$ as 3D bounding box b^{3D}_{img} in the camera space as the current state for detected car, shown in Figure 2

$$b^{3D}_{img} = [ry^{3D}_{img}, l^{3D}_{img}, w^{3D}_{img}, h^{3D}_{img}, x^{3D}_{img}, y^{3D}_{img}, z^{3D}_{img}]^T \quad (1)$$

3.2 Spatial Transform Module

When an autonomous vehicle is in a static scenario, which means that the ego-car is in the stationary situation, the moving cars viewed in the world coordinate system will be coherent with those viewed in the local sensor coordinate system, such as camera coordinate system. However, when tracking dynamic cars from a moving platform, the egomotion of the sensor on the autonomous vehicle will affect the relative location and orientation of the detected cars from the sensors' perspective. As a result, we develop a spatial transformation module to decouple the egomotion of the autonomous vehicle from the motion of the detected cars in the traffic scene.

In order to transform the location and orientation of the detected cars to the world coordinate system while the ego-vehicle is moving, we utilize the GPS/IMU data to estimate the ego-vehicle's 6 degrees of freedom (DoF) pose T^t_{t-1} in the IMU coordinate system (Osborne, 2008), followed by the transformation such that:

$$P^{world}_{img} = T^t_{t-1} * T^{imu}_{img} * P_{img} \quad (2)$$

Where P^{world}_{img} is the transformation from the camera space to the world space, T^t_{t-1} is the ego-vehicle's pose update from time $t-1$ to t in the IMU coordinate system, T^{imu}_{img} is the transformation from camera space to IMU space, P_{img} is the 3D location of detected car in the camera space.

3.3 3D Constant Acceleration Kalman Filter

A 3D constant acceleration Kalman filter is used for robust 3D car tracking, as the constant acceleration assumed will be useful to model target motion that is smooth in position and velocity changes. The state of vector is defined as $x_{world} = [ry, \dot{ry}, \ddot{ry}, l, w, h, x, \dot{x}, \ddot{x}, y, \dot{y}, \ddot{y}, z, \dot{z}, \ddot{z}]^T_{world}$ in the world coordinate system, and $[\dot{x},\dot{y},\dot{z}]$, $[\ddot{x},\ddot{y},\ddot{z}]$ are the velocity and the acceleration corresponding to x,y,z positions respectively. The state model and measurement model of the discrete linear time-invariant system are defined by Equations 3 and 4:

$$x_t = A \cdot x_{t-1} + w_{t-1} \quad (3)$$

$$z_t = C \cdot x_t + v_t \quad (4)$$

where t is the time index, A is the state transition matrix of the process from the state x_{t-1} to the state x_t, C is the observation matrix that connects the state vector and the measurement vector, w and v are the process noise and measurement noise, and they are all assumed to be mutually independent, zero-mean, white Gaussian noise, with covariance Q_t and R_t respectively. z_t is the measurement. The problem is to estimate $\hat{x}_{t+1}$ of x_{t+1} for the prediction of cars location and orientation, given the detection of cars in the next frame z_{t+1}.

3.4 Data Association

Data association is a typical problem encountered in MOT task matching the same identities in sequence. In the last decades, many methods have been proposed for data association, such as the Hungarian algorithm (Kuhn, 1955), the Joint Probabilistic Data Association (JPDA) (Fortmann et al., 1983), the Multiple Hypothesis Tracking (MHT) (Reid et al., 1979). In this paper, we use the Hungarian algorithm in our

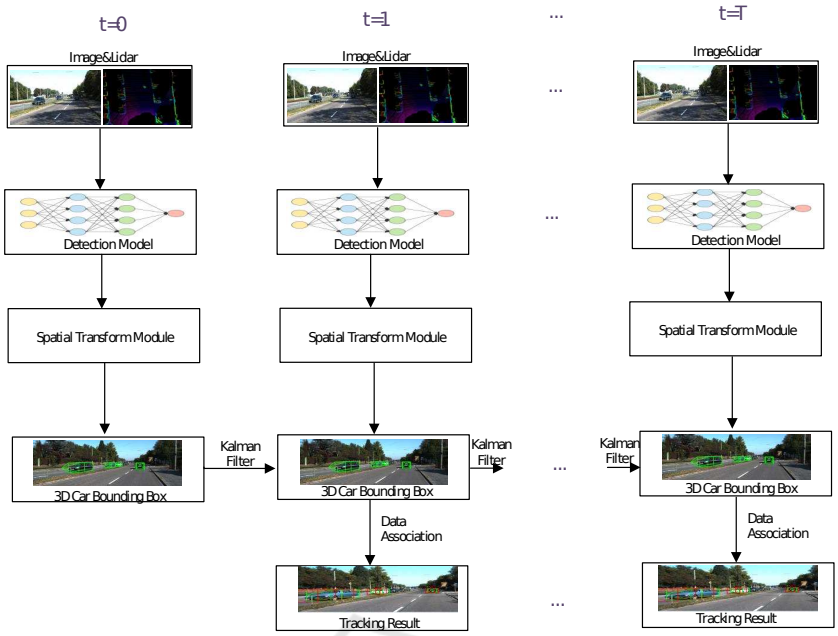

Figure 1: 3D Car Tracking Method Process.

Table 1: Previous State of art 3D Car Detection Networks.

Networks	Input Data	Forward Time	Easy	Moderate	Hard
			$AP(\%)$		
AVOD (Ku et al., 2017)	Image&Lidar	0.08s	73.59	65.78	58.38
MV3D (Chen et al., 2017)	Image&Lidar	0.36s	71.09	62.35	55.12
VoxelNet (Zhou and Tuzel, 2017)	Lidar	0.5s	77.47	65.11	57.73

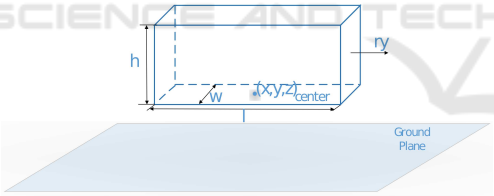

Figure 2: 3D Car Bounding Box.

work as it is a simple, efficient and linear data association algorithm that minimizes the total cost of similarity for each observation and hypotheses pair. The performance of data association is indicated from the MOT metrics of MOTA and IDS in Table 3.

The pre-fit residual between predicted quantity and measurement, and the corresponding residual covariance matrix are defined by Equation 5 and 6.

$$\tilde{y_k} = z_k - C_k \hat{x}_{k|k-1} \tag{5}$$

$$S_k = R_k + C_k P_{k|k-1} C_k^T \tag{6}$$

where P is the predicted state error covariance matrix. Considering the distance of state vector between measurement and prediction satisfies Gaussian distribution with mean of $\tilde{y_k}$ and covariance of S_k, the probability of which is defined as distance measure d_{ij} from each tracking target state vector to measurement. Then we apply the Hungarian algorithm (Kuhn, 1955), which can find an optimal assignment solution using a cost matrix, and solving the matching problem by minimizing the total cost E of Equation 7.

$$E = \sum_{i}^{N} \sum_{j}^{N} M_{ij} d_{ij} \tag{7}$$

where M is a permutation matrix that determines the total cost, N is the greater dimension of distance matrix. It is worth noting that the distance matrix d will be padded by 0 if it is not satisfied with square matrix.

4 EXPERIMENTAL RESULTS

In this section, we will apply our proposed tracking method to the training dataset from the KITTI tracking benchmark (Geiger et al., 2013). The dataset is captured from a VW station wagon equipped with 2 high-resolution color stereo-cameras, 2 high-resolution gray stereo-cameras, 1 Velodyne HDL-64E

rotating 3D laser scanner, 1 GPS/IMU navigation system. We extracted training sequences with ground truth of detections for cars in 2D RGB images and 3D point clouds from the dataset for evaluation. As the ground truth of the testing dataset is not available, we used a cross-validation method and split 7 sequences from a total of 21 training tracking sequences.

Evaluation policy is based on the CLEAR MOT metrics (Bernardin and Stiefelhagen, 2008). The performance metrics includes MOT precision (MOTP) indicating averaged total error of distance between the object and its corresponding matched object-hypothesis for all frames, MOT accuracy (MOTA) indicating total tracking accuracy for all frames. Besides, (Li et al., 2009) introduced other metrics, such as mostly tracked (MT), partly tracked (PL), mostly lost (ML), identity switches (IDS) indicating changing number of matched ground truth identity, fragmentations (FR) indicating the number of interrupted matched objects, MODA presents multiple object detection accuracy.

4.1 3D Car Detection Result

The 3D car detection AVOD model is trained based on KITTI dataset on an NVIDIA 1080Ti GPU, and the results are evaluated on the validation dataset using the 3D bounding box IOU, BEV AP (Average Precision) and global orientation angle. The best results are shown in Table 2, which are similar to the official results. The slight difference is normal due to the random dropout and random values of initial filters during training. Figure 3 presents several qualitative car detection results in 3D bounding box.

Table 2: Car Detection Average Precision (AP) of AVOD.

		Easy	Moderate	Hard
$AP_{3D}(\%)$	Official	73.59	65.78	58.38
	Ours	72.33	64.87	64.97
$AP_{BEV}(\%)$	Official	86.80	85.44	77.73
	Ours	86.18	77.98	78.05

4.2 3D Car Tracking Result

A clear advantage of Kalman filter can be seen in Table 3 when we track cars in the 3D bounding box in the world space, the visualization of the result is presented after transforming the location and rotation of the cars from world space to image coordinate system. It indicates that the tracking accuracy varies with the detection performance. In Figure 4, the green 3D bounding boxes with an arrow are detected cars, and the red bounding boxes are the predicted result from the Kalman filter from the previous frame.

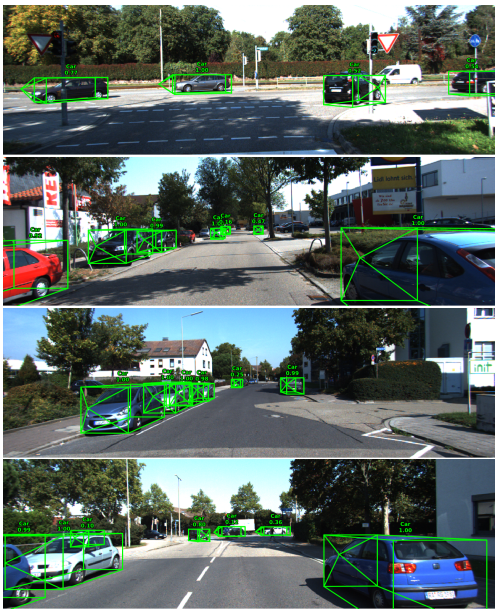

Figure 3: Car Detection Result.

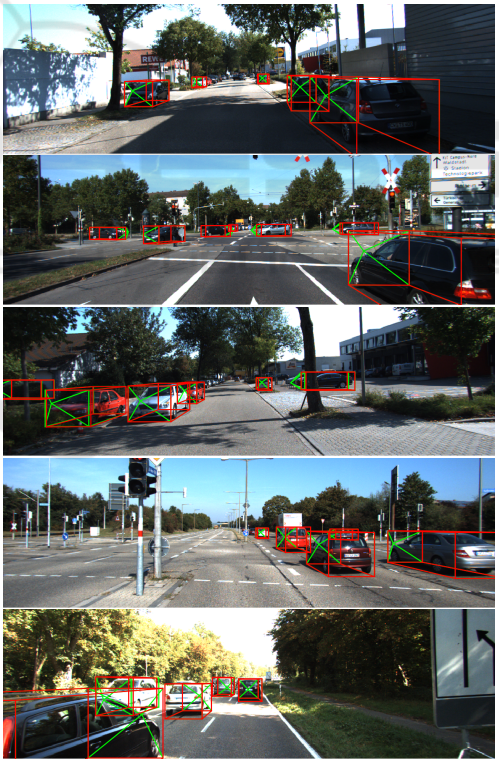

Figure 4: 3D Car Tracking Result.

5 CONCLUSIONS

In this paper, we presented a 3D dynamic car tracking method in the dynamic traffic scene on the moving

Table 3: 3D Car Tracking Result from Sequences from 0001 to 0007.

	Seq01	Seq02	Seq03	Seq04	Seq05	Seq06	Seq07	Overall
MODA (%)	59.46	38.90	73.65	57.16	69.60	85.20	77.73	65.77
MOTA (%)	58.83	39.90	73.35	56.51	69.02	85.20	77.38	65.38
MOTP (%)	83.75	75.31	79.22	75.95	83.81	74.14	84.24	81.34
IDS	7	0	1	5	7	0	7	27
Frag	31	17	14	38	32	8	42	182
MT (%)	42.31	26.67	50.00	19.23	36.36	72.73	77.36	48.48
PT (%)	30.77	53.33	50.00	69.23	57.58	27.27	22.64	40.40
ML (%)	26.92	20.00	0.0	11.54	6.06	0.0	0.0	11.11

ego-car. Firstly we utilized a state of art car detector AVOD to detect 3D cars from fused 2D RGB images and its corresponding point clouds data, then a spatial transform module maps the location and orientation of detected car from camera space to world coordinate system due to the effect of egomotion of the moving platform. At last, a constant acceleration Kalman filter was applied to estimate the state of dynamic cars in the world space for a smooth trajectory in the next frame. However, we noticed that the estimated tracking bounding boxes are disturbed significantly by the false rotation of the car that detected by the AVOD, and our future work will focus on improvement of rotation estimation of cars.

ACKNOWLEDGEMENTS

We thank all the reviewers for the comments and suggestions. This research is part of the HumanDrive project, funded by InnovateUK (Project ref: 103283) in the framework of the CAV2 call.

REFERENCES

Asvadi, A., Girao, P., Peixoto, P., and Nunes, U. (2016). 3d object tracking using rgb and lidar data. In *Intelligent Transportation Systems (ITSC), 2016 IEEE 19th International Conference on*, pages 1255–1260. IEEE.

Berclaz, J., Fleuret, F., Turetken, E., and Fua, P. (2011). Multiple object tracking using k-shortest paths optimization. *IEEE transactions on pattern analysis and machine intelligence*, 33(9):1806–1819.

Bernardin, K. and Stiefelhagen, R. (2008). Evaluating multiple object tracking performance: the clear mot metrics. *Journal on Image and Video Processing*, 2008:1.

Breitenstein, M. D., Reichlin, F., Leibe, B., Koller-Meier, E., and Van Gool, L. (2009). Robust tracking-by-detection using a detector confidence particle filter. In *Computer Vision, 2009 IEEE 12th International Conference on*, pages 1515–1522. IEEE.

Breitenstein, M. D., Reichlin, F., Leibe, B., Koller-Meier, E., and Van Gool, L. (2011). Online multiperson tracking-by-detection from a single, uncalibrated camera. *IEEE transactions on pattern analysis and machine intelligence*, 33(9):1820–1833.

Brendel, W., Amer, M., and Todorovic, S. (2011). Multiobject tracking as maximum weight independent set. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 1273–1280. IEEE.

Chen, L., Ai, H., Zhuang, Z., and Shang, C. (2018). Real-time multiple people tracking with deeply learned candidate selection and person re-identification. In *2018 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE.

Chen, X., Ma, H., Wan, J., Li, B., and Xia, T. (2017). Multi-view 3d object detection network for autonomous driving. In *IEEE CVPR*, volume 1, page 3.

Cheng, Y. (1995). Mean shift, mode seeking, and clustering. *IEEE transactions on pattern analysis and machine intelligence*, 17(8):790–799.

Choi, J., Kwon, J., and Lee, K. M. (2017). Visual tracking by reinforced decision making. *arXiv preprint arXiv:1702.06291*.

Choi, W. (2015). Near-online multi-target tracking with aggregated local flow descriptor. In *Proceedings of the IEEE international conference on computer vision*, pages 3029–3037.

Dequaire, J., Ondrúška, P., Rao, D., Wang, D., and Posner, I. (2018). Deep tracking in the wild: End-to-end tracking using recurrent neural networks. *The International Journal of Robotics Research*, 37(4-5):492–512.

Dequaire, J., Rao, D., Ondruska, P., Wang, D., and Posner, I. (2016). Deep tracking on the move: learning to track the world from a moving vehicle using recurrent neural networks. *arXiv preprint arXiv:1609.09365*.

Dewan, A., Caselitz, T., Tipaldi, G. D., and Burgard, W. (2016). Motion-based detection and tracking in 3d lidar scans. In *Robotics and Automation (ICRA), 2016 IEEE International Conference on*, pages 4508–4513. IEEE.

Durrant-Whyte, H. and Bailey, T. (2006). Simultaneous localization and mapping (slam): Part i the essential algorithms. robotics and automation magazine 13 (2): 99–110. doi: 10.1109/mra. 2006.1638022. Technical report, Retrieved 2008-04-08.

Ess, A., Leibe, B., Schindler, K., and Van Gool, L. (2009). Robust multiperson tracking from a mobile platform. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(10):1831–1846.

Fortmann, T., Bar-Shalom, Y., and Scheffe, M. (1983). Sonar tracking of multiple targets using joint probabilistic data association. *IEEE journal of Oceanic Engineering*, 8(3):173–184.

Geiger, A., Lenz, P., Stiller, C., and Urtasun, R. (2013). Vision meets robotics: The kitti dataset. *The International Journal of Robotics Research*, 32(11):1231–1237.

Geiger, A., Lenz, P., and Urtasun, R. (2012). Are we ready for autonomous driving? the kitti vision benchmark suite. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

Geiger, A., Roser, M., and Urtasun, R. (2010). Efficient large-scale stereo matching. In *Computer Vision–ACCV 2010*, pages 25–38. Springer.

Girshick, R. (2015). Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448.

Hong, S., You, T., Kwak, S., and Han, B. (2015). Online tracking by learning discriminative saliency map with convolutional neural network. In *International Conference on Machine Learning*, pages 597–606.

Hu, W., Li, X., Luo, W., Zhang, X., Maybank, S., and Zhang, Z. (2012). Single and multiple object tracking using log-euclidean riemannian subspace and block-division appearance model. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(12):2420–2440.

Kaestner, R., Maye, J., Pilat, Y., and Siegwart, R. (2012). Generative object detection and tracking in 3d range data. In *Robotics and Automation (ICRA), 2012 IEEE International Conference on*, pages 3075–3081. IEEE.

Kocić, J., Jovičić, N., and Drndarević, V. (2018). Sensors and sensor fusion in autonomous vehicles. In *2018 26th Telecommunications Forum (TELFOR)*, pages 420–425. IEEE.

Ku, J., Mozifian, M., Lee, J., Harakeh, A., and Waslander, S. (2017). Joint 3d proposal generation and object detection from view aggregation. *arXiv preprint arXiv:1712.02294*.

Kuhn, H. W. (1955). The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97.

Li, Y., Huang, C., and Nevatia, R. (2009). Learning to associate: Hybridboosted multi-target tracker for crowded scene.

Lu, W.-L., Ting, J.-A., Little, J. J., and Murphy, K. P. (2013). Learning to track and identify players from broadcast sports videos. *IEEE transactions on pattern analysis and machine intelligence*, 35(7):1704–1716.

Luo, H., Yang, Y., Tong, B., Wu, F., and Fan, B. (2018). Traffic sign recognition using a multi-task convolutional neural network. *IEEE Transactions on Intelligent Transportation Systems*, 19(4):1100–1111.

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

Nillius, P., Sullivan, J., and Carlsson, S. (2006). Multi-target tracking-linking identities using bayesian network inference. In *null*, pages 2187–2194. IEEE.

Osborne, P. (2008). The mercator projections.

Ošep, A., Hermans, A., Engelmann, F., Klostermann, D., Mathias, M., and Leibe, B. (2016). Multi-scale object candidates for generic object tracking in street scenes. In *Robotics and automation (icra), 2016 ieee international conference on*, pages 3180–3187. IEEE.

Osep, A., Mehner, W., Mathias, M., and Leibe, B. (2017). Combined image-and world-space tracking in traffic scenes. In *Robotics and Automation (ICRA), 2017 IEEE International Conference on*, pages 1988–1995. IEEE.

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

Reid, D. et al. (1979). An algorithm for tracking multiple targets. *IEEE transactions on Automatic Control*, 24(6):843–854.

Simonyan, K. and Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Xiang, Y., Alahi, A., and Savarese, S. (2015). Learning to track: Online multi-object tracking by decision making. In *Proceedings of the IEEE international conference on computer vision*, pages 4705–4713.

Xing, J., Ai, H., Liu, L., and Lao, S. (2011). Multiple player tracking in sports video: A dual-mode two-way bayesian inference approach with progressive observation modeling. *IEEE Transactions on Image Processing*, 20(6):1652–1667.

Yang, B., Huang, C., and Nevatia, R. (2011). Learning affinities and dependencies for multi-target tracking using a crf model. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 1233–1240. IEEE.

Zhang, L. and van der Maaten, L. (2013). Structure preserving object tracking. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1838–1845.

Zhou, Y. and Tuzel, O. (2017). Voxelnet: End-to-end learning for point cloud based 3d object detection. *arXiv preprint arXiv:1711.06396*.

Zollhöfer, M., Stotko, P., Görlitz, A., Theobalt, C., Nießner, M., Klein, R., and Kolb, A. (2018). State of the art on 3d reconstruction with rgb-d cameras. In *Computer Graphics Forum*, volume 37, pages 625–652. Wiley Online Library.