Towards Multi-object Detection and Tracking in Urban Scenario under Uncertainties

Achim Kampker[1], Mohsen Sefati[1,*,†], Arya S. Abdul Rachman[2,*], Kai Kreisköther[1]
and Pascual Campoy[3]

[1] *Chair of Production Engineering of E-Mobility Components, RWTH Aachen University, Aachen, Germany*

[2] *Delft Center for Systems and Control, Delft University of Technology, Delft, The Netherlands*

[3] *Computer Vision and Aerial Robotics Group, Centre of Automatics and Robotics,
Universidad Politécnica de Madrid, Madrid, Spain*

Keywords: Multi Object Tracking, Perception, 3D LIDAR, Autonomous Driving, Probabilistic Filtering.

Abstract: Automated vehicles in urban scenarios require a reliable perception technology to tackle the high amount of uncertainties. In this paper a real-time framework for multi-object detection and manoeuvre-aware tracking is presented, where the application of 3D LIDAR for a cluttered urban environment is demonstrated. Our approach combines sensor occlusion-aware detection method with computationally efficient rule-based filtering and adaptive probabilistic tracking to handle uncertainties arising from sensing limitation of 3D LIDAR and complexity of the targets' movement. The evaluation results using real-world pre-recorded data and comparison with state-of-the-art shows that the presented framework is capable of achieving promising tracking performance in the urban scenarios.

1 INTRODUCTION

Advanced Driver Assistance Systems (ADAS) and Automated Driving (AV) have been the focus of many research activities for several decades. Achieving higher automation levels for AVs also imposes higher requirements on environment perception. By advancing from highways to urban and intercity scenarios, further challenges have to be met, especially with respect to the tasks of detection and tracking. In a typical urban scene the AV is surrounded by multiple traffic objects of different types (pedestrians, cyclists, cars, trucks, etc.) with different skills and movement patterns. The AVs should be able to detect and associate these objects with corresponding context information from the modelled scene and predict their feature behaviour for the subsequent tasks such as decision making and trajectory planning. The basis for this is the ability to classify between dynamic and static objects and keeping the tracking of dynamic objects in a continuous manner. Consequently, multi-object detection and tracking become essential for AVs perception.

The recently introduced compact 3D LIDAR scanner (Velodyne, 2007) is especially suitable for multi-object detection and tracking task, since it enables far-reaching high fidelity acquisition of surrounding spatial information, which is not possible with conventional sensing technologies. LIDAR-based perception tasks geared toward autonomous vehicle is a widely discussed topic. Among others, (Luo et al., 2016) suggest real-time capable LIDAR detection and tracking, (Chen et al., 2015) introduce model-based detection for surrounding vehicles, (Himmelsbach and Wuensche, 2012) propose a top-down bottom-up approach to enhance detection and tracking result while simultaneously doing classification. Notwithstanding, there are comparably fewer literatures, which address the holistic integration of LIDAR perception tasks aimed toward practical use in the urban situation. (Zhang et al., 2011), (Wojke and Haselich, 2012), and (Choi et al., 2013) notably propose a complete scheme of Multi Object Tracking (MOT). However, these implementation does not specifically target the use-case of urban driving, and limitation of vehicle embedded computer is not necessarily taken into account.

[*] These authors contributed equally to the paper

[†] Corresponding author

Kampker, A., Sefati, M., Abdul Rachman, A., Kreisköther, K. and Campoy, P.
Towards Multi-Object Detection and Tracking in Urban Scenario under Uncertainties.
DOI: 10.5220/0006706101560167
In *Proceedings of the 4th International Conference on Vehicle Technology and Intelligent Transport Systems (VEHITS 2018)*, pages 156-167
ISBN: 978-989-758-293-6

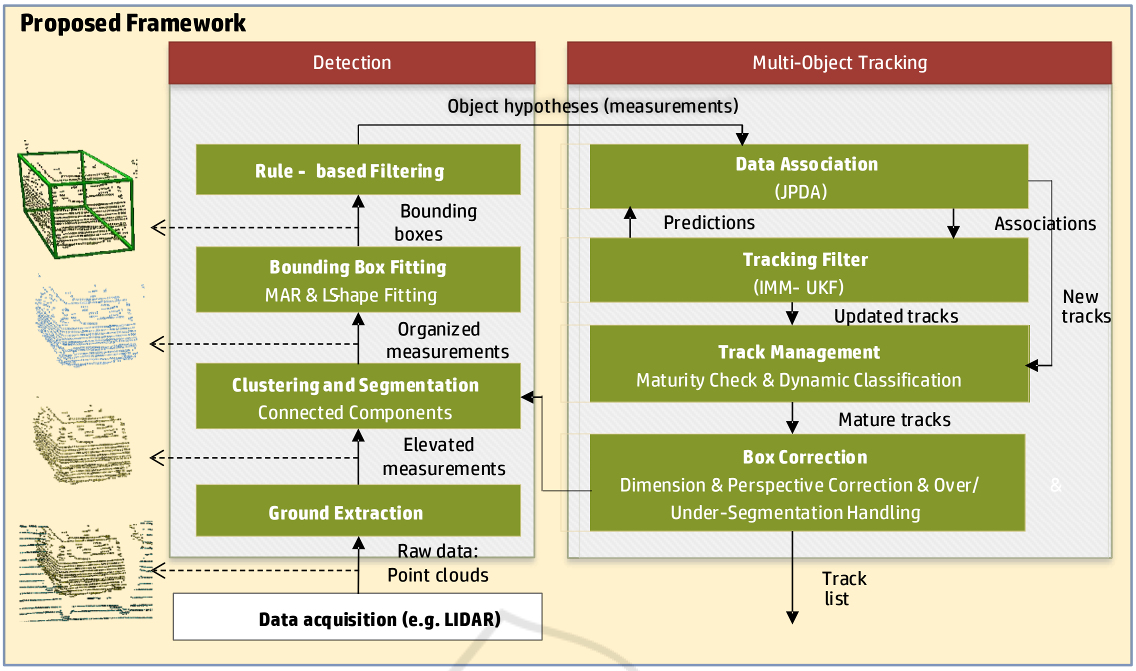

Figure 1: Structure of the proposed multi-object detection and tracking.

This paper presents a complete framework and pipeline for multi-target object detection and tracking for urban scenarios based on a hybrid approach, where both grid-based and object-based techniques are combined. The proposed framework is designed to cope with multiple targets, cluttered environment (i.e. the objects are close to each other), occlusions and uncertainties by applying a set of computationally efficient strategies. The main advantage of this framework is robustness against common uncertainties in urban scenarios with the application of probabilistic approaches for data association and tracking filters. In addition to that, promising tracking reliability with dynamic classification is achieved. The input of the framework is 3D LIDAR raw data in the form of point cloud, while the output is a track list of associated objects with their corresponding dynamic and geometric properties together with association probabilities. In this work, the framework is demonstrated by data acquired from a state-of-art Velodyne HDL-64E LIDAR sensor, because of its advantages regarding accuracy, field of view, and sampling rate of three-dimensional environmental measurements. However, this framework is also applicable to other sensor technologies, since it mainly relies on generic grid-based and object-based approaches.

The rest of the paper is structured as follow: an overview of the structure and main functions of the framework is given in section 2. Section 3 describes the detection part, where the non-ground measurements are extracted and object hypotheses are generated. The multi-object tracking with its main components are presented in section 4, and the further post-processing functions for dimension correction are presented in section 5. Finally, the framework is evaluated in section 6 by the use of raw data from KITTI data set (Geiger et al., 2013) and MOT16 (Milan et al., 2016) evaluation metrics.

2 SYSTEM ARCHITECTURE

The framework can be divided into two main function categories: detection and tracking. The input of the detection part is a 3D point cloud, which has to be divided into non-ground and elevated measurements. This is accomplished by a slope-based ground removal approach and a subsequent filtering process. In a further step, object hypotheses for the tracking targets are generated in a clustering step. The objects of interest are extracted by means of a subsequent feature-based bounding box fitting and a rule-based filtering.

The tracking is done based on centroid tracking of generated bounding boxes with four main steps: data association, tracking filters, tracking management and bounding-box correction. In the association step, a set of object hypotheses is determined, which correspond to the predicted measurements based on the already established tracks. In a case of a possible association, the track is updated with an associated measurement, otherwise a new track is created. The prediction and update steps are done by means of tracking filters. The track management maintains all tracks; labels their maturity and filters out the non-feasible and old ones. Finally, the bounding-box correction assign valid bounding box dimensions to the mature tracks and uses the track history in order to update this information with new measurements. Figure 1 illustrates the framework structure with its main components. The algorithmic implementation of the framework is discussed in Section 3 to 5.

3 DETECTION

3.1 Ground Extraction

The ground extraction is an important pre-processing step, in which all incoming 3D points are binary labelled into two groups of ground and non-ground elements. The term ground is considered as navigable and reachable area, which surrounds the ego-vehicle. The urban scenarios may have different types of terrain. Therefore, the ground extraction must be able to handle non-flat, sloped and uneven surfaces. This module is the first component in the whole framework and has to deal with entire data coming from the sensor. Thus, its computational performance is an important aspect. For this goal, a combination of channel-based and scan-based approach is used in this work. In the proposed approach, a slope-based and channel-wise classification is performed on a polar grid and by means of the modified technique from (Himmelsbach and Wuensche, 2012). After an initial estimation of the ground surface is achieved, the interrelationship between channels and the consistency of the estimated ground is checked subsequently. For this aim, the height comparison between the neighbour cells in a polar grid is applied. The estimated ground surface is smoothed and the missing spatial information are filled by applying a median filter. Figure 2 shows the result of ground extraction and the effect of consistency check and median filter.

3.2 Clustering

The first step towards the hypotheses generation is to divide an unorganised and non-ground point cloud into the smaller parts. This step is called clustering and can be done in 3D, 2.5D and 2D. Since the computational cost for 3D clustering is usually so high, the clustering problem is treated as a 2D-problem by mapping all elevated 3D points to a 2D grid as it is proposed by (Levinson et al., 2011) and (Himmelsbach et al., 2010). The clustering in this work is accomplished by applying the Connected Component Clustering (Pfaltz, 1966) in Cartesian grid representation. This approach has its origin in computer vision for clustering the 2D binary images. However it has also been used for 3D LIDAR point cloud (cf. (Rubio et al., 2013). The Connected Component Clustering is applied to the point cloud based on the row-to-row approach. This approach makes two passes: 1) storing equivalences and assigning the temporary label for ”connectedness” of cells and 2) determining the relation between the equivalence classes and replacing the temporary labels.

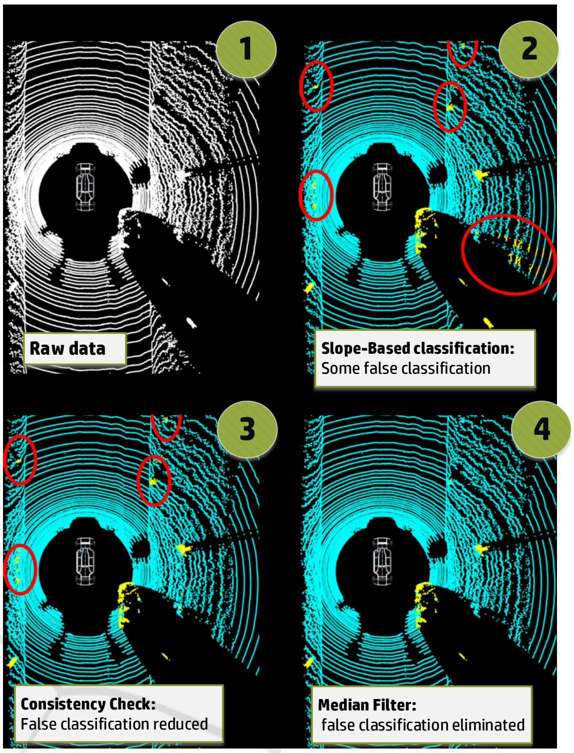
1
Raw data
2
Slope-Based classification:
Some false classification
3
Consistency Check:
False classification reduced
4
Median Filter:
false classification eliminated

Figure 2: Main steps of the ground extraction: 1) raw data, 2) result of the ground estimation, 3) result of consistency check and 4) result of median filter.

Initially, the whole grid is checked for the occupancy and the cells are assigned with two initial states for empty (0) and occupied (-1). Each cell in the grid is examined for the connectivity by checking the occupied neighbour cells and using the spatial kernel K_s with size s. If the target cell belongs to the same region as the neighbour cells, the same cluster ID is assigned to it. Otherwise, the new ID is created by incrementing the ID by one. If the connected neighbour cells are already assigned with different cluster IDs, the minimum ID will be chosen as the target cell. After all occupied cells are assigned with a cluster ID, the second pass uses a union-find data structure to replace each cell label with its equivalent class and avoid multiple labels for a single connected region. Since the size of the kernel defines the maximum spatial distance between two connected cells within the same cluster, it is responsible for over- and under-segmentation error, which is taken into account later in section 5.3. Figure 3 shows three partial snapshots of the same scene, where the ground classification, ground removal and clustering have been applied subsequently. Different colours in the left part of the figure refer to the different clusters in the scene.

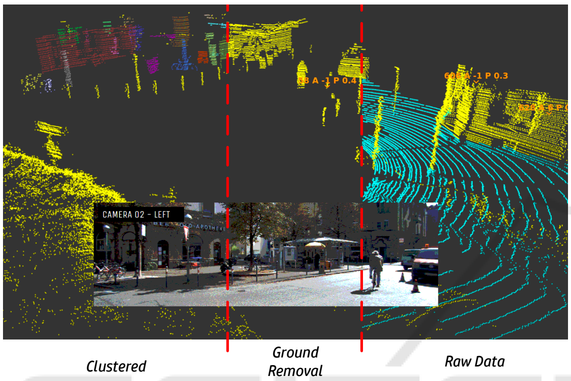

Figure 3: Measurement pre-processing: ground removal and clustering.

3.3 Bounding Box Fitting

The clusters of 3D points which are recognised as objects provide limited information about the pose of the objects. Moreover, some parts of the objects might be seen only partially with LIDAR. Thus, a further process is required in order to formulate a better hypothesis about each object. In order to tackle this, a 3D bounding box representation, which gives better information about the dimension and orientation of the detected objects is chosen. Generally, there are two groups of approaches for bounding box fitting: feature-based and model-based approaches (Chen et al., 2015). The model-based approaches offer more accuracy and better results than the feature-based approaches, due to the use of rectangle or cuboid models together with the application of optimization or sampling techniques. However, they suffer from high computational cost and are therefore not suitable for urban scenarios with a high number of detected objects. Thus, a feature-based method is proposed for this work, where its result is improved continuously by integrating the tracking results back to the detection.

First, the Minimum Area Rectangle (MAR) (Freeman and Shapira, 1975) is applied to the 2D cluster in order to create the initial bounding box. The height information of each cluster is retained, by deriving the difference between the highest and the lowest point. This information can be used for forming of 3D oriented bounding box (cf. Figure 4). The MAR approach is sufficient for most of well-defined clusters. However, this approach might fail for occluded objects and leads to erroneous heading angle, where there are not enough measurement points available. To tackle this issue, a feature-based L-shape fitting approach is applied, which corrects the box orientation. Similar to (Ye et al., 2016), the L-shape fitting is done by extracting the outer contour of the cluster. As it is shown in Figure 5, the farthest outlier points x_1 and x_2 are selected, which are laying on the opposite side of the object facing the LIDAR sensor. The line L_d is drawn between two points and an orthogonal line L_0 is obtained with the maximum distance d_{max} and angle close to 90 deg by applying the Iterative End Point (IEPF) algorithm. The corner point x_3 can be found near to the L_0, which forms together with x_1 and x_2 an L-shape polygon. The heading is then described by the longest line of the L-shape which is a valid assumption for the most of traffic objects such as vehicles and cyclist.

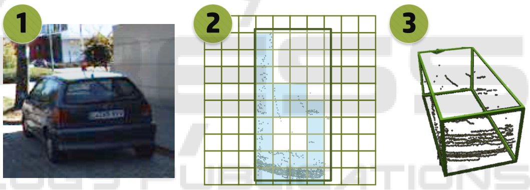

Figure 4: MAR box fitting with embedded height.

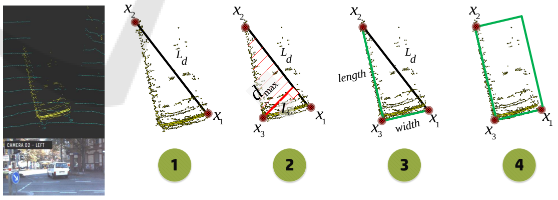

Figure 5: L-shape fitting for a more accurate bounding box fitting.

4 MULTI-OBJECT TRACKING

4.1 Tracking Algorithm

The object tracking refers to the problem of determining the number of objects of interest, their identities and their states based on sensor measurements. In this work, the states are position, velocity and yaw angle and yaw rate. The result of a tracking algorithm relies mainly on two parts of data association and tracking filter. There are three important aspects, which are needed to be taken into account by selection of optimal data association for urban scenarios: 1) handling multiple objects with different movement patterns, 2) handling cluttered environment and 3) computational efficiency.

Based on these requirements the Joint Probabilistic Data Association (JPDA) (Bar-Shalom and Li, 1995) is chosen for this work. For the filtering part, typical approaches are based on Bayesian Filtering such as Kalman and Particle Filters, which deal with a single motion model to predict and update object states. Even by the presence of a perfect motion model representing the object trajectory, there is no guarantee, that the object always follows this model. The objects in urban traffic may have different movement patterns and switch between different maneuvers described by different models. Therefore, a maneuver-aware tracking approach which is capable of dealing with multiple motion models has to be applied. Among different maneuver-aware target tracking algorithms, the Interacting Multiple Model (IMM) (Genovese, 2001) based on an optimal Kalman Filter shows a promising performance. Beside an improvement in the filtering process, an additional advantage of IMM is the dynamic classification.

Further advantages can be achieved by application of non-linear models, which also requires non-linear estimation filters such as Extended Kalman Filter (EKF), Unscented Kalman Filter (UKF) or Particle Filter (PF). (Gao et al., 2012) and (Djouadi et al., 2005) have shown that IMM-UKF has a better performance than IMM-EKF. Thus, a tracking algorithm is proposed in this work based on a coupled filter JPDA-IMM-UKF for three motion models: Constant Velocity, Constant Turn Rate, and Random Motion, which can deal with the tracking of multiple manoeuvring objects in a cluttered environment. It can be noted, that there are already similar implementations for coupled filters such as IMM-UK-PDA(Schreier et al., 2016a), IMM-UK-MHT (Blackman, 2004) and IMM-PF (Wang et al., 2015). However, the JPDA-IMM-UKF is not applied for the LIDAR and urban scenarios yet, which is the contribution of this work.

The JPDA-IMM-UKF algorithm consists of four main steps: 1) Interaction, 2) Prediction-and-Measurement Validation Step, 3) data Association-and-Model-Specific Filtering Step and 4) Mode Probability Update-and-Combination Step. Compared to existing closely related implementation PDA-IMM-UKF applied to RADAR (see (Schreier et al., 2016b)), JPDAF is used instead of the conventional PDAF since we are performing multi-object tracking and considering the presence of clutter. The association probability β between each track t and measurement j considering all feasible joint association events θ across all measurements Z_k is given by (Bar-Shalom and Li, 1995):

$$\beta_{jt}(l) \equiv = \sum_{\theta:\theta_{jt}\in\theta} P\{\theta_k | Z_k\} \tag{1}$$

With computed Kalman gain $\mathbf{K}$, and innovation term $\mathbf{z}_k - \hat{\mathbf{z}}_k^-$; the updated system states $\hat{\mathbf{x}}$ become:

$$\begin{aligned} \hat{\mathbf{x}}_{k|k} &= \hat{\mathbf{x}}_{k|k-1} + \mathbf{K}_k(\mathbf{v}_k) \\ \text{with } \mathbf{v}_k &= \sum_{m=1}^{N_v} \beta_{m,k}(\mathbf{z}_k - \hat{\mathbf{z}}_k^-) \end{aligned} \tag{2}$$

4.2 Track Management

There are two main objectives considered for track management in this work: 1) maturity check and pruning, and 2) dynamic classification.

4.2.1 Maturity Check and Pruning

Each track is assumed as ”immature” once it is initialised based on the first association with a measurement. Subsequently, the status will be changed to ”mature” after it is seen for a more than $n_i = 3$ consecutive time frames. As long as the track is not initialised or associated with a wrong measurement out of the valid range, its state is defined as ”invalid” and set to zero. Once it is initialised, the state is set to ”initialising” and incremented by one. After it is seen in multiple consecutive frames, it is assumed as mature and its status is set to ”tracking” and incremented further. The track enters the ”drifting” status by a further state increment, as soon as its measurement is lost at the next time step. Once a feasible measurement is found in the next frame the status is changed back to ”tracking” and the state is decremented. Otherwise, the state is incremented up to $n_d = 3$ frames, where the status is reset to ”immature”.

One of the undesirable traits of JPDA filter is its tendency to coalesce when the neighbouring track shares the same measurement. In order to prevent duplicate tracks associated with the same measurement, a hybrid pruning approach is developed based on track history and Euclidean distance. The track is considered as duplicate if the cumulative sum of standard deviation is less than a predefined threshold called history gating level. In this case, the track with shorter life time is deleted. Furthermore, the Euclidean distance between each track pair is calculated and checked against the physically possible distance in urban scenes. If the distance is less than a threshold, the newer track will be deleted.

4.2.2 Dynamic Classification

Classifying the dynamic objects is a non-trivial task due to the presence of measurement and detection noise, occlusion and therefore jumping object frames. Thus, the velocity thresholds are not sufficient for dynamic classification and further information has to be taken into account. Similar to (Schreier, 2017) the classification is done by incorporating both velocity thresholds and IMM probabilities. The object is classified as static, when it has a zero or close to zero velocity together with a higher probability of a Random Motion Model. Since the estimated velocity is not necessarily smooth, an average velocity of previous $n = 3$ frames is taken into account for the classification.

Figure 6 shows an intersection in an urban scene with different traffic objects waiting behind the red traffic light, a traffic object turning to the right and a further traffic object crossing the intersection. It can be seen that the waiting traffic objects are classified as static, while the turning object is assigned with a ”dynamic” state. It can also be seen that the crossing vehicle is in a ”drifting” state, since there is no measurement available at this frame for an association.

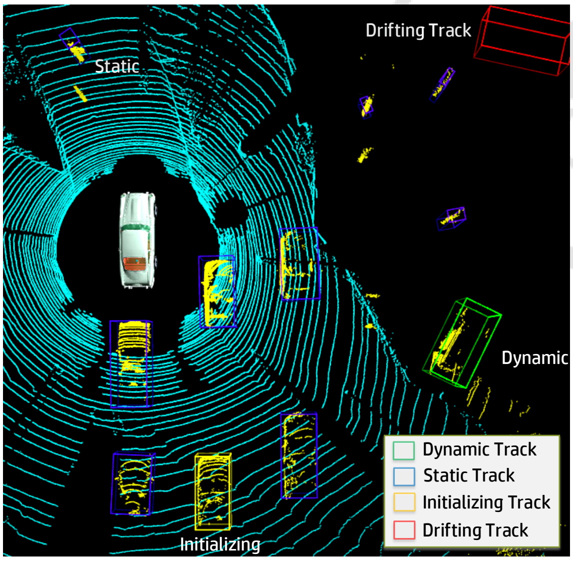

Figure 6: Track Management: colour-coded classification of track maturity and dynamic classification. ”dynamic” track (green) indicates the object is moving, ”static” track (blue) indicates the object is stopping together with ego-vehicle, ”initializing” track (yellow) indicates the track is not yet mature since the object has just entered the sensor frame, and ”drifting” track (red) indicates the track is about to be lost because it is entering the blind spot area.

5 BOUNDING BOX CORRECTION

The JPDA-UKF-IMM algorithm is designed to track the centre of the fitted bounding box, which is technically a position tracker. Since the bounding box dimensions are not among the filtered states, a further step is required in order to associate the correct geometric features of the box. The LIDAR sensor is not able to see the whole object in each frame due to occlusion caused by the target object itself (i.e. self-occlusion) or a nearby blocking object. This may lead to over- or under-segmentation as well as dimension changes over time. In order to tackle this problem, the result of tracking algorithm can be used to improve the bounding box fitting in three steps explained in the following subsections.

5.1 Dimension Updating

The dimensions of bounding boxes may change due to object occlusions or changes in observation positions of ego vehicle as it is shown in Figure 7. A dimension history can be integrated for monitoring the dimension changes over time and allowing an update for mature tracks with ”tracking” status under two main assumptions: 1) the bounding box is not allowed to shrink and reduce its width and length and 2) the bounding box is not allowed to have sudden changes in heading angle or moving direction. If there are more than one bounding boxes associated with a single track, the one with higher association probability is taken. In a case of equal probabilities, the one in the nearest Euclidean distance is chosen. Furthermore, it is checked, if there is an approximately same number of points in the track and associated measurement. The dimension information is kept and stored for the track until the next update for each mature track.

5.2 Perspective Correction

In addition to the dimension of the bounding box, its position has to be updated based on observation viewing angle. For this goal, the new bounding box is shifted with respect to nine reference points proposed by (Schueler et al., 2012) illustrated in Figure 8 by the green dots. The reference points describe the best seen corner or edge and distinguish between the front, rear and side of the target. The shifting process is done under the assumption that the reference points of the old and the new bounding box significantly overlaps. An instance of perspective correction is shown in Figure 9.

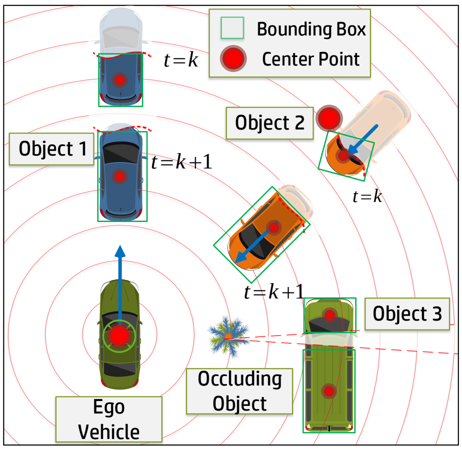

Figure 7: Effect of occlusion on bounding box fitting.

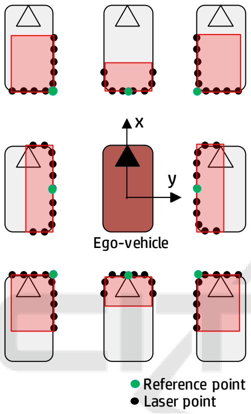

Figure 8: Changes in the dimension of a fitted bounding box of the target with respect to viewing angle. The dark points indicate the LIDAR points; the green points indicate the reference points and red boxes show the fitted bounding boxes.

5.3 Over- and Under-segmentation Handling

The occlusion objects in traffic scenes especially in urban scenarios may cause over- and under-segmentation. This error may also be caused by clustering process, where the kernel size is not adjusted well to the current scene. In order to solve this problem, the top-down approach from (Himmelsbach and Wuensche, 2012) is applied, where the tracking information is re-used in clustering and box fitting. An over-segmentation can be identified by inspecting if the new-found clusters overlap with the predicted position of the bounding box. In case of significant overlaps, the clusters have to be merged (c.f. Figure 10). An under-segmentation occurs, when predicted tracks are within one clustered region. In this case, the kernel size of the region has to be reduced iteratively until the correct number of expected clusters is achieved.

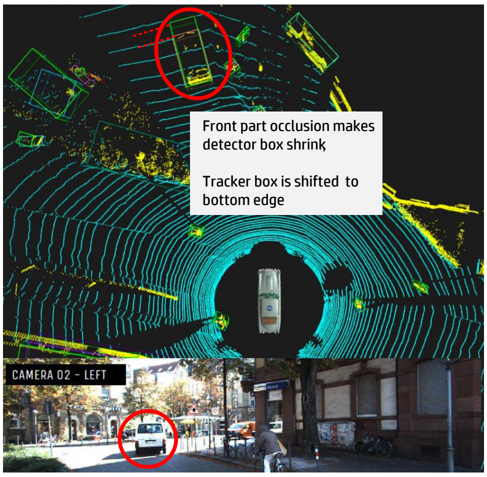
Front part occlusion makes detector box shrink
Tracker box is shifted to bottom edge
CAMERA 02 - LEFT

Figure 9: Instance of perspective correction. Bounding box of a self-occluded van located in front of ego-vehicle is shifted downward so that the van dimension is retained.

6 EVALUATION

6.1 RAW Data and Ground Truth: KITTI Dataset

The proposed multi-object detection and tracking algorithm can be evaluated in real world scenarios by using non-synthetic data. For this purpose, the KITTI datasets (Geiger et al., 2013) are used. This public benchmark provides the recordings of Velodyne HDL-64E sensor, among other sensors in different urban scenarios in the city of Karlsruhe, Germany. It also includes real-world traffic situations and range from highways over rural areas to inner-city scenes with high-quality hand-labelled annotation.

In order to evaluate the relevant urban scenarios, KITTI datasets within category ”City” are used. The collection of datasets consist of 10 different driving scenarios with the cumulative frame number of 2111 frames and 188 unique traffic objects. The composition of each dataset is represented in Table 1 and Figure 11. Note that the datasets from ”City” category contain lots of vulnerable road users compared to other sets and thus more representative for urban scenarios.

6.2 Benchmark Results and Discussion

The tracker performance is evaluated by using MOT16 benchmark method (Milan et al., 2016) which combines the CLEAR quantitative metrics (Bernardin and Stiefelhagen, 2008) augmented with set of Track Quality Measures (Wu and Nevatia, 2007). It is important to note the used datasets are not uniform: the driving

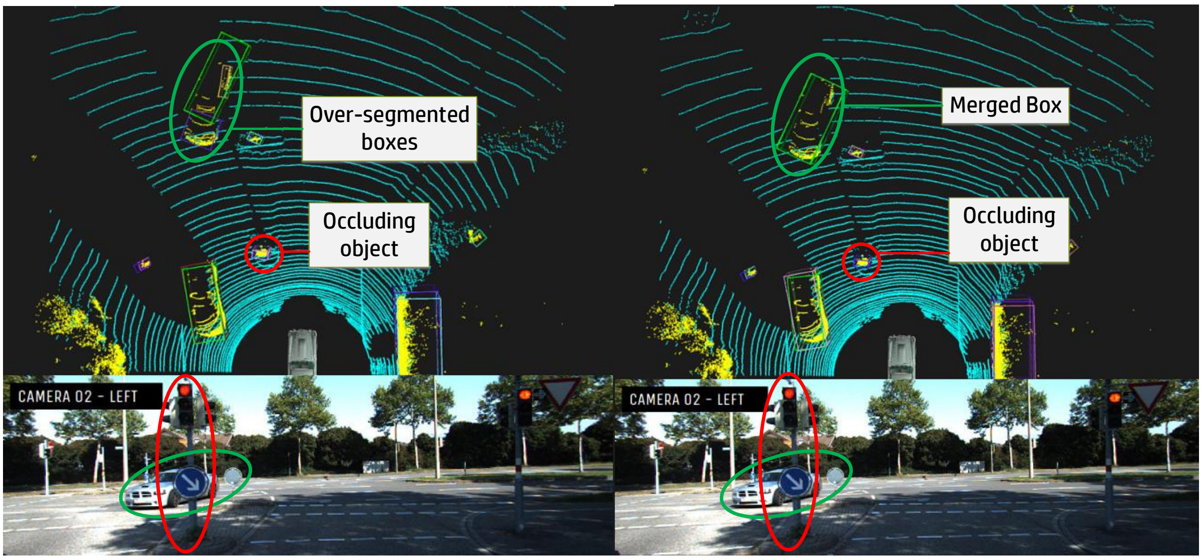

Figure 10: Example of over-segmentation handled by box merging.

Table 1: Evaluation Datasets.

Dataset	Frame count	Unique obj.	No. of box
0001	106	11	142
0002	75	2	45
0005	152	14	473
0009	441	82	1413
0013	142	4	101
0017	112	5	84
0018	268	12	196
0048	20	7	81
0051	436	40	381
0057	359	12	471
Sum	**2111**	**189**	**3387**

scenario along with the object compositions and movement types may vary significantly as dataset changes. This is intentional as tracking methods can be heavily overfitted on one particular dataset and potentially introduce evaluation bias (Milan et al., 2016). Therefore, the individual dataset evaluation result is a more representative indicator to reflect the framework performance. Nevertheless, it is still useful to view the averaged score as shown in Table 1 to provide the reader with the information about tracker overall performance.

The Multi Object Tracking Accuracy (MOTA) score reflects that the tracker has a reasonable high degree of accuracy with a 86% overall score. The score is lowered mainly by the number of False Negative (FN), since the number of False Positive (FP) and ID Switch (IDSW) are comparatively low. The Multi Object Tracking Precision (MOTP) score is limited to 91%

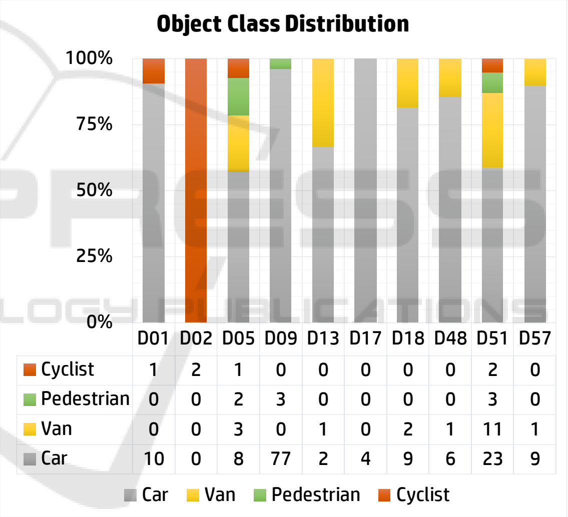

Figure 11: Distribution of object classes across all evaluation datasets.

which is an expected result, since despite a perfect tracking, only partial dimensional information can be derived when an object enters the sensor frame from a far distance; so the tracking precision would always be low for the first few time frames.

A significant deviation across all datasets can be seen due to datasets contain varying urban scenario. Nevertheless, the average results highlight that the tracker yield higher rate of Mostly Tracked (MT) than Mostly Lost (ML). The recall-rate (i.e. sensitivity) and precision (i.e. positive predictive value) indicate that the tracker hypotheses possess a high degree of relevance to the actual object, where the lower recall rate is consistent with the number of FNs counted in total. Fi-

Table 2: Overall evaluation result.

(a) CLEAR Metrics

Metrics	Value
MOTA	86.12% ± 6.00
MOTP	91.01% ± 5.03
FP (sum)	65
FN (sum)	406
IDSW (sum)	75
Total Obj. Instances	3387
Total Frame	2111

(b) Track Quality Measures.

Metrics	Value
Mostly Tracked	70.64% ± 17.47
Mostly Lost	9.33% ± 8.10
Recall rate	88.92% ± 10.18
Precision rate	98.43% ± 2.73
Fragmentation	211

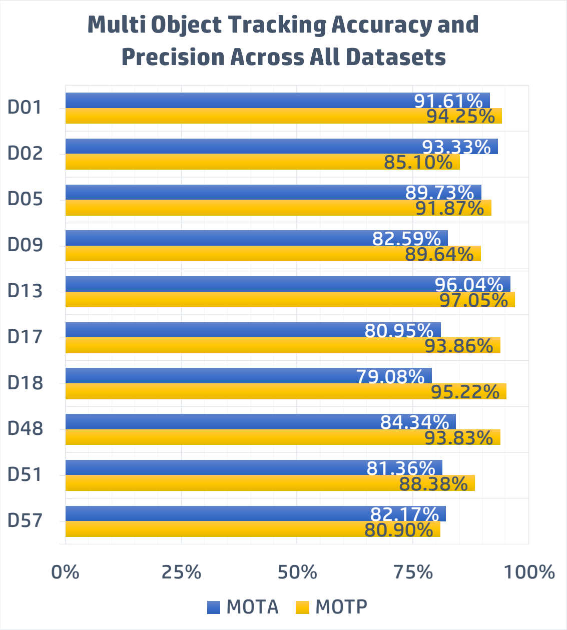

Figure 12: Per dataset MOTA and MOTP scores.

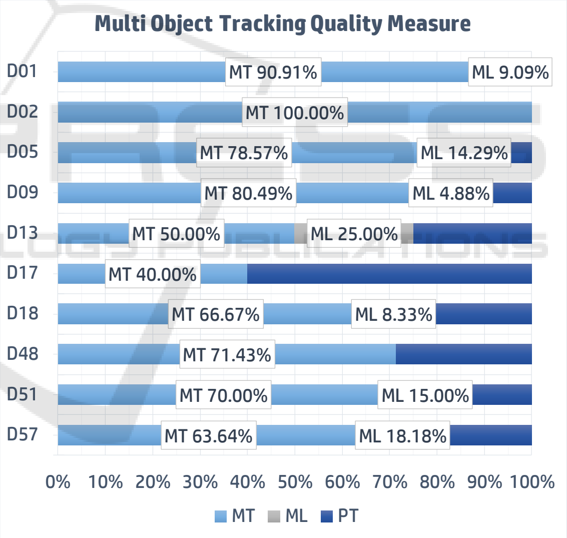

Figure 13: Per-dataset Track Quality Measures. ”PT” refers to Partially Tracked, which is not classified as either MT or ML.

nally, the number of Fragmentation is a subset of the FNs; here we see that more than half of FNs is caused by track lost. Note that the lost tracks are recoverable (i.e. not all FNs are the results of complete detection failure across all frames). The tracking performance (MOTP and MOTA) for individual dataset are shown in Figure 12, the Quality Measures are shown in Figure 13, and the base metric scores are listed in Table 2.

Some datasets are discussed in details, in order to provide the reader with physical meaning of the results: In Dataset 0005, 78% of tracks are considered to be MT, while a relatively large number of ML tracks are available. Here the ego vehicle is moving in a curved urban road which causes a constant change in a reference sensor frame. Combination of self-occlusion (cf. Figure 9 for visual depiction) and a relative fast turning rate of the target objects increases the uncertainties of the target spatial position. Therefore, reduced tracking accuracy and fragmentation errors can be seen in this scenario. Dataset 0009 represents a typical complex detection and tracking scenario: it has a large number of frame counts with a great number of unique objects compared to other datasets. In this dataset, the ego vehicle made a 90-degree turn (i.e. a sudden change of sensor frame) and stopped at a 4-way junction with a persistently occluding object. The situation can be observed in Figure 10.

Nevertheless, handling uncertainties is one of the main contributions of this work: here we see the MOTA score reflects that the use of JPDA filter enables the tracker to form hypotheses with sufficient accuracy. The results are obtained despite cluttered environment and manoeuvring targets. Performance reduction is found in certain scenarios (Dataset 0017 and 0051). However, since more than 80% of the tracker hypotheses are considered as MT with only 5% considered as ML, adequate robustness against persistent- and self-occlusion of the target objects can be found, regardless of sensor frame change, turning cars and other occluding objects.

Table 3: CLEAR comparison of state-of-art 3D LIDAR trackers.

Method	MOTA	MOTP	FN	FP
Proposed Framework	86.12 %	n/a. in m	11.89 %	1.92 %
Tracking Circle (Ye et al., 2016) (averaged)	86.5%	< 0.2 m	3.5%	8.0%
Energy-based (Xiao et al., 2016)	84.2 %	< 0.12 m	5.8 %	2.77 %
BUTD (Xiao et al., 2016)	89.1 %	< 0.16 m	2.6 %	7.6 %
Generative (Kaestner et al., 2012)	77.7 %	< 0.14 m	8.5 %	10.1 %

As a summary, the benchmarking process yields a better understanding of the tracker performance in a large variation of urban scenarios with different classes of traffic objects. Cars, vans and pedestrians are tracked reliably by an average of above 86% with the proposed framework. Quality Measures support the scores of the CLEAR metrics: MT tracks outnumber ML tracks by a significant margin in all datasets, including datasets with complex scenarios. These scenarios contain constant sensor frame change, persistent occluding object and actively-manoeuvring targets. Note that we see that the tracking accuracy and performance may decrease as the number of objects increases. However, in this situation the Quality Measures indicate that the majority of objects are still covered adequately by the tracker.

7 COMPARISON TO STATE-OF-THE-ART

The use of both established MOT metrics and public datasets are also useful to enable objective comparison to the performance of state-of-art trackers. The utilised metrics, namely the MOTP, MOTA, MT, ML, FN and FP are common measures for tracking performance. Publicly ranked benchmarks (see KITTI Object Tracking Evaluation 2012 (Geiger et al., 2013) and 2017 MOT challenge (Leal-Taixé et al., 2017)) use these metrics, as well as numbers of MOT-related literatures such as (Zheng et al., 2012; Bernardin and Stiefelhagen, 2008; Piao et al., 2016; Wen et al., 2015).

Compared to camera tracking, there is notably fewer LIDAR literature which put significant concern on evaluation using established metrics. Some notable publications which use both Velodyne and CLEAR as metrics are that of (Ye et al., 2016) which use geometric-based tracking circle method, (Xiao et al., 2016) which use point assignment task based on energy function, (Spinello et al., 2011) which use Bottom-Up Top-Down Detector (BUTD), and (Kaestner et al., 2012) which use Generative Object Detection and Tracking. The comparison results can be seen in Table 3. These works use different criteria to compute the MOTP. Our approach takes into account the position and dimensional integrity of the tracked objects, thus the bounding box overlap ratio is used. Meanwhile, these works consider only the precision of centre point of a detected object, so the MOTP is based on Euclidean distance error instead. In addition, only work of (Ye et al., 2016) deals with a sensor mounted on a moving car; the other three use the dataset recorded on ETH Zurich Polyterrasse, which deals with a static reference frame in university canteen scenery and populated mainly with pedestrians. While results of (Ye et al., 2016) would be the best control comparison to this thesis work, it only uses 2 datasets with unspecified ground truth details.

A general overview indicates, that our proposed approach has a comparable accuracy (± 3% differences) to state-of-art, but accompanied with quite larger percentage of FN (11.89 % vs 2.6% with that of BUTD). In the previous section, it has been found that a large number of FN is contributed by the datasets with complex scenario (mainly Dataset 0017). Nevertheless, if we inspect other datasets individually, the FN rate would be on par (2-7%) with other approaches. Therefore, a comparison with standardised datasets are needed to give more insight, if the compared state-of-art works exhibit a similar performance reduction in significantly complex urban situations.

8 CONCLUSION

An integrated multi-object detection and tracking framework has been introduced in this paper. The framework is especially designed for the use of environment perception in the urban scenarios with the associated uncertainties of 3D LIDAR sensor measurements. However, this framework can be used for other sensor technologies as well.

The detector is able to cope with occlusion and handle under/over-segmentation, by receiving the additional information from the tracker. The tracking algorithm itself employs probabilistic data association and filtering based on a coupled IMM-UKF-JPDA filter, which allows a manoeuvre-aware multi-object tracking under uncertainties in a cluttered environment. Moreover, geometric properties of the tracks are updated in a post-processing part by means of computationally low demanding rule-based filtering and the the use of box frame history.

Finally, the framework is evaluated with the help of established MOT16 metrics, which shows that the tracking performance is favourable in a variety of prerecorded real-world urban scenarios. Since the framework is designed and found to run in real-time (under 100 ms), we expect that our framework is applicable for autonomous vehicles. However, the performance of this framework can be increased in future works by further code optimisation, applying parallel programming techniques and further fitting algorithm for V and U shape traffic objects.

REFERENCES

Bar-Shalom, Y. and Li, X. R. (1995). *Multitarget-multisensor Tracking: Principles and Techniques*. Yaakov Bar-Shalom.

Bernardin, K. and Stiefelhagen, R. (2008). Evaluating multiple object tracking performance: The CLEAR MOT metrics. *Eurasip J. Image Video Process.*, 2008.

Blackman, S. (2004). Multiple hypothesis tracking for multiple target tracking. *IEEE Aerosp. Electron. Syst. Mag.*, 19(1):5–18.

Chen, T., Dai, B., Liu, D., Fu, H., Song, J., and Wei, C. (2015). Likelihood-Field-Model-Based Vehicle Pose Estimation with Velodyne. *IEEE Conf. Intell. Transp. Syst. Proceedings, ITSC*, 2015-Octob:296–302.

Choi, J., Ulbrich, S., Lichte, B., and Maurer, M. (2013). Multi-Target Tracking using a 3D-Lidar sensor for autonomous vehicles. *IEEE Conf. Intell. Transp. Syst. Proceedings, ITSC*, (Itsc):881–886.

Djouadi, M., Sebbagh, A., and Berkani, D. (2005). IMM-UKF algorithm and IMM-EKF algorithm for tracking highly maneuverable target: a comparison. *Proc. 7th WSEAS Int. Conf. Autom. Control. Model. Simul.*, pages 283–288.

Freeman, H. and Shapira, R. (1975). Determining the minimum-area encasing rectangle for an arbitrary closed curve. *Commun. ACM*, 18(7):409–413.

Gao, L., Xing, J., Ma, Z., Sha, J., and Meng, X. (2012). Improved IMM algorithm for nonlinear maneuvering target tracking. *Procedia Eng.*, 29:4117–4123.

Geiger, a., Lenz, P., Stiller, C., and Urtasun, R. (2013). Vision meets robotics: The KITTI dataset. *Int. J. Rob. Res.*, 32(11):1231–1237.

Genovese, A. F. (2001). The interacting multiple model algorithm for accurate state estimation of maneuvering targets. *Johns Hopkins APL Tech. Dig. (Applied Phys. Lab.*, 22(4):614–623.

Himmelsbach, M., von Hundelshausen, F., and Wuensche, H. (2010). Fast segmentation of 3D point clouds for ground vehicles. *Iv*, pages 560–565.

Himmelsbach, M. and Wuensche, H. J. (2012). Tracking and classification of arbitrary objects with bottom-up/top-down detection. *IEEE Intell. Veh. Symp. Proc.*, pages 577–582.

Kaestner, R., Maye, J., Pilat, Y., and Siegwart, R. (2012). Generative object detection and tracking in 3D range data. *2012 IEEE Int. Conf. Robot. Autom.*, pages 3075–3081.

Leal-Taixé, L., Milan, A., Reid, I., Roth, S., and Schindler, K. (2017). MOT17 Results.

Levinson et al. (2011). Towards fully autonomous driving: Systems and algorithms. *IEEE Intell. Veh. Symp. Proc.*, (Iv):163–168.

Luo, Z., Habibi, S., and Mohrenschildt, M. (2016). LiDAR Based Real Time Multiple Vehicle Detection and Tracking. 10(6):1083–1090.

Milan, A., Leal-Taixe, L., Reid, I., Roth, S., and Schindler, K. (2016). MOT16: A Benchmark for Multi-Object Tracking. pages 1–12.

Pfaltz, J. L. (1966). Sequential Operations in Digital Picture Processing. *J. ACM*, 13(4):471–494.

Piao, S., Sutjaritvorakul, T., and Berns, K. (2016). Compact Data Association in Multiple Object Tracking: Pedestrian Tracking on Mobile Vehicle as Case Study. *9th IFAC Symp. Intell. Auton. Veh.*, 49(15):175–180.

Rubio, D. O., Lenskiy, A., and Ryu, J. H. (2013). Connected components for a fast and robust 2D lidar data segmentation. *Proc. - Asia Model. Symp. 2013 7th Asia Int. Conf. Math. Model. Comput. Simulation, AMS 2013*, (September 2015):160–165.

Schreier, M. (2017). *Bayesian environment representation, prediction, and criticality assessment for driver assistance systems*. PhD thesis.

Schreier, M., Willert, V., and Adamy, J. (2016a). Compact Representation of Dynamic Driving Environments for ADAS by Parametric Free Space and Dynamic Object Maps. *IEEE Trans. Intell. Transp. Syst.*, 17(2):367–384.

Schreier, M., Willert, V., and Adamy, J. (2016b). Compact Representation of Dynamic Driving Environments for ADAS by Parametric Free Space and Dynamic Object Maps. *IEEE Trans. Intell. Transp. Syst.*, 17(2):367–384.

Schueler, K., Weiherer, T., Bouzouraa, M. E., and Hofmann, U. (2012). 360 Degree multi sensor fusion for static and dynamic obstacles. *2012 IEEE Intell. Veh. Symp.*, pages 692–697.

Spinello, L., Luber, M., and Arras, K. O. (2011). Tracking people in 3D using a bottom-up top-down detector. *Proc. - IEEE Int. Conf. Robot. Autom.*, pages 1304–1310.

Velodyne (2007). Velodyne's HDL-64E: A High Definition Lidar Sensor for 3-D Applications. 2007. *White Pap.*, page 7.

Wang, D. Z., Posner, I., and Newman, P. (2015). Model-free detection and tracking of dynamic objects with 2D lidar. *Int. J. Rob. Res.*, 34(7):1039–1063.

Wen, L., Du, D., Cai, Z., Lei, Z., Chang, M.-C., Qi, H., Lim, J., Yang, M.-H., and Lyu, S. (2015). UA-DETRAC: A New Benchmark and Protocol for Multi-Object Detection and Tracking.

Wojke, N. and Haselich, M. (2012). Moving Vehicle Detection and Tracking in Unstructured Environments. *2012 IEEE Int. Conf. Robot. Autom.*, pages 3082–3087.

Wu, B. and Nevatia, R. (2007). Detection and tracking of multiple, partially occluded humans by Bayesian combination of edgelet based part detectors. *Int. J. Comput. Vis.*, 75(2):247–266.

Xiao, W., Vallet, B., Schindler, K., and Paparoditis, N. (2016). Simultaneous Detection and Tracking of Pedestrian From Panoramic Laser Scanning Data. *ISPRS Ann. Photogramm. Remote Sens. Spat. Inf. Sci.*, III-3(July):295–302.

Ye, Y., Fu, L., and Li, B. (2016). Object Detection and Tracking Using Multi-layer Laser for Autonomous Urban Driving.

Zhang, L., Li, Q., Li, M., Mao, Q., and Nüchter, A. (2011). Multiple Vehicle-like Target Tracking Based on the Velodyne LiDAR. (2005).

Zheng, W., Thangali, A., Sclaroff, S., and Betke, M. (2012). Coupling detection and data association for multiple object tracking. *Comput. Vis. Pattern Recognit. (CVPR), 2012 IEEE Conf.*, pages 1948–1955.