Camera-based Semantic Enhanced Vehicle Segmentation for Planar LIDAR

Chen Fu[1], Peiyun Hu [2], Chiyu Dong[1], Christoph Mertz[2] and John M. Dolan[2]

***Abstract*— Vehicle segmentation is an important step in perception for autonomous driving vehicles, providing object-level environmental understanding. Its performance directly affects other functions in the autonomous driving car, including Decision-Making and Trajectory Planning. However, this task is challenging for planar LIDAR due to its limited vertical field of view (FOV) and quality of points. In addition, directly estimating 3D location, dimensions and heading of vehicles from an image is difficult due to the limited depth information of a monocular camera. We propose a method that fuses a vision-based instance segmentation algorithm and LIDAR-based segmentation algorithm to achieve an accurate 2D bird's-eye view object segmentation. This method combines the advantages of both camera and LIDAR sensor: the camera helps to prevent over-segmentation in LIDAR, and LIDAR segmentation removes false positive areas in the interest regions in the vision results. A modified T-linkage RANSAC is applied to further remove outliers. A better segmentation also results in a better orientation estimation. We achieved a promising improvement in average absolute heading error and 2D IOU on both a reduced-resolution KITTI dataset and our Cadillac SRX planar LIDAR dataset.**

I. Introduction

Object detection is one of the core functions in the perception system of autonomous vehicles. Each kind of sensor has its own properties. Cameras provide dense texture information and state-of-the-art computer vision algorithms can detect different kinds of objects. A vision system provides adequate semantic information about objects in an environment, but it lacks detailed measurements for a detected object, i.e., after detecting an object, it is difficult to locate and measure the dimension and heading of the object. Radar detects distance and speed of an object relative to the host vehicle. However, the RADAR only provides range and range-rate instead of accurate vehicle shape and heading. LIDAR provides accurate position measurement of an object, but raw points must be segmented to reflect the shape and pose. Thus a robust segmentation algorithm for LIDAR is required. A high-performance LIDAR provides a dense 3D-point cloud which contains detailed information about the shape and pose of a object, whereas a planar LIDAR provides a limited horizontal resolution with noise, from which it is hard to identify objects [1].

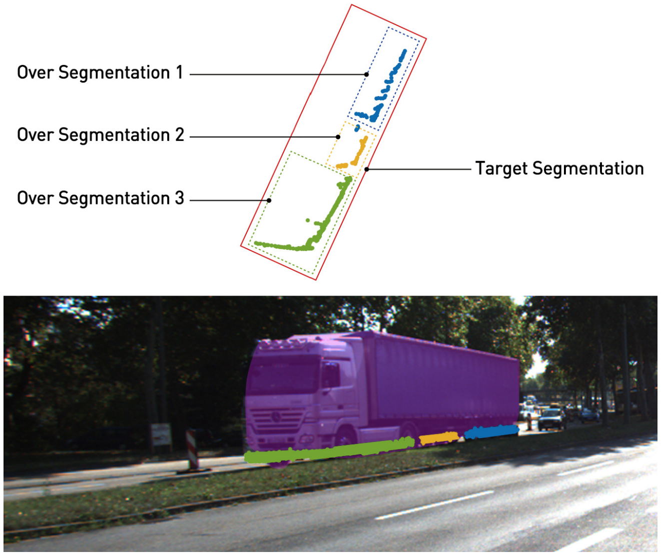


Fig. 1: Example of over-segmentation in 2D spatial segmentation. Ground truth truck is over-segmented into three smaller sub-segments by the LIDAR-based spatial segmentation method. We solve this problem by applying semantic information to overcome the over-segmentation, which enhances the 2D spatial segmentation on LIDAR point.

Using planar LIDAR can greatly reduce the expense of an autonomous driving car, which benefits the commercialization of self-driving cars. Considering its limited sensing capability, merely using the planar LIDAR does not satisfy the perceptual requirements to guarantee safety. Current planar LIDAR-based vehicle fitting methods usually have two main steps [2]: 1) Segmentation: separate the point cloud into groups which represent different objects; 2) Heading estimation: fit each segment of the point cloud with a proper geometry model, such as an L-shape, and estimate the corresponding heading angle. The performance of such a method is limited by these two steps. In the segmentation step, only the distances between LIDAR points are used to identify their clusters. But there is no guarantee that close points come from a single object, especially in a dense traffic condition where cars keep smaller gaps. Therefore, a LIDAR-only segmentation algorithm is easily affected by these outliers. In addition, over-segmenting is also a problem in LIDAR-only methods, i.e., two clusters of LIDAR points which are apart from each other may be identified as different objects, even if they come from a single object. Since a vision system can provide more object-level texture information than a LIDAR point cloud, a vision-based semantic segmentation can be used to separate points on different objects, hence eliminating outliers to refine the LIDAR segmentation.

Given an object detected by both LIDAR and camera, we fuse these detections together to achieve a better estimate of the shape and orientation of the object. To evaluate the performance of the segmentation, Intersection-Over-Union (IOU) rate and average absolute heading error (AAHE)

[1] Chen Fu and Chiyu Dong are with the Department of Electrical and Computer Engineering, Carnegie Mellon University, Pittsburgh, PA 15213, USA {cfu1, chiyud}@andrew.cmu.edu

[2] Peiyun Hu, Christoph Mertz and John M. Dolan are with the Robotics Institute, Carnegie Mellon University, Pittsburgh, PA 15213, USA peiyunh, mertz, jmd@cs.cmu.edu

are considered. This paper is organized as follows: Section II briefly reviews the prior work in vehicle detection and segmentation, using vision or LIDAR; Section III-A introduces Mask R-CNN for image-based detection; Section III-B describes our method for LIDAR 2D spatial segmentation; Section III-C discusses our LIDAR and vision fusion method and outlier removal algorithm using T-linkage RANSAC, which eliminates under-segmentation and over-segmentation; [3]. Sections IV and V discuss the experimental results of our methods and give conclusions.

II. Related Work

In 2007, the "Boss" vehicle developed by CMU outperformed other self-driving vehicles and won the DARPA Grand Urban Challenge. To accurately detect and track surrounding objects, multiple high-performance sensors including a 64-channel Velodyne LIDAR, camera and RADAR were mounted on the vehicle [4]. However, those high-performance range sensors are not affordable for the commercial autonomous vehicles aimed at average consumers. The latest version of CMU's autonomous driving test platform is a Cadillac SRX with multiple planar LIDARs, RADARs and cameras integrated on the vehicle. With automotive-grade sensors and an only slightly modified appearance, the platform can navigate and pilot itself on highways as well as in urban areas [5]. The challenge of planar LIDAR segmentation lies in the limited vertical field of view (FOV) as well as occlusion between objects. The limitation results in over-segmentation and under-segmentation of vehicles, which is quite ambiguous for the tracking and planning system.

In [6], the author proposed a Ramer algorithm based on planar LIDAR, which models a set of scanning points using several line segments. However, this method is unable to generalize to multiple-layer LIDAR sensors, as it needs the sequential information of the scanning points. The other family of distance-based algorithms considers the spatial information of scanning points [7], [8]. In [9], the clustering threshold is formulated into a linear equation, which considers ego-vehicle velocity, angle resolution and distance between LIDAR sensor and scanning points. Though these methods get rid of the assumption that the LIDAR points are scanned sequentially, it is computationally expensive to measure the distance between all pairs of data points. To solve this problem, [2] creatively combined a K-D tree with the distance-based algorithm, which improves the efficiency. However, the performance of these purely LIDAR-based segmentation algorithms highly relies on the quality of data points, which is incapable of handling the high-occlusion scenario. Our proposed method overcomes these shortcomings by fusing vision semantic information with LIDAR segmentation. Taking advantage of the object-level semantic information provided by the vision-based method, our method reduces the over-segmentation rate and boosts the 2D IOU and vehicle heading estimation.

Due to the application of deep convolution networks in the computer vision community, promising improvement has been achieved in vision-based object detection and segmentation [10]. In Mask R-CNN [11], multi-task frameworks are introduced to simultaneously detect, segment and classify objects. Though this family of deep architectures shows good performance in detection and segmentation tasks, it is difficult for these architectures to predict location and heading of objects in 3D real-world coordinates. To overcome these problems, the Deep3DBox framework proposed a new architecture that improves the orientation and dimension estimation in 3D with a modified multi-task VGG network [12]. However, such camera-based methods cannot recover accurate depth from projective transformation. In fact, as shown by most recent studies, camera-based methods still under-perform dense LIDAR-based ones at predicting 3D bounding boxes by a large margin (18.2% vs. 87.7% in terms of 3D IOU) [13]–[16]. Such a gap exists likely due to the inherent difficulty of estimating depth through images. Even with a stereo setup, the baseline is often limited (e.g. 0.54m for KITTI). On the other hand, LIDAR provides direct measurement of spatial shape and depth. To bridge the gap, our work tries combining camera-based semantic understanding and LIDAR-based shape to get the best of both worlds.

Taking advantage of the capabilities of different sensors, multiple-sensor fusion provides a more robust perception system for autonomous vehicles [17], [18]. In [19], [20], multiple planar LIDAR sensors, RADAR and cameras are fused together to detect and track on-road objects. To solve the over-segmentation problem in image segmentation, [21] fuses the front-view camera with high-end Velodyne LIDAR, which provides more efficient super-surfaces using geometric information. Even though these methods improve the detection and tracking performance of previous perception systems, they are incapable of correctly estimating the heading and dimension of detected vehicles. In this work, we combine the state-of-the-art vision-based instance segmentation algorithm and LIDAR-based vehicle segmentation, which further improves heading and dimension estimation.

III. Semantic Enhanced vehicle Segmentation

Accurately segmenting LIDAR data points is crucial for the perception system of an autonomous driving vehicle. Over-segmentation will mislead the tracking process, as multiple observations can be associated with one confirmed track. The ambiguity in data association prevents the Bayesian filter from converging, which gives inaccurate vehicle states to decision making and motion planning. In this section, we explain in detail how semantic information from vision enhances the conventional 2D spatial segmentation algorithm in planar LIDAR.

A. Vision Based Object Segmentation and LIDAR Point Projection

The performance of vision-based object detection and semantic segmentation results has increased dramatically recently, due to the application of deep convolution networks. Mask R-CNN is a parallel deep instance segmentation

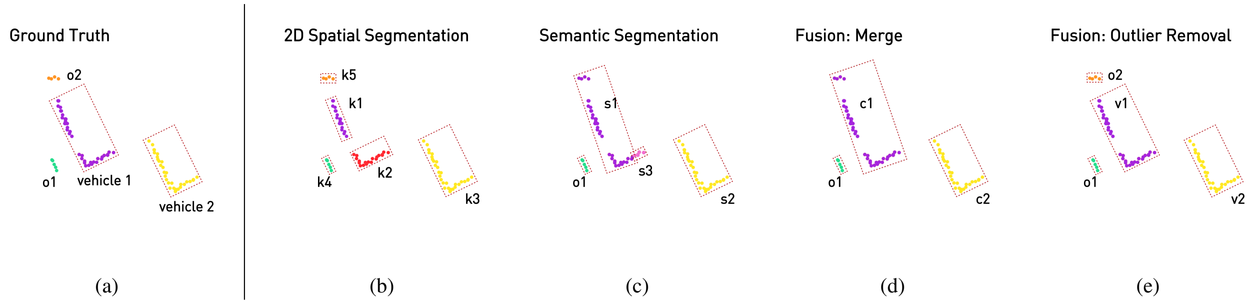


Fig. 2: Example of our proposed semantic enhanced method. (a) The ground truth segmentation on 2D planar LIDAR. (b) Over-segmentation case of 2D spatial segmentation algorithm. (c) Under-segmentation and over-segmentation caused by Mask R-CNN. (d)(e) Combined LIDAR and camera information applying our fusion method

framework which can simultaneously predict the class label, bounding box and pixel-wise classification of all objects in an input image [11]. The main contribution of Mask R-CNN is applying a Fully Convolutional Network (FCN) to the Region of Interest (ROI), which generates a pixel-wise mask in parallel with the object classification branch and bounding box regression branch. We take advantage of the joint object detection and segmentation ability of Mask R-CNN to detect and segment vehicles using a LIDAR-synchronized camera. The semantic segmentation result regions indicate the occupancy of vehicles in the image plane. A sample result of Mask R-CNN vehicle detection and segmentation from an on-road camera image is shown in Fig. 3a. To associate the LIDAR points in 3D world coordinates with pixels in the image, we project the LIDAR points onto the corresponding image using calibration matrices. By applying an instance segmentation method and geometry projection of the LIDAR points in world coordinates, we can generate the semantic segmentation of each object in terms of LIDAR points. The result of Mask R-CNN and LIDAR semantic segmentation is shown in Fig 3b.

B. 2D Spatial Segmentation in LIDAR Bird's-Eye View

In sparse LIDAR data without sequence information, a K-D tree-based adaptive segmentation algorithm has been shown to be an efficient method for low-dimensional clustering [2]. In this method, 2D spatial information is considered to build up a K-D tree with LIDAR points as leaves. The range factor r is controlled by the distance between the LIDAR sensor and the object, as the angle resolution grows with the scanning distance. The algorithm purely segments all the data points from 2D coordinates into clusters which represent objects in the real world without additional heuristics on the number or shape of the objects.

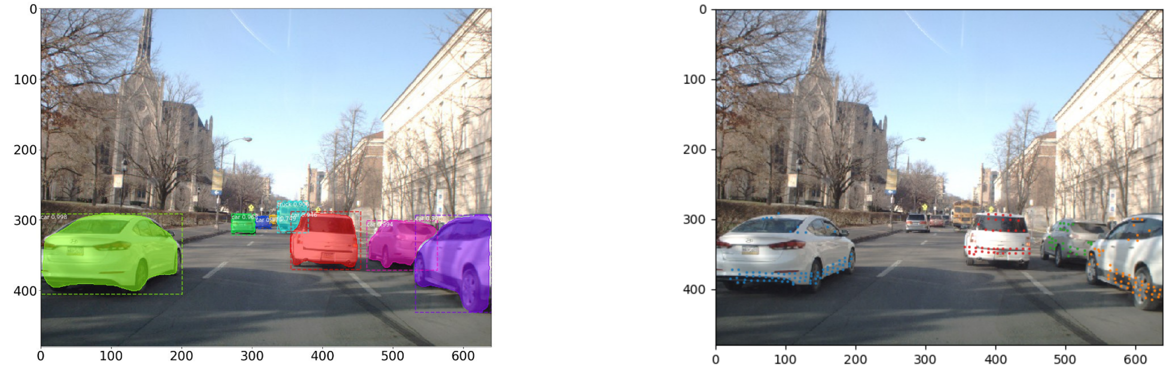
C. Fusion of LIDAR segmentation and Vision Semantic

In 2D spatial segmentation, the performance is limited by the over-segmentation caused by the occlusion of different objects as well as limited vertical FOV and quality of LIDAR data. One example is shown in Fig. 2b, where vehicle 1 has been over-segmented into two sub-clusters k_1 and k_2, caused by the occlusion of object $o1$. However, vision-based semantic segmentation and LIDAR point projection, as we described in III-A, suffer from ambiguity caused by inaccurate mask boundaries. As shown in Fig. 2c, vehicle 1 is under-segmented by the semantic information, which merges o_2 into the cluster. In this section, we introduce a novel optimization-based fusion method to solve these problems. To further improve the segmentation accuracy and estimate the vehicle heading, we propose a modified T-linkage RANSAC outlier removal algorithm as shown in 2e.

By applying 2D spatial segmentation in Section III-B, we obtain a list of N clusters defined as K, where a particular cluster in the list is k_n (spatial cluster). We define the clustering result of semantic-based LIDAR segmentation in Section III-A as S and a cluster in the list as s_m (semantic cluster), where $m \in \{1 \ldots, M\}$. We define an association matrix A, which is a square binary matrix indicating whether two spatial clusters belong to one object. We define β as the merging threshold of two spatial clusters. The optimization problem can be summarized as follows:

$$\begin{aligned} \text{Maximize} \quad & \sum_{i=1}^{N}\sum_{j=1}^{N} A_{ij} \\ \text{Subject to} \quad & A_{ij} = \mathbf{1}_{(k_i \cup k_j) \cap s_m \geq \beta} \; i,j = 1,\ldots,N \\ & A_{ij} \in \{0,1\} \end{aligned} \tag{1}$$

However, due to the inaccuracy of the mask and ambiguity on the mask boundary, some clusters of scanning points in the background or from nearby obstacles might be merged into the vehicle segments. A sample result is shown in Fig. 2c. This under-segmentation of vision semantics leads to an outlier, which misleads the vehicle dimension and heading estimation. To solve this problem, we introduce a modified T-linkage RANSAC algorithm with model-based iterative outlier rejection. The conventional T-linkage RANSAC algorithm [22] makes no assumption on the number of lines. Instead, it assumes that all the outliers are minority and distributed uniformly [23]. However, in planar LIDAR small clusters are sometimes reasonable inliers and relatively large clusters might be outliers caused by inaccurate vision under-segmentation. Therefore, we apply model-based iterative outlier rejection techniques to the the T-linkage algorithm. The proposed outlier rejection algorithm iteratively merges line segments C_i and C_j together to form a vehicle model based on a distance metric. The algorithm is shown in Alg. 1.



(a) Object detection and semantic segmentation result using Mask R-CNN. The bounding box, object category and corresponding mask are shown in the figure.

(b) Semantic segmentation result of LIDAR points for each vehicle. Points are selected within each segmentation region and clustered together.

Fig. 3: Sample results of Mask R-CNN and vision-based semantic segmentation of planar LIDAR points.

Algorithm 1: Model-based iterative outlier rejection algorithm

```
Input  : A list C of lines
Output: Object segmentation
while ($\hat{d} \leq \epsilon$) do
    Find $\hat{i}, \hat{j} : \hat{d} = \arg\min_{i,j} dist(C_i^k, C_j^{k^*})$;
    $C_{\hat{i}}$ = merge($C_{\hat{i}}$, $C_{\hat{j}}$);
    C.delete($C_{\hat{j}}$);
end
```

Algorithm 2: T-linkage RANSAC for vehicle heading estimation with model-based iterative outlier rejection

```
Input  : Fused cluster $V$
Output: Fitted Rectangle and vehicle Heading
SizeV = $V$.length();
H = HypoGenerator(hypoModel, $\beta \times SizeV$, $V$);
for ($h_i \in H$) do
    for ($v_j \in V$) do
        $R(i,j) = dist(h_i, v_j)$;
    end
end
$P$ = zeros($R$.size());
index = find($R \leq \epsilon$);
$P$(index) = Gaussian($R$(index), $\epsilon$);
$C$ = zeros(sizeS);
while ($\ni P(:,i) \not\perp P(:,j)$) do
    Find
     $p, q : d_T(p,q) = \arg\min_{i,j} dist_T(P(:,i), P(:,j))$;
    $C$ = Merge($C$, $p$, $q$);
    [$C$, $P$] = Update($P$, $V$, $C$);
end
$\hat{C}$ = outlierRejection($C$, $V$);
RectangleFitting($\hat{C}$, $V$);
```

We choose to use 'line' as the model hypothesis of the T-linkage algorithm, according to the appearance of the vehicle in the planar LIDAR scan. We can represent each point in the previous fused segment V by a preference vector based on a set of randomly generated line hypotheses H. Each element in the preference vector refers to the Euclidean distance of point v_i from line hypothesis h_j. We define P as the preference matrix with each column representing a point. The similarity of two points is calculated by the Tanimoto distance [24]. Similar points are clustered together into line segments C_i and the corresponding preference vector of the cluster is updated. We fit a rectangle model of each segment to estimate the dimension and heading of vehicle. One example of the outlier removal algorithm is shown in Fig. 2e. The overall T-linkage RANSAC algorithm with the model-based iterative outlier rejection algorithm is shown in Algorithm 2.

IV. EXPERIMENTAL RESULTS

In this section, we compare the performance of our proposed method with planar LIDAR-based vehicle segmentation method and vision-based method, using reduced-resolution KITTI LIDAR dataset [25] and our Cadillac SRX dataset.

A. KITTI Dataset Preprocessing

The Velodyne HDL-64E is a 64-channel, 360° FOV LIDAR sensor with adjustable data update rate. Unlike the planar LIDAR that are installed at the bumper height of the SRX, Velodyne LIDAR are usually mounted on top of the vehicle. The powerful Velodyne sensor has a horizontal angular resolution of around 0.09° and 26.8° of vertical FOV with approximate 0.4° angular resolution. To reduce the resolution of the dense KITTI Velodyne data, we select data points 0.8 m to 1.0 m below the center of the Velodyne sensor. In this way, the reduced-resolution Velodyne data have similar point cloud features to the planar LIDAR. We

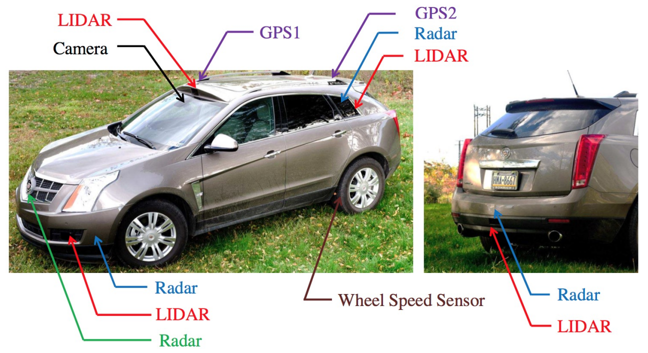


Fig. 4: Sensor integration of CMU Cadillac SRX. The sensors are FLEA camera, Radar and LUX LIDAR [5].

TABLE I: Performance comparison of 2D spatial segmentation and proposed Semantic Enhanced spatial segmentation over different categories of vehicles on pre-processed KITTI dataset. For precision and recall, we use a LIDAR point-wise metric.

Methods	Metric	Car	Truck	Van	Overall
2D Spatial	precision	0.998	0.930	0.995	0.995
	recall	0.927	0.720	0.926	0.918
	2D IOU	0.371	0.250	0.402	0.376
Semantic Enhanced	precision	0.993	0.879	0.998	0.987
	recall	0.975	0.852	0.969	0.968
	2D IOU	0.444	0.412	0.461	0.448

also use the RGB images to obtain semantic information.

B. Cadillac SRX autonomous vehicle and Dataset

CMU's autonomous vehicle testing platform is a Cadillac SRX equipped with six planar IBEO LUX sensors, which provide a 360° field of view (FOV) of the surrounding environment [5]. The LUX LIDAR sensor has a horizontal FOV of 85° with 0.125° angular resolution and vertical FOV of 3.2° with four horizontal planes separated by 0.8°vertically. To capture the front view of the vehicle, a FLEA camera is mounted under the rear-view mirror with a FOV of 36°, and is synchronized and calibrated with LIDAR. A detailed illustration of the sensor installation is shown in Fig. 4. We drove the SRX platform around the main CMU campus in Pittsburgh to collect urban traffic data.

C. Performance on Reduced-resolution KITTI LIDAR Dataset

We reduce the resolution of the KITTI LIDAR data to approximate the scenario in which a planar LIDAR is being used. We evaluate our proposed method on the reduced-resolution LIDAR bird's-eye view from the KITTI dataset and compare with previous methods. We compare the point-wise precision and recall of LIDAR point segmentation, as well as the 2D intersection over union (IOU) of the ground truth bird's-eye segmentation bounding box with the actual footprint of the vehicle. In Table I, the previous 2D spatial segmentation method achieves an average precision and average recall of 0.995 and 0.918. In this case, the previous method has a low performance in recall, which is caused by over-segmentation. By applying our semantic enhanced segmentation method, we achieve a 0.968 recall, which is much better than previous 2D spatial segmentation with a slight decrease in precision. The drop in precision is mainly caused by the corner cases when outliers are considered in the line model.

TABLE II: Average-Absolute-Heading Error (AAHE) of vision-based method, 2D spatial segmentation and proposed Semantic Enhanced segmentation over different categories of vehicles on pre-processed KITTI dataset. RGB images are used as input for Deep3DBox.

(°)	Method	Car	Truck	Van	Overall
AAHE	Deep3DBox [12]	7.72	10.79	9.08	8.26
	2D Spatial	6.67	4.16	4.53	5.92
	Semantic Enhanced	5.94	2.26	4.17	5.26

To compare the results with the 2D spatial segmentation method in detail, we also provide the precision and recall based on vehicle type. For the vehicle type truck, the 2D spatial segmentation method only achieves a recall of 0.720, since it is difficult for the planar LIDAR to capture the truck's whole shape. In all, we achieve a 5% increase for the car and van vehicle types and a 14% increase for the truck vehicle type. In Table II, we compare the AAHE of the 2D spatial segmentation method and the proposed semantic enhanced segmentation method. The proposed method improves the performance of overall AAHE by 11%. For the truck category, we even achieve a 45% increase. In compared with the single vision-based method Deep3DBox [12], our method achieves a 36.42% improvement in overall AAHE. Our method outperforms Deep3DBox in the Truck and Van categories by around 79% and 54% for heading estimation.

We also compared the 2D IOU of the 2D spatial segmentation algorithm and single vision-based method Deep3DBox with the proposed semantic enhanced LIDAR segmentation method at different distances. As the reduced-resolution LIDAR has a lot of occlusions at far distances, its 2D IOU performance decreases with distance. However, our semantic enhanced LIDAR segmentation method helps to achieve a better 2D IOU compared to 2D spatial segmentation methods at far distances. Compared to the Deep3DBox, the proposed method performs better at short distances which are more crucial than far regions, as we need more accurate dimension and location information of nearby obstacles for decision making and motion planning. The result is shown in Fig. 5. All these results indicate a better prediction in vehicle dimension and heading estimation by applying the proposed semantic enhanced method.

D. Performance on Cadillac SRX Dataset

We also compare our proposed semantic enhanced LIDAR segmentation method with a 2D spatial segmentation algorithm on the Cadillac SRX dataset. To evaluate the LIDAR point-wise segmentation accuracy, we compare the precision and recall by a point-wise metric. For our semantic enhanced method, we achieve a 3% increase in recall with slight decrease in precision. A better segmentation can boost the performance of vehicle heading and dimension estimation. Our semantic enhanced segmentation method achieved a

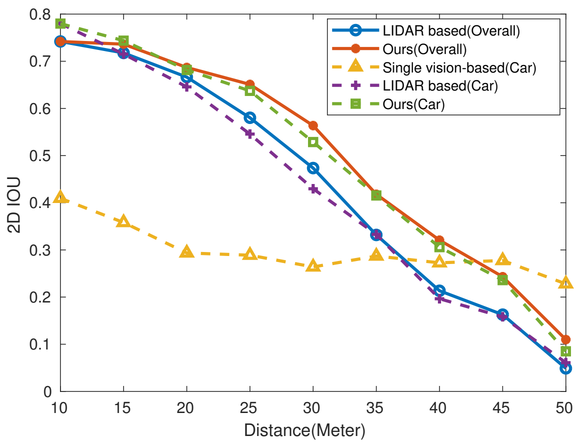


Fig. 5: Comparison of 2D IOU of LIDAR-based 2D spatial segmentation, vision-based Deep3DBox method and our proposed semantic enhanced LIDAR segmentation method with different distances from ego-vehicle to surrounding vehicles. We provide a comparison on all vehicles as well as the particular vehicle category 'Car'.

TABLE III: Performance comparison of 2D spatial segmentation and proposed Semantic Enhanced segmentation on Cadillac SRX dataset.

Methods(Overall)	Precision	Recall	2D IOU	AAHE (°)
2D Spatial	0.997	0.896	0.256	6.48
Semantic Enhanced	0.992	0.923	0.357	4.79

0.357 2D IOU, compared with 0.256 for the 2D spatial segmentation method. We also improved the AAHE of vehicle heading estimation from 6.48° to 4.79°. The results are shown in Table III. The comparison results show that our proposed semantic enhanced LIDAR segmentation algorithm performs better than the 2D spatial segmentation method on planar LIDAR.

V. Conclusion

In this paper, we propose a semantic enhanced LIDAR segmentation method with a modified T-linkage RANSAC outlier rejection algorithm. The proposed method solves the over-segmentation caused by the limitations of 2D spatial segmentation algorithm. The outlier rejection algorithm removes the outliers and ambiguities of the vision-based semantic mask. The proposed method is tested on a reduced-resolution KITTI LIDAR dataset as well as our own dataset. We achieved improvement in 2D IOU and average absolute heading error estimation (AAHE) on the reduced-resolution KITTI dataset as well as our Calldilac SRX dataset. Further work will be solving corner cases by improving outlier removal. We will also consider explicit occlusion reasoning.

References

[1] Z. Qiao, K. Muelling, J. M. Dolan, P. Palanisamy, and P. Mudalige, "Automatically generated curriculum based reinforcement learning for autonomous vehicles in urban environment," in *Intelligent Vehicles Symposium (IV), 2018 IEEE*. IEEE, 2018, pp. 1233–1238.

[2] X. Zhang, W. Xu, C. Dong, and J. M. Dolan, "Efficient l-shape fitting for vehicle detection using laser scanners," in *Intelligent Vehicles Symposium (IV), 2017 IEEE*. IEEE, 2017, pp. 54–59.

[3] C. Gao and J. R. Spletzer, "On-line calibration of multiple LIDARs on a mobile vehicle platform," in *Proceedings - IEEE International Conference on Robotics and Automation*, 2010, pp. 279–284.

[4] M. S. Darms, P. E. Rybski, C. Baker, and C. Urmson, "Obstacle detection and tracking for the urban challenge," *IEEE Transactions on Intelligent Transportation Systems*, vol. 10, no. 3, pp. 475–485, 2009.

[5] W. Junqing, S. Jarrod M., K. Junsung, D. John M., R. Raj, and L. Bakhtiar, "Towards a viable autonomous driving research platform," in *Intelligent Vehicles Symposium, 2013. Proceedings. IEEE*. IEEE, 2013, pp. 763–770.

[6] N. Fawzi and B. Alexandre, "Laser-based vehicles tracking and classification using occlusion reasoning and confidence estimation," in *Intelligent Vehicles Symposium, 2008. Proceedings. IEEE*. IEEE, 2008.

[7] D. R, O. E, A. J, and G. J, Villagra andC, "Multi-target detection and tracking with a laser scanner," in *Intelligent Vehicles Symposium, 2004. Proceedings. IEEE*, 2004, pp. 796–801.

[8] W. Stefan, S. Michael, K. Nico, and D. Klaus, "Classification of laserscanner measurements at intersection scenarios with automatic parameter optimization," in *Intelligent Vehicles Symposium, 2005. Proceedings. IEEE*, June, 2005, pp. 94–99.

[9] J. Levinson, J. Askeland, J. Becker, J. Dolson, D. Held, S. Kammel, J. Z. Kolter, D. Langer, O. Pink, V. Pratt *et al.*, "Lidar based perception solution for autonomous vehicles," in *Intelligent Systems Design and Applications (ISDA), 2011 11th International Conference on*, 2011.

[10] J. R. del Solar, L. Patricio, and S. Naiomi, "A survey on deep learning methods for robot vision," *CoRR*, vol. abs/1803.10862, 2018.

[11] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Proceedings of the International Conference on Computer Vision (ICCV)*, 2017.

[12] M. Arsalan, A. Dragomir, F. John, and K. Jana, "3d bounding box estimation using deep learning and geometry," in *CVPR*, 2017.

[13] X. Chen, H. Ma, J. Wan, B. Li, and T. Xia, "Multi-view 3d object detection network for autonomous driving," in *IEEE CVPR*, 2017.

[14] X. Chen, K. Kundu, Z. Zhang, H. Ma, S. Fidler, and R. Urtasun, "Monocular 3d object detection for autonomous driving," in *IEEE CVPR*, 2016.

[15] X. Chen, K. Kundu, Y. Zhu, A. Berneshawi, H. Ma, S. Fidler, and R. Urtasun, "3d object proposals for accurate object class detection," in *NIPS*, 2015.

[16] F. Chabot, M. Chaouch, J. Rabarisoa, C. Teulire, and T. Chateau, "Deep manta: A coarse-to-fine many-task network for joint 2d and 3d vehicle analysis from monocular image," in *CVPR*, 2017.

[17] D. L. D. L. Hall, S. Member, and J. Llinas, "An introduction to multisensor data fusion," *Proceedings of the IEEE*, vol. 85, no. 1, pp. 6–23, 1997.

[18] "Multisensor data fusion: A review of the state-of-the-art," pp. 28–44, 2013.

[19] H. Cho, Y. W. Seo, B. V. Kumar, and R. R. Rajkumar, "A multi-sensor fusion system for moving object detection and tracking in urban driving environments," in *Proceedings - IEEE International Conference on Robotics and Automation*, 2014, pp. 1836–1843.

[20] R. O. Chavez-Garcia and O. Aycard, "Multiple Sensor Fusion and Classification for Moving Object Detection and Tracking," *IEEE Transactions on Intelligent Transportation Systems*, vol. 17, no. 2, pp. 525–534, 2016.

[21] M. H. Daraei, A. Vu, and R. Manduchi, "Region Segmentation Using LiDAR and Camera," *2017 IEEE 20th International Conference on Intelligent Transportation (ITSC)*, pp. 1177–1182, 2017.

[22] M. Luca and F. Andrea, "T-linkage: A continuous relaxation of j-linkage for multi-model fitting," in *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 3954–3961.

[23] C. V. Stewart, "MINPRAN: A new robust estimator for computer vision," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 17, no. 10, pp. 925–938, 1995.

[24] T. T., "An elementary mathematical theory of classification and prediction," in *Internal IBM Technical Report*, 1957.

[25] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets robotics: The kitti dataset," *International Journal of Robotics Research (IJRR)*, 2013.