A Semantic-Guided LiDAR-Vision Fusion Approach for Moving Objects Segmentation and State Estimation

Songming Chen, Haixin Sun, Vincent Frémont

A Semantic-Guided LiDAR-Vision Fusion Approach for Moving Objects Segmentation and State Estimation

Songming Chen[1], Haixin Sun[1] and Vincent Frémont[1]

***Abstract*— Moving Objects Segmentation (MOS) is critical and indispensable for secure intelligent vehicle operation in the dynamic environment. For the state estimation task which is based on the assumption of static surroundings, to identify and filter out the moving objects plays an important role in robust ego-motion estimation. In this paper, a LiDAR-Vision fusion approach is developed to segment moving objects in the scene, which utilizes the LiDAR-based semantic segmentation as a prior and vision-based geometric information for validation. The effectiveness of our approach to segment moving objects is highlighted by the comparison with the traditional robust kernel-based outlier rejection methods. Our approach is benchmarked with three city category sequences in the KITTI dataset, which outperforms the kernel-based methods and achieves the leading results of 77.9% average fitness and 7.65 cm RMSE respectively.**

I. Introduction

Exteroceptive sensors are crucial for the intelligent transportation system to perceive the surroundings and realize self-state estimation in the dynamic and unknown environments. Among them, the Light Detection and Ranging (LiDAR) and the camera sensors are commonly employed to measure the changes in the environment, on which basis, the high-level tasks such as object detection, state estimation and obstacle avoidance could be performed. On the one hand, the visual sensor captures the rich representation of the scene, which includes the color, texture and semantic information. However, the frame-based cameras are very sensitive to the illumination changes and do not work well in the aggressive motion situation. Besides, the scale ambiguity remains another challenge for the vision-based perception and state estimation. On the other hand, the LiDAR sensors are illumination-invariant and provide all-round Field Of View (FOV) perception. They could also accurately acquire the depth information for scale-aware estimation. Nevertheless, the the performance of LiDAR sensors degrades a lot in the extreme weather conditions, such as undergoing dense fog and heavy rain. Another drawback of LiDAR-based perception is the sparsity of 3D point clouds, which brings the difficulty for feature extraction and data association. The complementary features of the visual and range sensors encourage us to adaptively combine them for robust perception and state estimation, which efficiently compensates the individual sensory modality weakness.

With the advent of deep learning, object detection neural networks could be applied to predict the object position and class on the image plane [1] and 3D point clouds [2] [3] in an end-to-end manner. These neural inference frameworks are mature and can achieve the real-time performance for object detection, which facilitates the on-board integration. Nonetheless, for the objects with the same class, their states of motion (static or dynamic) are not distinguished during the neural inference. Moving objects are considered as the most unstable traffic participants, which will corrupt the state estimation and mapping process. Thus, more attention should be given to them when we design an intelligent transportation system. In this paper, we focus on the problem of moving objects segmentation and investigate their impact on the state estimation. A LiDAR-Vision fusion approach is proposed to combine the LiDAR-based semantic cues and vision-based geometric clues to identify the objects position and state of motion jointly. These non-permanent moving objects are then filtered out beforehand for the efficient state estimation and scene reconstruction. The contributions of this paper are listed as:

1) An efficient moving objects segmentation approach with the complementary LiDAR and visual sensors is proposed, which adaptively fuses the semantic and geometric information.
2) The extensive evaluation (qualitative and quantitative) of our approach has been made, which is shown to outperform the conventional robust kernel-based methods for outliers rejection.
3) The robust semantic-guided state estimation and scene reconstruction has been achieved in the highly dynamic scenarios.

The rest of this paper has the outline as follows: In Section II, the related work and recent hybrid methods for moving objects segmentation will be covered. Then, the proposed methodology will be presented in Section III. After that, the experimental results and the corresponding analysis will be shown in Section IV. Finally, a brief conclusion and future perspectives will be given in Section V.

II. Related work

A. Vision-based method

The visual sensors provide dense texture and color information of the scene, which facilitates the geometric correspondence establishment and contextual understanding. In [4], a purely geometric mono-vision based approach is proposed for moving objects segmentation in challenging urban

* This work was carried out in the framework of the NExT Senior Talent Chair DeepCoSLAM, which was funded by the French Government (ANR-16-IDEX-0007). The first-author was also funded by China Scholarship Council.

[1] The authors are with Nantes Université, École Centrale de Nantes, LS2N, UMR 6004, 44000 Nantes, France. {songming.chen, haixin.sun, vincent.fremont}@ec-nantes.fr

areas. The epipolar, trifocal tensor, and structure consistency constraints are flexibly combined to classify the pixel-wise non-static points. The dynamic pixels are then clustered by the connected components labeler for instance-level moving objects segmentation. The optical flow consistency analysis also helps to segment the moving objects from the static background. In [5], the dynamic objects are identified with optical flow-based point trajectories clustering and these moving objects are then excluded from dense SLAM estimation in dynamic environments. In [6], the Flow Vector Bound (FVB) constraint is combined with graph-based clustering for incremental motion segmentation. Nonetheless, the aforementioned methods only leverage geometric information for the object clustering, which often fails in complex scenes. A novel Semantic-Guided RANSAC approach is thus presented in [7] for moving objects segmentation in heavy traffic scenarios. The semantic constraint provides potential moving objects prior and the geometric epipolar residuals are used for the final moving objects verification, which exhibits promising results. Despite of the efficient moving objects segmentation on the visual image plane, the scale metric remains ambiguous, which can be solved by the integration of stereo vision system [8] or range-based sensors [9].

B. LiDAR-based method

For LiDAR-based perception and state estimation, robust kernels are commonly adopted to ease the negative impact of outliers. The identified outliers could then be clustered to construct the moving objects in the scene. It is shown in [10] that, the outlier filters such as Tukey, Huber and Cauchy kernels could greatly mitigate the outliers effect for point clouds registration. Besides, the data-driven ResNet50-based method is proposed in [11] to infer the point-wise probability of being dynamic with only a single frame. On this basis, the scene reconstruction module takes the network output of dynamic objects probability for static components mapping. Then, the SpSequenceNet is designed in [12] to operate directly on 4D point clouds (consecutive 3D point clouds) for moving objects segmentation. Both the spatial and temporal information of LiDAR point clouds are exploited to extract the motion status. However, the SpSequenceNet training and prediction are computationally intensive due to the massive point clouds size. Recently, an innovative range image-based algorithm named Removert is presented in [13]. In the Removert framework, the dynamic objects are pruned from the query LiDAR scans via scan-to-map consistency check. Meanwhile, the pre-built map is corrected with the multi-scale false prediction reverting. As the prior map is not trivially accessible, the map-free method in [14] inputs the inter-frame range-image residuals to the semantic segmentation networks for the real-time class-agnostic moving objects segmentation. Nonetheless, the end-to-end network needs the ground truth binary masks for training that are quite time-consuming to prepare and refine.

C. Hybrid Methods

In order to overcome the individual sensor limitations, the hybrid methods which take advantage of the LiDAR and visual sensors are proposed in [9] [15] [16] [17]. The stereo vision systems are adopted in [9] to improve the object detection and tracking results of the LiDAR-based perception. Specifically, the vision-based system confirms the surrounding objects existence and their dynamic behavior are better modeled due to the dense visual measurements. In [15], the vision-based segmentation result is fused with the planar LiDAR-based prediction, which achieves an improving 2D Intersection-Over-Union (IOU) rate on the Bird-Eye-View (BEV) plane. Besides, the RGB images are converted to a polar-grid representation in [16], which augments the LiDAR point clouds with the color information for the efficient semantic segmentation. A novel architecture to fuse the precise LiDAR depth information and ERFNet-based visual semantics is presented in [17], which is shown to obtain satisfying objects segmentation results both on the image and BEV plane.

III. PRESENTATION OF THE METHOD

In this section, the proposed LiDAR-Vision fusion approach for real-time moving objects segmentation is detailed. And its overall pipeline is shown in Fig. 1. To start with, the LiDAR measurements are used to extract the 3D regions of interest (ROI) for movable objects prediction (see Section III-A). Then, with the given calibration parameters, 2D ROI on the image plane can be generated via the 3D-2D perspective projection. In order to determine the state of motion for the potentially moving objects, the temporal consistency check is conducted via the optical flow tracking and epipolar geometry (see Section III-B). Subsequently, the instance-level moving objects are back-projected and removed in the LiDAR point clouds for efficient state estimation and static scene mapping (see Section III-C).

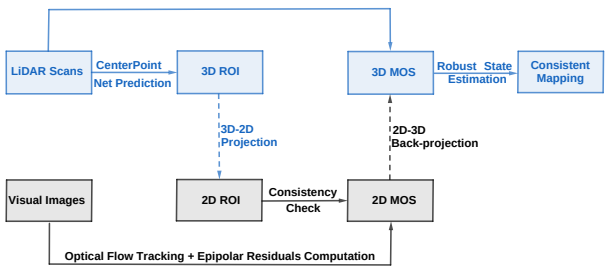

Fig. 1. Overview of the proposed LiDAR-Vision fusion approach for moving objects segmentation and state estimation

A. LiDAR Sensor-Based MOS Prediction

1) 3D MOS Prediction: In order to locate the movable objects and identify their semantic classes in the LiDAR point clouds, the center-based framework CenterPoint [3] is adopted for MOS prediction. The CenterPoint network follows the well-known encoder-decoder architecture pipeline, where the point clouds height and intensity information are

encoded in the Bird-Eye-View (BEV) map representation. A resnet-based backbone is used to extract features on the flattened BEV images, which is followed by the center heatmap head and property regression heads. Specifically, the center heatmap head helps to locate the object centers and infer their semantic classes. And the attributes of objects' 3D size and yaw orientation can be obtained from the up-scaled feature maps with the property regression heads.

The outputs of the CenterPoint network are expressed as $\{(C_x^i, C_y^i, C_z^i), (L^i, W^i, H^i), \theta^i, S^i\}_{i=1...n}$, which represent center coordinates, dimensions, heading angles and semantic classes (vehicles, cyclists or pedestrians) of n detected objects respectively. The center-based 3D object detection gives absolute range and captures real-scale shapes of the objects, which is perspective distortion-free and facilitates the interactive perception and state estimation. Compared with anchor-based 3D object detection methods, the center-based method is more robust to predict the heading orientation through the rotation-invariant points fitting, especially during the ego-turning phases[1] (see Fig. 2).

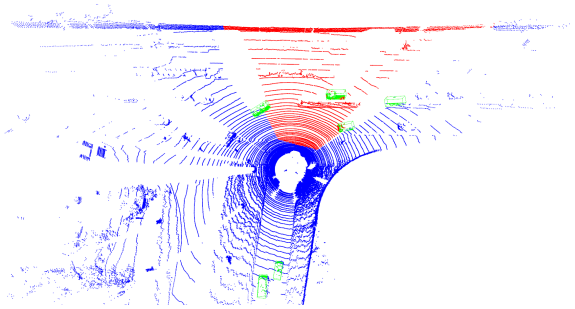

Fig. 2. The LiDAR-based 3D object detection results (rendered in green) with the CenterPoint network at the road intersection

2) 3D-2D Projection: In order to develop the LiDAR-Vision fusion approach, it is essential to transform the LiDAR-based detection results from the LiDAR frame $\mathscr{F}_L$ to the image frame $\mathscr{F}_I$ (see Fig. 3). We assume that the sensors are well synchronized and pre-calibrated with known extrinsic and intrinsic parameters. The perspective projection first takes eight corners ${}^{\mathscr{F}_L}\{\mathbf{x}_i\}_{i=1...8}$ of the 3D bounding box expressed in $\mathscr{F}_L$, and left-multiplies them with the LiDAR-Camera rigid transformation matrix ${}^{C}\mathbf{T}_L$ and image projection matrix ${}^{I}\mathbf{P}_C$ sequentially to get the corner coordinates in $\mathscr{F}_I$.

$$ {}^{\mathscr{F}_I}\mathbf{x}_i = {}^{I}\mathbf{P}_C * {}^{C}\mathbf{T}_L * {}^{\mathscr{F}_L}\mathbf{x}_i \tag{1} $$

Then the 2D bounding box boundaries can be extracted trivially from the span of the corner coordinates ${}^{\mathscr{F}_I}\{\mathbf{x}_i\}_{i=1...8}$. It is notable that the perspective projection constructs the one-to-one mapping correspondences for the 3D bounding box in $\mathscr{F}_L$ and 2D bounding box in $\mathscr{F}_I$, which provides the possibility for the bounding box back-projection. And during the 3D-2D perspective projection, the semantic labels of detected objects remain unchanged.

[1]The LiDAR scan and RGB-image are extracted from the KITTI 2011_09_26_drive_0014_sync sequence

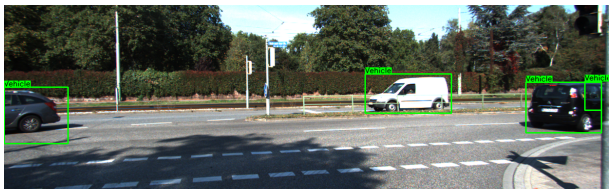

Fig. 3. The LiDAR-based 3D object detection results mapped on the image plane through the perspective projection

B. Visual Sensor-Based MOS Validation

1) 2D Point-Wise MOS: Given the predicted regions of interest with the LiDAR measurements, the visual multi-view geometry provides a sanity check for the moving objects segmentation validation. To start with, the Shi-Tomasi corners features, which remain invariant under the rotation, translation and scaling operation, are detected in the image frame. Then, the detected features are associated with the pyramidal Lucas-Kanade optical flow tracking between two consecutive image frames. And the optical flow backward check is also implemented to reduce the risks of mismatching. During the optical flow-based tracking, if the features lying on the movable objects (rendered in green, shown in Fig. 3) are not semantically consistent across two frames, they will be directly identified as dynamic points. After that, the matched features which belong to the background (rendered in blue, see Fig. 4), are used to estimate the fundamental matrix $\hat{\mathbf{F}}$ within the RANSAC framework. Since the movable objects points (rendered in green, see Fig. 4) are excluded from the estimation, the RANSAC process will converge quickly and provide a reliable fundamental matrix estimation. With the paired background points $(\mathbf{x}_i, \mathbf{x}_i')_{i=1...n}$ and the estimated fundamental matrix $\hat{\mathbf{F}}$, the corresponding epipolar line $\mathbf{l}_\mathbf{i}' \sim \hat{\mathbf{F}}\,\mathbf{x}_i$ with reduced coefficients $[a_i, b_i, c_i]^T$ can be obtained. And the Signed Epipolar Distance (SED) d_i^{SED} from the point $\mathbf{x}_i' = (u_i', v_i')$ to line $\mathbf{l}_\mathbf{i}'$ is calculated as:

$$ d_i^{SED} = \frac{a_i u_i' + b_i v_i' + c_i}{\sqrt{a_i^2 + b_i^2}} \tag{2} $$

Assuming that the measurement noise is normally distributed, then the calculated signed epipolar distance $\{d_i^{SED}\}_{i=1...n}$ will follow the Gaussian distribution as shown in Fig. 5, which lays the basis for outlier rejection. As the fundamental matrix transformation compensates the inter-frame ego-motion on the image plane, the static points will have close to zero (noise corruption) SEDs. On the contrary, the points on moving objects tend to get the SEDs exceeding the 3-sigma bounds, which will be classified as outliers and segmented from the static background. Nonetheless, it also needs to be mentioned that for objects following the degenerate motions within the epipolar plane, the epipolar constraint alone is not sufficient. This kind of degenerate motion usually happens when the ego-vehicle is following the moving object forward and constantly maintains the straight-line motion. In this case, the Flow Vector Bound (FVB) constraint [18] can be leveraged to detect such moving

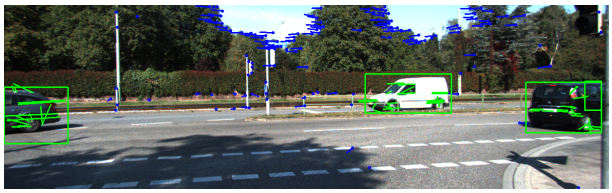

Fig. 4. The background corner points (rendered in blue) are tracked with sparse optical flow and are used for robust fundamental matrix estimation

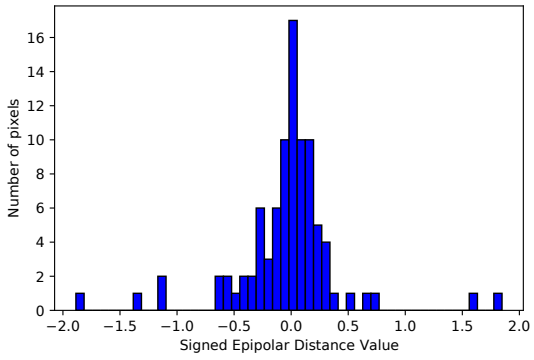

Fig. 5. The Signed Epipolar Distance distribution of background points (rendered in blue)

points with low SEDs. Given the sequential images, the pixels parallax d_i^{FVB} of paired points $(\mathbf{x}_i, \mathbf{x}_i')$ between two consecutive frames can be computed as:

$$\begin{aligned} \mathbf{x}_i' - \mathbf{KRK}^{-1}\mathbf{x}_i &= \frac{1}{z}\mathbf{Kt} \\ d_i^{FVB} &= \left|\mathbf{x}_i' - \mathbf{KRK}^{-1}\mathbf{x}_i\right| \end{aligned} \tag{3}$$

where the scalar z represents depth value and the matrices $\mathbf{K}$, $\mathbf{R}$ and $\mathbf{t}$ stand for the camera intrinsics, inter-frame rotation and translation respectively. With a set of matched background points $\{(\mathbf{x}_i, \mathbf{x}_i')_{i=1\ldots n}\}$, their parallax bound $[d_{min}^{FVB},\ d_{max}^{FVB}]$ can be easily found by Eq. 3. For the points with degenerate motions, if their parallax value is not within the interval of $[d_{min}^{FVB},\ d_{max}^{FVB}]$, they will also be labeled as dynamic outliers.

2) 2D-3D Instance-Level MOS: The vision-based MOS validation further exploits the underlying geometric and semantic cues, to identify the truly dynamic points lying on the movable objects. The combination of semantic, epipolar and FVB constraints allows for better MOS recognition even in degenerated cases. From the statistical point of view, the likelihood of an object being dynamic depends on proportion of outliers lying inside, on which basis, we can obtain the instance-level MOS probability. In our case, the threshold of 50% is set for the instance-level MOS decision making. It means that for each movable object, if there are more than 50% of the points inside are classified as mobile, the object itself will be considered as a dynamic object. Then, the object bounding box 2D-3D back-projection as depicted in Fig. 7, is implemented to get the depth information. And all the points inside the truly moving objects (rendered in red, shown in Fig. 6) will be cleared for the following robust state estimation and consistent scene mapping. In brief, 2D-3D instance-level MOS algorithm can be summarized as:

Algorithm 1 2D-3D Instance-Level MOS Algorithm

Input: Predicted Movable Objects ROI $\{\mathbf{R}_i\}_{i=1\ldots r}^{2D}$
Output: Validated MOS $\{\mathbf{S}_i\}_{i=1\ldots s}^{2D,\ 3D}$

1: Detect the Shi-Tomasi corners and track them with LK-optical flow between two consecutive image frames.
2: Estimate the fundamental matrix $\hat{\mathbf{F}}$ with the feature matches belonging to the background.
3: Compute the SED residual distribution $(\mu^{SED}, \sigma^{SED})$ using Eq. 2 and the flow vector bound $[d_{min}^{FVB},\ d_{max}^{FVB}]$ using Eq. 3 for the background points.
4: Check the motion status of points inside $\{\mathbf{R}_i\}_{i=1\ldots r}^{2D}$ based on the semantic, epipolar and FVB constraints.
5: Classify the movable object $\{\mathbf{R}_i\}^{2D}$ as validated dynamic object $\{\mathbf{S}_i\}^{2D}$ if the proportion of outliers in $\{\mathbf{R}_i\}^{2D}$ exceeds the threshold 50%.
6: Implement the $\{\mathbf{S}_i\}^{2D}$ back-projection to obtain $\{\mathbf{S}_i\}^{3D}$.
7: Exclude the points inside $\{\mathbf{S}_i\}^{3D}$ for the following robust state estimation and consistent scene mapping.

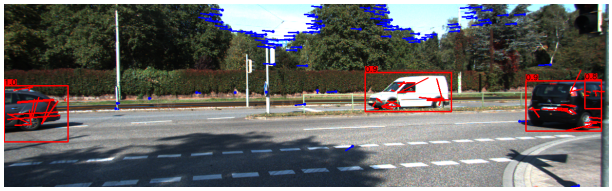

Fig. 6. The 2D instance-level moving objects segmentation (rendered in red), along with their probability of being dynamic

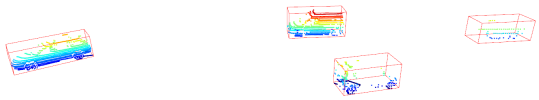

Fig. 7. The 3D instance-level moving objects segmentation via back-projection, and all the points inside will be classified as outliers

C. Robust State Estimation and Mapping

The robust state estimation is driven by the reliable correspondence matches across frames, where the estimated transformation ${}^j\mathbf{T}_i$ tends to minimize the overall distance between the paired correspondences $\{(\mathbf{p}_i, \mathbf{p}_j) \in \mathbf{M}\}$. In our pipeline, the 3D LiDAR scans are iteratively matched with the point-to-plane metric to deduce the vehicle ego-motion, which reaches centimeter-level precision. However, the existence of dynamic objects in the scene tends to cause the scan misalignment, thus degrading the registration accuracy of sequential LiDAR scans. In order to mitigate the impact of outliers in the objective function $\mathbf{F}({}^j\mathbf{T}_i)$ minimization, the

robust kernel functions $\rho(\mathbf{r}_{ij})$ adaptively adjust the weights of the matches with large residuals.

$$\mathbf{F}({}^{j}\mathbf{T}_i) = \sum_{(\mathbf{p}_i,\mathbf{p}_j)\in\mathbf{M}} \rho((\mathbf{p}_j - {}^{j}\mathbf{T}_i\mathbf{p}_i)\vec{\mathbf{n}}_j) \tag{4}$$

where $(\mathbf{p}_i, \mathbf{p}_j)$ are the correspondences belonging to the matched points set $\mathbf{M}$, and $\vec{\mathbf{n}}_j$ is the normal vector around $\mathbf{p}_j$ for calculating the point-to-plane distance. Nevertheless, the kernel-based Iterative Closest Point (ICP) method is not sufficient to handle the constant dynamic objects corruption. To solve this problem, our approach distinguishes the instance-level moving objects with the semantic and geometric information fusion. With the segmented moving objects back-projected to the 3D LiDAR scans, the weighting coefficients $\rho(\mathbf{r}_{ij})$ for the paired points lying inside the dynamic objects are uniformly assigned as zero, which further reduces the influence of dynamic objects in challenging scenarios. In order to align the LiDAR scans in the global frame, the poses from the state estimation thread will then be leveraged for static point clouds registration.

IV. EXPERIMENTAL RESULTS

In this section, the effectiveness of the proposed sensor fusion-based MOS and state estimation system is validated. The experimental evaluations are conducted with the challenging city category sequences[2] in the KITTI dataset [19], which were recorded in heavy traffic hours. The LiDAR-Camera sensor setup is adopted with the known calibration parameters, where the 64-layer Velodyne HDL-64E LiDAR gives accurate range information and the RGB-camera provides more contextual knowledge of the scene. In the KITTI dataset, only the front-view images are provided. So we focus only on the moving objects segmentation and static scene mapping within the visual sensor field of view. Extensive qualitative and quantitative results demonstrate that, the proposed semantic-guided MOS helps to robustify the pose estimation process in challenging heavy traffic scenarios. Moreover, the ghosting effect in scene reconstruction process is remarkably eliminated due to the moving objects removal.

A. Evaluation Metrics

The semantic-guided MOS efficiently reduces the outliers effect in the 3D LiDAR scan matching. In order to quantify the performance of LiDAR scan matching, the evaluation metrics of Relative Fitness (RF) and Relative Root Mean Square Error (RMSE) of the inlier correspondences $\{\mathbf{I}\}$ are used. They are defined as follow:

$$RF = \frac{|\mathbf{I}|}{|\mathbf{M}|}, \quad RMSE = \frac{1}{|\mathbf{I}|}\sum_{(\mathbf{p}_i,\mathbf{p}_j)\in\mathbf{I}} \sqrt{\|\mathbf{p}_i - \mathbf{p}_j\|^2} \tag{5}$$

where $|\cdot|$ is taking the cardinal number of a set. On the one hand, Relative Fitness (RF) measures the proportion of associated inliers $\{\mathbf{I}\}$ among all the matched point pairs $\{\mathbf{M}\}$, and higher relative fitness represents better scan matching results. On the other hand, RMSE measures the root mean square errors of all inlier correspondences $\{(\mathbf{p}_i, \mathbf{p}_j) \in \mathbf{I}\}$, and lower RMSE stands for greater scan alignment.

[2]http://www.cvlibs.net/datasets/kitti/raw_data.php?type=city

B. Results Analysis

1) Qualitative results: The ability to segment dynamic components in surrounding environments is essential for the intelligent transportation system. In our approach, the movable objects ROI are predicted with the CenterPoint neural network. Then, the visual multi-view geometry constraints provide a sanity check for instance-level MOS validation. Such a combination allows for better recognition, which is capable of detecting tiny objects (see Fig. 8) and even partially occluded objects (see Fig. 6). The right-side vehicle in Fig. 6 is occluded on the image plane, which is quite difficult to detect with only visual hints. Nonetheless, the high-resolution LiDAR sensor receives the reflection from part of occluded vehicle and manages to predict its existence, as depicted in Fig. 7. Moreover, it is demonstrated in Fig. 8 that, the epipolar constraint compensates the vehicle ego-motion which accurately classifies the parked car as static. And the flow vector bound constraint efficiently helps to identify the dynamic vehicle performing degenerate motions on the lane, which facilitates better contextual understanding of the driving area[3].

Fig. 8. The tiny moving object with degenerate motions (move along the epipolar plane) is successfully segmented with the FVB constraint

Fig. 9. The static car parked on the roadside (in blue) and dynamic car driving on the lane (in red) are distinguished and back-projected to the 3D LiDAR scan

Since the moving objects are not temporally consistent, they do not belong to the permanent components of the scene. Therefore, moving objects should be eliminated from the mapping process in order to build a consistent representation of the scene, as shown in Fig. 10. The reconstructed static map will promote high-level tasks such as map-based localization and path planning.

2) Quantitative results: It is shown in Tab I and Tab II that, our semantic-guided MOS approach achieves the leading results of 77.9% average fitness and 7.65 cm RMSE respectively. Since the optimization-based state estimation is usually built upon the static environment assumption. The

[3]The LiDAR scan and RGB-image are extracted from the KITTI 2011_09_26_drive_0013_sync sequence

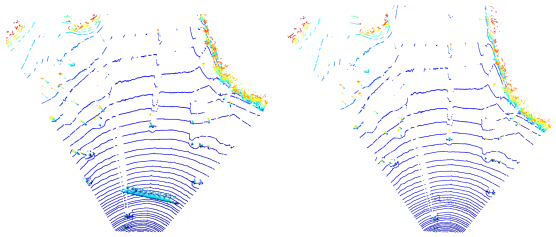

Fig. 10. The ghosting effect of a moving car is greatly reduced due to the semantic-guided moving objects segmentation and removal

presence of dynamic objects in the scene may degrade the ego-motion estimation and complicate the map maintenance task. A two-stage prediction-then-validation pipeline is thus designed to segment instance-level objects in the scene. It is more efficient than the traditional kernel-based methods, since we concentrate on the ROIs instead of the whole points clouds. Compared to the end-to-end DL-methods, our approach relieves from motion segmentation ground truth annotation and training. The multiple constraints combination also ensures the robustness of dynamic outliers rejection in complex situations, such as handling objects with degenerated motions.

TABLE I

REGISTRATION RESULTS BENCHMARKING WITH RF(%)

Methods / Dataset	Tukey	Huber	Cauchy	Ours
2011_09_26_13	57.2%	55.6%	55.6%	**57.7%**
2011_09_26_17	94.7%	94.7%	94.7%	**95.7%**
2011_09_26_18	78.9%	79.5%	79.1%	**80.2%**
Average	76.9%	76.6%	76.5%	**77.9%**

TABLE II

REGISTRATION RESULTS BENCHMARKING WITH RMSE (CM)

Methods / Dataset	Tukey	Huber	Cauchy	Ours
2011_09_26_13	10.18	10.37	10.39	**9.98**
2011_09_26_17	5.73	5.75	5.75	**5.66**
2011_09_26_18	7.37	7.41	7.41	**7.31**
Average	7.76	7.84	7.85	**7.65**

V. CONCLUSION AND FUTURE WORK

In this paper, we propose a novel pipeline to perform the efficient MOS that robustifies state estimation and facilitates consistent scene mapping in dynamic environments. The complementary range and visual sensors are combined to efficiently detect the truly moving objects, even in degenerated cases. The qualitative and quantitative experimental results demonstrate that, the proposed approach can outperform the traditional kernel-based methods in the complex traffic scenes. The future work plan will be devoted to onboard multi-camera and LiDAR fusion, where the front, side and rear camera images will be stitched for panoramic perception. The ablation study of different 3D object detection neural networks will also be conducted, which ensures the MOS robustness in various scenarios such as high-speed motion and mutual occlusion.

REFERENCES

[1] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2016, pp. 779–788.

[2] M. Simon, K. Amende, A. Kraus, J. Honer, T. Samann, H. Kaulbersch, S. Milz, and H. Michael Gross, "Complexer-yolo: Real-time 3d object detection and tracking on semantic point clouds," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2019, pp. 0–0.

[3] T. Yin, X. Zhou, and P. Krahenbuhl, "Center-based 3d object detection and tracking," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 11 784–11 793.

[4] V. Frémont, S. A. R. Florez, and B. Wang, "Mono-vision based moving object detection in complex traffic scenes," in *Proceedings of the IEEE Intelligent Vehicles Symposium*, 2017, pp. 1078–1084.

[5] Y. Wang and S. Huang, "Towards dense moving object segmentation based robust dense rgb-d slam in dynamic scenarios," in *Proceedings of the IEEE International Conference on Control Automation Robotics and Vision*, 2014, pp. 1841–1846.

[6] R. K. Namdev, A. Kundu, K. M. Krishna, and C. Jawahar, "Motion segmentation of multiple objects from a freely moving monocular camera," in *Proceedings of the IEEE International Conference on Robotics and Automation*, 2012, pp. 4092–4099.

[7] S. Chen, H. Sun, and V. Frémont, "Mono-vision based moving object detection using semantic-guided ransac," in *Proceedings of the IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems*, 2021, pp. 1–6.

[8] B. Wang, S. A. Rodríguez Florez, and V. Frémont, "Multiple obstacle detection and tracking using stereo vision: Application and analysis," in *Proceedings of the IEEE International Conference on Control Automation Robotics and Vision*, 2014, pp. 1074–1079.

[9] S. A. Rodríguez F, V. Frémont, P. Bonnifait, and V. Cherfaoui, "Visual confirmation of mobile objects tracked by a multi-layer lidar," in *Proceedings of the IEEE International Conference on Intelligent Transportation Systems*, 2010, pp. 849–854.

[10] P. Babin, P. Giguere, and F. Pomerleau, "Analysis of robust functions for registration algorithms," in *Proceedings of the IEEE International Conference on Robotics and Automation*, 2019, pp. 1451–1457.

[11] P. Ruchti and W. Burgard, "Mapping with dynamic-object probabilities calculated from single 3d range scans," in *Proceedings of the IEEE International Conference on Robotics and Automation*, 2018, pp. 6331–6336.

[12] H. Shi, G. Lin, H. Wang, T.-Y. Hung, and Z. Wang, "Spsequencenet: Semantic segmentation network on 4d point clouds," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 4573–4582.

[13] G. Kim and A. Kim, "Remove, then revert: Static point cloud map construction using multiresolution range images," in *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2020, pp. 10 758–10 765.

[14] X. Chen, S. Li, B. Mersch, L. Wiesmann, J. Gall, J. Behley, and C. Stachniss, "Moving object segmentation in 3d lidar data: A learning-based approach exploiting sequential data," *IEEE Robotics and Automation Letters*, vol. 6, no. 4, pp. 6529–6536, 2021.

[15] C. Fu, P. Hu, C. Dong, C. Mertz, and J. M. Dolan, "Camera-based semantic enhanced vehicle segmentation for planar lidar," in *Proceedings of the IEEE International Conference on Intelligent Transportation Systems*, 2018, pp. 3805–3810.

[16] K. El Madawi, H. Rashed, A. El Sallab, O. Nasr, H. Kamel, and S. Yogamani, "Rgb and lidar fusion based 3d semantic segmentation for autonomous driving," in *Proceedings of the IEEE Intelligent Transportation Systems Conference*, 2019, pp. 7–12.

[17] R. Barea, C. Pérez, L. M. Bergasa, E. López-Guillén, E. Romera, E. Molinos, M. Ocaña, and J. López, "Vehicle detection and localization using 3d lidar point cloud and image semantic segmentation," in *Proceedings of the IEEE International Conference on Intelligent Transportation Systems*, 2018, pp. 3481–3486.

[18] A. Kundu, K. M. Krishna, and J. Sivaswamy, "Moving object detection by multi-view geometric techniques from a single camera mounted robot," in *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2009, pp. 4306–4312.

[19] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets robotics: The kitti dataset," *The International Journal of Robotics Research*, vol. 32, no. 11, pp. 1231–1237, 2013.