Semantic-Direct Visual Odometry

Yaoqi Bao, Zhe Yang, Yun Pan, *Member, IEEE*, and Ruohong Huan

***Abstract*—Traditional direct SLAM methods formulate the camera pose and map estimation as minimization of the photometric error, which is tackled by the Gauss-Newton algorithm or the Levenberg-Marquardt algorithm in the optimization. However, the convexity of the photometric error only holds in a small region due to the characteristics of the convexity for grayscale images. Thus, the system may be stuck in sub-optimal local minima when tracking points have large displacement. Unlike grayscale images, the semantic probabilities omit the details inside the semantic objects while mainly reforming on the boundary of semantic objects, which has better convexity for large displacement. In this letter, we propose a novel semantic-direct visual odometry (SDVO), exploiting the direct alignment of semantic probabilities. By constructing the joint error function based on grayscale images and semantic probabilities, the joint error function achieves better convexity contrary to the photometric error. Consequently, the proposed system moves towards optima with steady steps in the optimization iterations. Experimental results on the challenging real-world dataset demonstrate a significant improvement over the baseline by integrating the direct alignment of semantic probabilities.**

***Index Terms*—Autonomous vehicle navigation, semantic scene understanding, SLAM.**

I. Introduction

SIMULTANEOUS localization and mapping (SLAM) is a fundamental building block for plenty of artificial intelligence applications, such as autonomous driving, unmanned aerial vehicles (UAVs), virtual reality, and augmented reality. Driven by these appealing applications, SLAM has been a hot research topic in the last two decades among the Computer Vision and Robotics communities. Visual SLAM can be roughly divided into feature-based (indirect) methods [1]–[3] and direct methods [4]–[6]. Direct methods optimize the photometric error based on the direct alignment of grayscale images, while feature-based methods optimize the reprojection error based on the correspondences between repeatable features in consecutive frames. For a long time, the field was dominated by feature-based methods. In general, feature-based methods are more accurate, while direct methods are more robust in texture-less environments [5]. One of the inherent issues for direct methods is that the convexity of grayscale images only holds in a small region. Fig. 1 shows the 3D visualization of the grayscale value for one image in the KITTI odometry dataset [7] and the corresponding semantic probabilities generated by HR-Net [8]. The left column is the grayscale image (top) and the probabilities for the road class (bottom); the right column is the 3D visualization of the corresponding yellow box at the left column. As shown in the right column of Fig. 1, the grayscale image preserves the details of the object while the probabilities of road mainly reform on the boundary of the road. For points on the boundary of semantic objects, the convexity of the semantic probabilities holds in a larger region than the grayscale image.

In this letter, we propose a novel semantic-direct visual odometry (SDVO), which integrates the direct alignment of semantic probabilities into LDSO. By constructing the joint error function based on grayscale images and semantic probabilities, the joint error function achieves better convexity contrary to the photometric error. Consequently, the proposed system moves towards optima with steady steps in the optimization iterations. With the help of the direct alignment of semantic probabilities in multiple semantic channels, the discrimination of points is improved. Therefore, the proposed system establishes more robust data associations between consecutive frames, against the direct alignment of grayscale images alone. Most direct methods evaluate the weighted sum of squared difference (SSD) over a small neighborhood of points based on the strong assumption that the sensor values of object points are constant over time, which is not the case in reality. The direct alignment of semantic probabilities enhances the robustness to illumination due to the invariance of semantics for illumination variations.

The main contributions of this letter:

- According to our knowledge, SDVO is the first visual monocular SLAM system that exploits the direct alignment of semantic probabilities.
- By integrating the direct alignment of semantic probabilities into LDSO, our method achieves improved localization performance and outperforms ORB-SLAM2. Experimental results demonstrate the effectiveness of the proposed on the KITTI odometry dataset.

The remainder of this letter is organized as follows. Section II provides an overview of related works concerning semantic SLAM. Then, Section III presents SDVO in detail. The experimental results and analysis are given in Section IV. Finally, a brief conclusion is drawn in Section V.

Manuscript received January 26, 2022; accepted May 7, 2022. Date of publication May 23, 2022; date of current version May 31, 2022. This letter was recommended for publication by Associate Editor X. Zuo and Editor J. Civera upon evaluation of the reviewers' comments. This work was supported in part by the Zhejiang Provincial Natural Science Foundation of China under Grant LY19F020032, and in part by the Zhejiang Provincial Key Research and Development Program of China under Grant 2021C03027. *(Corresponding author: Yun Pan.)*

Yaoqi Bao, Zhe Yang, and Yun Pan are with the College of Information Science and Electronic Engineering, Zhejiang University, Hangzhou 310027, China (e-mail: baoyaoqi@zju.edu.cn; yangzhevlsi@zju.edu.cn; panyun@zju.edu.cn).

Ruohong Huan is with the College of Computer Science and Technology, Zhejiang University of Technology, Hangzhou 310023, China (e-mail: huanrh@zjut.edu.cn).

Digital Object Identifier 10.1109/LRA.2022.3176799

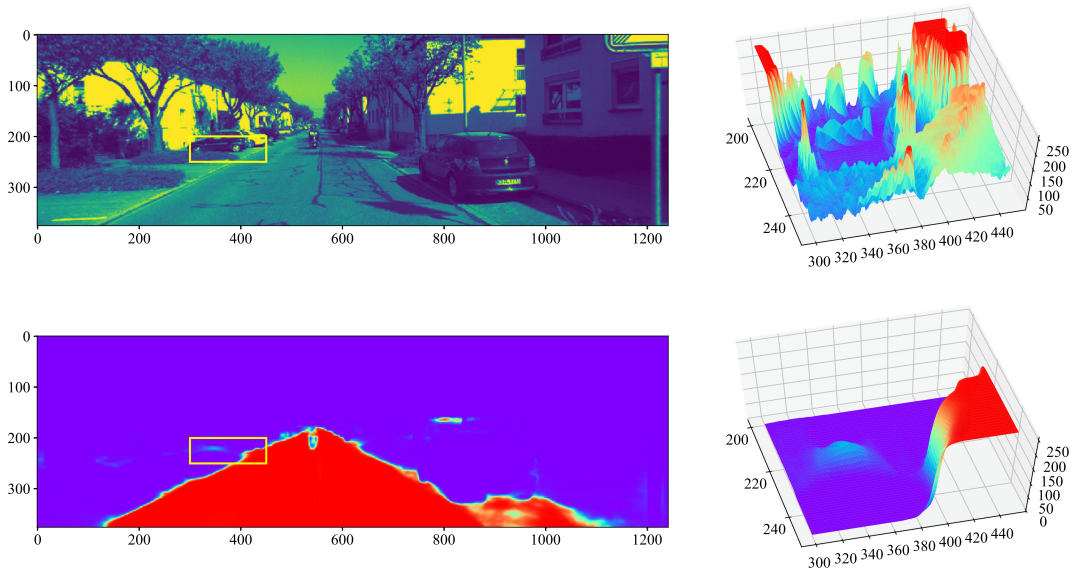


Fig. 1. The 3D visualization of the grayscale value for one image in the KITTI odometry dataset and the corresponding semantic probabilities generated by HRNet. The left column is the grayscale image (top) and the probabilities for the road class (bottom); the right column is the 3D visualization of the corresponding yellow box at the left column.

II. Related Works

In the last decade, there emerges a large number of works in the field of semantic SLAM or SfM: semantic visual odometry methods [9]–[17] that leverage semantic information to improve tracking accuracy or robustness; semantic mapping methods [18]–[24] that fuse high-level semantic information with geometric information to construct 3D semantic maps; semantic relocalization methods [25]–[31] that utilize the invariance of semantics to appearance variations to improve the precision of loop detection under challenging conditions; semantic alignment methods [17], [32]–[39] that exploit the alignment of semantics in camera pose estimates or map construction. For this letter, most related works are semantic visual odometry methods and semantic alignment methods.

A. Semantic Visual Odometry Methods

Recently, high-level semantic information like lines, planes, and objects are leveraged to improve the robustness in dynamic environments. Zhong *et al.* [9] employ the single shot multibox detector [40] to detect dynamic objects and discard all features on dynamic objects regardless of the state of features, which may suffer a loss in information. Recently, the epipolar geometry is widely used to distinguish moving objects and static objects. For robust self-localization in dynamic streets, Chen *et al.* [10] filter out moving dynamic objects with the help of YOLOv2 [41] and epipolar geometry. For robust self-localization in indoor environments, Yu *et al.* [11] combine the output of SegNet [42] with the epipolar geometry to reduce the impact of moving dynamic objects on tracking while building a dense semantic map. Similarly, Bescos *et al.* [12] adopt the output of Mask-RCNN [43] and multi-view geometry to filter out moving dynamic features. Another direction is to get initial estimates based on points of static class first. Brasch *et al.* [44] exploit semantic information extracted from static scene parts within an explicit probabilistic model to maximize the probability for both tracking and mapping. Cui *et al.* [13] utilize the output of SegNet to get a reliable fundamental matrix based on static features, which is further cooperated with epipolar geometry to filter out moving dynamic features.

Moreover, high-level semantic information is also leveraged to add extra constraints to improve tracking accuracy. Under the Manhattan world assumption, Kim *et al.* [15] estimate drift-free rotational motion jointly from both lines and planes by exploiting environmental regularities. Zhang *et al.* [16] adopt detected road traffic sign features in bundle adjustment to improve tracking accuracy. Our previous work [14] proposes a point reselection strategy based on coarse semantic plane constraints in urban environments, which discards static points inconsistent with the nearby co-plane points of the same semantic class. According to our knowledge, semantic probabilities have not been exploited for visual odometry yet.

B. Semantic Alignment Methods

Cohen *et al.* [32] employ symmetries and semantic information to stitch multiple sides of a building together. Taneja *et al.* [33] estimate initial poses of spherical panoramic images with respect to the cadastral 3D by aligning the building outlines. Toft *et al.* [34] estimate camera poses by aligning semantically segmented images to a pre-build sparse 3D model consisting of semantically labeled points and curves. In their later work [35], they propose a semantic consistency score that measures the individual correspondences by projecting the semantically labeled 3D points from the SfM model into semantically segmented images. These works are based on semantic image-to-model alignment, which is not suitable for visual odometry, as it requires a pre-built 3D model. For medium-term continuous

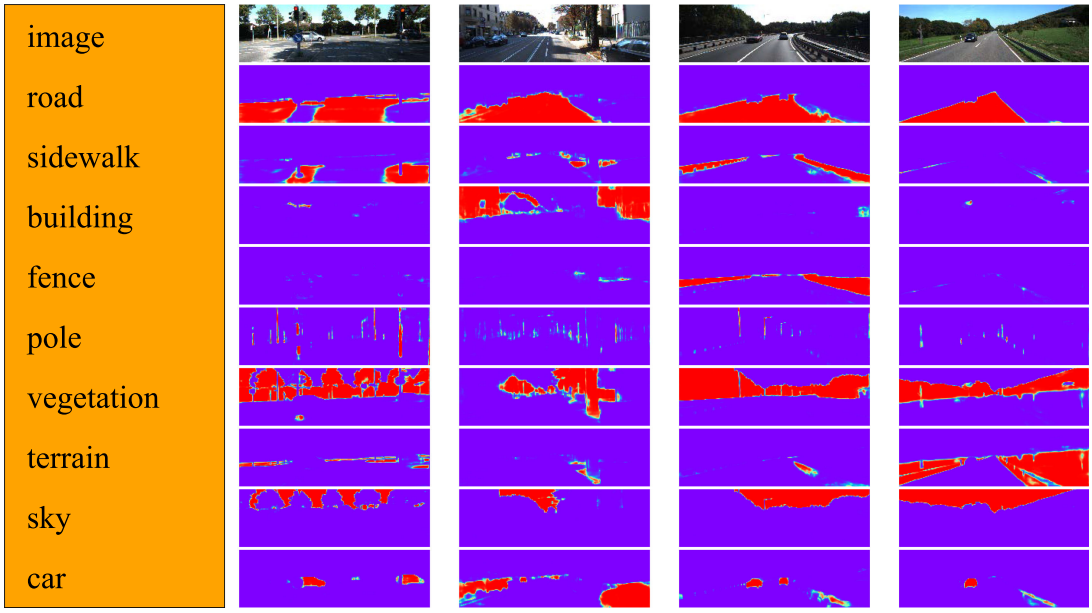


Fig. 2. The visualization of 9 selected channels for 4 representative scenes in the KITTI semantic segmentation benchmark. The top row is the original images, from the second row to the last row is the visualization of corresponding semantic channels.

tracking of points, VSO [17] proposes the semantic reprojection error, which measures points' distance to the nearest region with the same semantic label in the way of image-to-image alignment. The integration of the semantic reprojection error helps preserve camera-point associations for longer intervals, which results in the reduction of the translational drift. Apart from the alignment of semantics, the alignment of deep features has also attracted researchers' focus. To enhance the robustness of direct methods in challenging conditions, [36]–[39] employ the idea of algorithms for non-linear least squares problems to train deep features tailored for direct alignment.

In our perspective, the most closely related work is [17]. In contrast, our work is based on the direct alignment of semantic probabilities, not semantic labels. Semantic probabilities are continuous values that can be derivated directly, while semantic labels are discrete values that can not be derivated. Semantic labels need to be transformed to be integrated into the system error, like [17]. In this letter, semantic probabilities are directly aligned in the image-to-image style to estimate camera poses.

III. Methodology

The main idea of this letter is to leverage the better convexity of semantic probabilities in the optimization of direct methods. To do that, we integrate the direct alignment of semantic probabilities into LDSO. The main modifications are introduced in this section.

A. Semantic Probabilities and Point Selection

In this letter, we employ HRNet to extract semantic probabilities from images, due to its high-resolution characteristics. In machine learning, the softmax function [45] is widely used for assigning probabilities to multiple outputs. For semantic segmentation neural networks, the logits layer after softmax operation is regarded as probabilities of corresponding semantic class.

For each input image I_i, a dense pixel-wise semantic segmentation is generated by HRNet $C_i : \mathbb{R}^2 \rightarrow \mathcal{C}$, where each pixel is labeled as one of the semantic classes from set $\mathcal{C}$. For point p_k, the probability that p_k is of semantic class c can be calculated as:

$$p(s_{p_k} = c|C_i) = \frac{e^{x_c}}{\sum_{c \in \mathcal{C}} e^{x_c}} \tag{1}$$

where x_c is the value of p_k in the logits layer for the corresponding semantic channel c. Now, the sum of probabilities of all semantic classes has been normalized to 1.

The logits layer of HRNet has 19 channels corresponding to 19 semantic classes. However, not all semantic channels are abundant in information and tracking hints in the KITTI odometry dataset [7]. We only adopt the 9 semantic channels that frequently occur in the KITTI odometry dataset, including road, sidewalk, building, fence, pole, vegetation, terrain, sky, and car. Fig. 2 shows the visualization of 9 selected channels for 4 representative scenes in the KITTI semantic segmentation benchmark [46]. As shown in Fig. 2, the boundaries of semantic objects are well-captured, especially the boundaries of road, vegetation, sky, and car.

To maximize the strength of the direct alignment of semantic probabilities, the main goal of our point selection strategy is to select points with high semantic gradients. Normally, points on the boundary of semantic objects have both high gradients in semantic channels and the grayscale image. LDSO uses the Shi-Tomasi score [47] to detect corners in the grayscale image, while the proposed method computes the sum of the Shi-Tomasi score for 9 selected channels to detect semantic corners. Fig. 3

Fig. 3. The point selection in LDSO and SDVO for two scenes in the KITTI odometry dataset. The left column is the point selection in LDSO, while the right column is the point selection in SDVO.

shows the point selection in LDSO and SDVO. As shown in Fig. 3, SDVO mainly selects points alongside the boundaries of semantic objects, while LDSO selects points both inside objects and alongside the boundaries.

B. Model Formulation

Before presenting the semantic alignment error and the joint error of SDVO, we briefly introduce the photometric error. Let $\mathcal{X} = \{T_1, \ldots T_m, p_1, \ldots, p_n\}$ be the m SE(3) keyframe poses and n points (inverse depth parameterization) in the sliding window. The photometric error of point p_k detected in reference frame i, observed in a target frame j, is defined as:

$$e^I_{i,j,k} = \sum_{p \in \mathcal{N}_{p_k}} w_p \left\| (I_j[p'] - b_j) - \frac{t_j e^{a_j}}{t_i e^{a_i}} (I_i[p] - b_i) \right\|_\gamma \tag{2}$$

where $\mathcal{N}_{p_k}$ is the neighborhood pattern of p_k; a and b are the affine light transform parameters; t is the exposure time; I denotes the grayscale image; $\|.\|_\gamma$ is the Huber norm; w_p is a heuristic weighting factor; p' is the reprojected pixel of p on I_j calculated by:

$$p' = \Pi \left(R \Pi^{-1}(p, d_{p_k})) + T \right) \tag{3}$$

where Π and Π^{-1} are the projection function and the back-projection function; R and T are the relative rotation and translation between frame i and j; d_{p_k} is the inverse depth of p_k. More details can be found in [5], [6].

Similarly, the semantic alignment error of point p_k in semantic channel c, detected in reference frame i, observed in a target frame j, is defined as:

$$e^{S_c}_{i,j,k} = \sum_{p \in \mathcal{N}_k} w_{p_s} \left\| S_{c,j}[p'] - S_{c,i}[p] \right\|_\gamma \tag{4}$$

where $S_{c,i}$ is the semantic probabilities of semantic channel c for frame i and w_{p_s} is the heuristic weighting factor for the semantic channel c. In a word, the main difference between the photometric error and the semantic alignment error is the input. The photometric error measures the direct alignment of grayscale images, while the semantic alignment error measures the direct alignment of semantic probabilities for selected semantic channels.

Finally, the joint error E_{joint} to be minimized in the optimization:

$$E_{\text{joint}} = \sum_{T_i, Tj, p_k \in \mathcal{X}} \left(e^I_{i,j,k} + \lambda_s \sum_{c \in \mathcal{N}_c} e^{S_c}_{i,j,k} \right) \tag{5}$$

where $\mathcal{N}_c$ is the set of selected semantic channels; λ_s is the weight of the semantic alignment error for all selected semantic channels.

C. Sliding Window Optimization

Like [6], we optimize the joint error E_{joint} using the Levenburg-Marquardt algorithm. Let $x = [x_p^T, x_d^T]^T$ denote all the variables to be optimized, with x_p including camera intrinsics, affine brightness parameters, and camera poses, while x_d including the inverse depth of points. Normally, the windowed optimization problem:

$$H \delta_x = b \tag{6}$$

$$H = (1 + \lambda) J^T W J \quad \text{and} \quad b = -J^T W r \tag{7}$$

where J is the Jacobian of the residual r and W is the weighting matrix. For the proposed method, the residual r has 2 main parts, the residual of the grayscale image $r^I_{i,j,k}$ and the the residuals of the semantic probabilities in selected semantic channels $r^{S_c}_{i,j,k}$.

$$r^I_{i,j,k} = (I_j[p'] - b_j) - \frac{t_j e^{a_j}}{t_i e^{a_i}} (I_i[p] - b_i) \tag{8}$$

$$r^{S_c}_{i,j,k} = S_{c,j}[p'] - S_{c,i}[p] \tag{9}$$

The Jacobian of $r^I_{i,j,k}$ and $r^{S_c}_{i,j,k}$ are computed individually, with J_I denotes the Jacobian of $r^I_{i,j,k}$ and J_{S_c} denotes the Jacobian of $r^{S_c}_{i,j,k}$. Then (7) can be rewritten as:

$$H = (1 + \lambda) \left(J_I^T W_I J_I + \sum_{c=1}^{N_c} J_{S_c}^T W_{S_c} J_{S_c} \right) \tag{10}$$

$$b = -\left(J_I^T W_I r^I_{i,j,k} + \sum_{c=1}^{N_c} J_{S_c}^T W_{S_c} r^{S_c}_{i,j,k} \right) \tag{11}$$

IV. Experimental Results

We evaluate the proposed method on the KITTI benchmark suite. To the best of our knowledge, KITTI is the only real-world dataset that provides data for both odometry [7] and pixel-level semantic segmentation [46]. However, KITTI semantic segmentation benchmark only has 200 semantically annotated images, which is not sufficient for training complex models like HRNet. To get accurate semantic probabilities, we fine-tune HRNet on the KITTI semantic segmentation benchmark based on the model pre-trained on the Cityscapes dataset [48]. The main focus of this letter is to illustrate the benefits of the direct alignment of semantic probabilities for visual odometry. For sequences with re-visited places, the loop closure thread can also reduce errors of camera pose estimates, which is complementary to the direct alignment of semantic probabilities. The tracking performance with the loop closure thread de-activated is also valuable since the occurrence of re-visited places can not be controlled. To fully illustrate the effectiveness of the direct alignment of semantic probabilities, experimental results with loop closure thread de-activated are provided. Moreover, experimental results with loop closure thread activated are also provided to evaluate the proposed method from a full perspective.

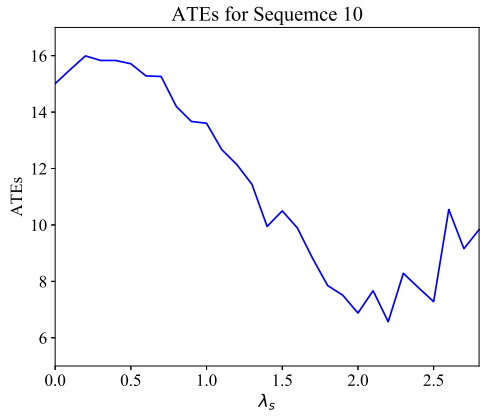


Fig. 4. ATEs under different λ_s for Sequence 10.

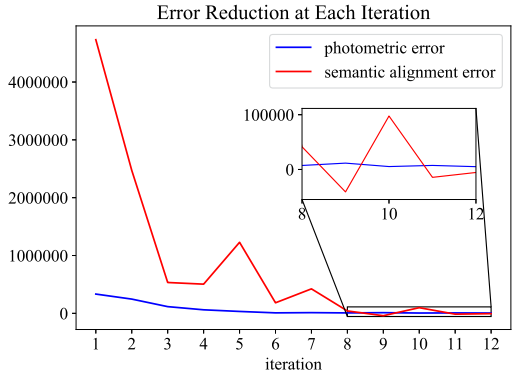


Fig. 5. An example of error reduction in the optimization ($\lambda_s = 2.2$).

A. *Configuration for λ_s*

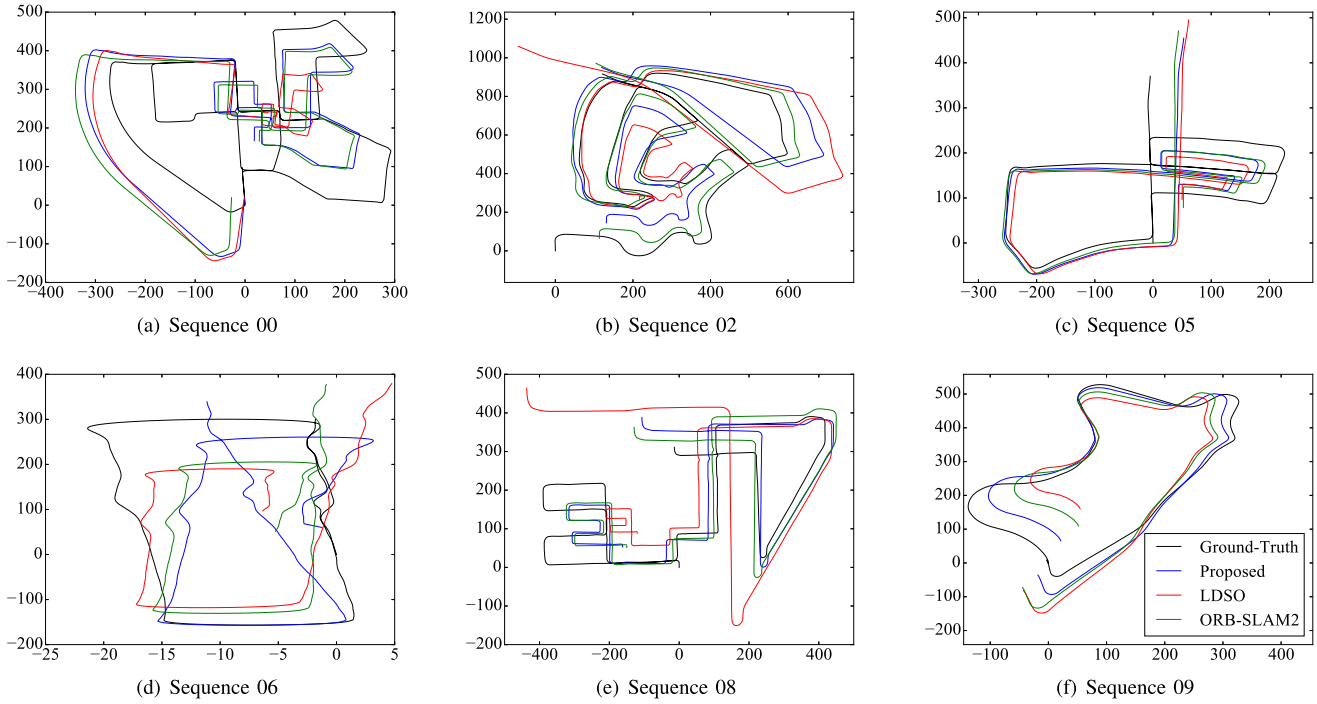
Before we compare the tracking performance of the proposed system with other methods, we need to find a suitable configuration for the weight of semantic alignment error for all selected semantic channels λ_s. We choose Sequence 10, which is a typical short sequence without re-visited places in the KITTI odometry dataset, for λ_s selection. To have a relatively comparable starting point, the semantic probabilities are rescaled to 255, which makes the value range similar to the grayscale image. Fig. 4 shows absolute trajectory errors (ATEs) under different λ_s for Sequence 10. The ATEs are computed by performing Sim(3) alignment to the ground truth on estimated trajectories. As shown in Fig. 4, the best configuration for λ_s is 2.2. The direct alignment of semantic probabilities plays a more important role as λ_s grows, and there is a decreasing trend in ATEs (before $\lambda_s = 2.2$) due to the better convexity that comes from the direct alignment of semantic probabilities. However, as λ_s goes beyond 2.2, the direct alignment of grayscale images may gradually sink in the direct alignment of semantic probabilities as the weight of the direct alignment of grayscale images becomes insufficient in the optimization. The consecutive frames in urban environments look similar in semantic probabilities since semantic probabilities omit details and mainly reform on the boundary of different semantic objects. Therefore, the direct alignment of the semantic probabilities alone may only provide rough estimates for the camera poses and the depths of points, even though the convexity of the semantic probabilities holds in a larger region than the grayscale image. On the contrary, the convexity of the grayscale image only holds in a small region, but the grayscale image is abundant in details of information, which is useful in the last iterations of the optimization for accurate estimates. Fig. 5 shows an example of error reduction in the optimization with $\lambda_s = 2.2$. As shown in Fig. 5, the reduction of the semantic alignment error is significantly higher than the reduction of the photometric error in the initial iterations. As the optimization goes on, the gap is gradually narrowing down. In the last iterations, the reduction of the semantic alignment error fluctuates around zero, while the reduction of the photometric error is stable (>0). Under the best configuration $\lambda_s = 2.2$, the optimization of the proposed method works in a coarse-to-fine style, which mainly relies on the direct alignment of semantic probabilities in the initial iterations but focuses on the direct alignment of grayscale images in the last iterations. The performance of HRNet varies from sequence to sequence, causing the best configuration for λ_s to differ from sequence to sequence. However, we still apply $\lambda_s = 2.2$ to all the sequences in the KITTI odometry dataset for a fair comparison.

TABLE I
COMPARISON WITH OTHER METHODS WITH LOOP COLSURE THREAD DE-ACTIVATED

Seq	ORB-SLAM2	VSO	LDSO	Proposed	Improvement to LDSO(%)
00	79.32	**45**	122.11	70.81	42.01
01	x	x	28.84	**21.23**	26.39
02	39.02	**23**	136.72	78.32	42.71
03	**1.11**	2.1	2.89	1.93	33.33
04	0.92	1.9	1.17	**0.39**	66.67
05	41.93	**19**	53.83	42.49	21.06
06	51.55	40.5	60.00	**20.46**	65.90
07	17.77	**12.5**	18.42	15.34	16.75
08	51.38	**42**	128.62	53.36	58.52
09	60.23	43	75.99	**30.67**	59.64
10	7.36	7.7	17.26	**6.85**	60.31

B. *Loop Thread De-Activated*

A comparison with LDSO, ORB-SLAM2, and VSO is given in Table I. Table I shows the ATEs of four methods with loop closure thread de-activated on all sequences of the training set in the KITTI odometry dataset. From the second column to the fifth column are the tracking result of ORB-SLAM2, VSO, LDSO, and the proposed method. For a fair comparison, we run each sequence 10 times to obtain the average ATE, except VSO. VSO proposes the semantic reprojection error based on the distance to the nearest region with the same semantic label, which can be integrated into traditional visual odometry to enable medium-term continuous tracking of points using semantics. We use the tracking result of integrating the semantic reprojection into mono-ORB-SLAM2 that is presented in [17]. The last column is the improvement of the proposed method compared to LDSO. The second column and the fourth column show that ORB-SLAM2 is more accurate than LDSO, while LDSO is more robust (ORB-SLAM2 fails on Sequence 01 due to the texture-less characteristic of the high-way scene). The



(a) Sequence 00 (b) Sequence 02 (c) Sequence 05

(d) Sequence 06 (e) Sequence 08 (f) Sequence 09

Fig. 6. Trajectories for ORB-SLAM2, LDSO, and the proposed method with loop closure thread de-activated.

fourth column and the fifth column show that the proposed method outperforms LDSO in each sequence. Moreover, the improvement of the proposed method compared to LDSO is quite obvious. For example, the improvement on Sequence 00, 02, 04, 06, 08, 09, and 10 are all over 42%. The proposed method achieves the best performance on Sequence 01, 04, 06, 09, and 10; VSO achieves the best performance on Sequence 00, 02, 05, 07, and 08; ORB-SLAM2 achieves the best performance on Sequence 03. According to [17], the weight of the semantic reprojection error is chosen empirically per sequence not consistent for all sequences. The semantic reprojection error in VSO requires points to have the same semantic label in consecutive frames. However, for points on the boundary of different objects, the semantic label is unstable. VSO takes the semantic label of points as parameters to optimize, using expectation maximization (EM). In comparison, the proposed method selects points with high semantic gradients (mainly on the boundary) and exploits the direct alignment of semantic probabilities. Meanwhile, the performance of the proposed method and ORB-SLAM2 is quite close on Sequence 05 and 08, the difference is less than 4%. Fig. 6 shows the trajectories estimated by ORB-SLAM2, LDSO, and the proposed method with loop closure thread de-activated for 6 sequences in the KITTI odometry dataset. The trajectories estimated by ORB-SLAM2 and the proposed method are both closer to the ground truth than trajectories estimated by LDSO in all sequences. The trajectories estimated by the proposed method are closer to the ground truth than ORB-SLAM2 in Sequence 00, 06, and 09, while the trajectories estimated by ORB-SLAM2 are closer to the ground truth in Sequence 02. The difference of estimated trajectories between the proposed method and ORB-SLAM2 is ambiguous in Sequence 05 and 08.

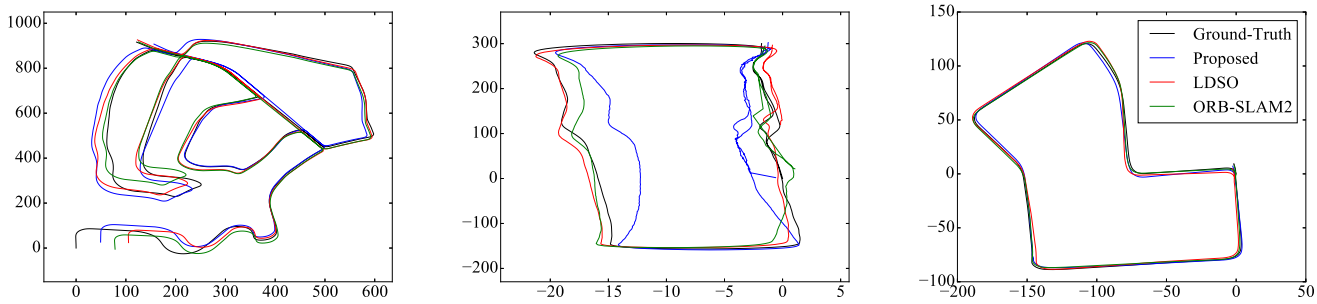
TABLE II
COMPARISON WITH OTHER METHODS WITH LOOP COLSURE THREAD ACTIVATED

Seq	ORB-SLAM2	LDSO	[14]	Proposed	Improvement to LDSO(%)
00	**7.02**	10.06	7.87	7.38	26.63
01	x	28.86	**6.70**	21.29	26.22
02	27.05	25.92	**22.83**	24.73	4.59
03	**1.14**	2.90	1.45	1.93	33.27
04	1.04	1.17	1.17	**0.39**	66.70
05	5.42	4.57	3.89	**3.66**	19.90
06	17.19	13.52	12.51	**5.40**	60.04
07	1.93	2.80	**1.69**	1.97	29.53
08	**52.06**	128.92	109.67	53.36	58.61
09	47.08	76.02	68.54	**30.73**	59.58
10	7.93	17.25	13.47	**6.84**	60.34

In summary, the integration of the direct alignment of semantic probabilities can surely improve the tracking accuracy of LDSO. With the loop closure thread de-activated, the proposed achieves better or comparable performance in most sequences of the KITTI odometry dataset (except Sequence 02) compared to ORB-SLAM2, while maintaining the robustness in texture-less environments.

C. Loop Thread Activated

Another comparison with LDSO, ORB-SLAM2, and our previous work [14] is given in Table II. Table II shows the ATEs of four methods with loop closure thread activated on all sequences of the training set in the KITTI odometry dataset (Sequence 00, 02, 05, 06, 07, and 09 have re-visited places, while Sequence 01, 03, 04, 08, and 10 do not have re-visited places). With loop



(a) Sequence 02 (b) Sequence 06 (c) Sequence 07

Fig. 7. Trajectories for ORB-SLAM2, LDSO, and the proposed method with loop closure thread activated.

closure thread activated or de-activated, the tracking result does not have much difference on sequences without re-visited places, like Sequence 01, 03, 04, 08, and 10, as the loop closure thread does not take effect on these sequences. For Sequence 09, the loop closure thread of LDSO and the proposed method does not work at all, while the loop closure thread of ORB-SLAM2 occasionally works. Therefore, the tracking results of LDSO and the proposed method for Sequence 09 in Table II are similar to Table I. From the second column to the fifth column are the tracking result of ORB-SLAM2, LDSO, the tracking result presented in [14], and the proposed method. The last column is the improvement of the proposed method compared to LDSO. The third column and the fifth column show that the proposed method outperforms LDSO in each sequence. In general, the improvement of the proposed method compared to LDSO on sequences with re-visited places is less than the improvement on sequences without re-visited places. The proposed method achieves the best performance on Sequence 04, 05, 06, 09, and 10. Meanwhile, the performance of the proposed method is close to the best performance on Sequence 00, 02, 07, and 08. Our previous work adopts the priori of flat planes in urban environments for point reselection, which limits the application scenarios. On the contrary, the proposed method does not have such limitations for application scenarios as long as accurate semantic probabilities are obtained. For long sequences without re-visited places (like Sequence 08 and 10), the improvement of our previous work compared to LDSO is not as obvious as the improvement of the proposed method. Fig. 7 shows the trajectories estimated by ORB-SLAM2, LDSO, and the proposed method with loop closure thread activated for 3 sequences with re-visited places in the KITTI odometry dataset. As shown in Fig. 7(a) and (c), the difference of the trajectories estimated by three methods is hard to tell. For Fig. 7(b), it looks like trajectories estimated by LDSO and ORB-SLAM2 are closer to the ground truth than the trajectory estimated by the proposed method. The reason for this illusion is the huge scale difference between the 2 axises. The trajectory estimated by the proposed method is actually closer to the ground truth than the trajectories estimated by the other 2 methods.

In summary, the improvement for the direct alignment of semantic probabilities decreases (compare Table I and Table II) as the complementary characteristic between the loop closure thread and the direct alignment of semantic probabilities. With loop closure thread activated, the proposed achieves better or comparable performance in all sequences of the KITTI odometry dataset compared to ORB-SLAM2, while maintaining the robustness in texture-less environments.

D. Runtime Results

We run all the experiments on an Ubuntu 16.04 LTS desktop with an AMD Ryzen Threadripper 1950X CPU and 64 G memory. Exclude the network inference time, taking Sequence 10 in the KITTI odometry dataset as an example, the proposed method needs 270.96 ms to process each frame on average, while LDSO needs 130.32 ms to process each frame on average.

V. Conclusion

In this letter, we proposed a novel semantic-direct visual odometry (SDVO), exploiting the direct alignment of semantic probabilities, inspired by the better convexity of semantic probabilities against grayscale images. Experimental results on the KITTI odometry dataset illustrate the effectiveness of the direct alignment of semantic probabilities. However, there are some points to address in the future. The weight of semantic alignment error λ_s is set to be consistent for all sequences, regardless of the semantic segmentation performance. An adaptive assignment of λ_s, which is aware of the semantic segmentation accuracy and depth variance, can further improve the tracking accuracy and robustness. Moreover, a multi-thread structure designed for multi-semantic channel processing, channel selection per point, and a lightweight network can be employed to improve the computational performance.

References

[1] A. J. Davison, I. D. Reid, N. D. Molton, and O. Stasse, "MonoSLAM: Real-time single camera SLAM," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 6, pp. 1052–1067, Jun. 2007.
[2] G. Klein and D. Murray, "Parallel tracking and mapping for small ar workspaces," in *Proc. 6th IEEE ACM Int. Symp. Mixed Augmented Reality*, 2007, pp. 225–234.
[3] R. Mur-Artal, J. M. M. Montiel, and J. D. Tardos, "ORB-SLAM: A versatile and accurate monocular SLAM system," *IEEE Trans. Robot.*, vol. 31, no. 5, pp. 1147–1163, Oct. 2015.
[4] J. Engel, T. Schöps, and D. Cremers, "LSD-SLAM: Large-scale direct monocular SLAM," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 834–849.

[5] J. Engel, V. Koltun, and D. Cremers, "Direct sparse odometry," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 3, pp. 611–625, Mar. 2018.

[6] X. Gao, R. Wang, N. Demmel, and D. Cremers, "LDSO: Direct sparse odometry with loop closure," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2018, pp. 2198–2204.

[7] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? The KITTI vision benchmark suite," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2012, pp. 3354–3361.

[8] K. Sun *et al.*, "High-resolution representations for labeling pixels and regions," 2019, *arXiv:1904.04514*.

[9] F. Zhong, S. Wang, Z. Zhang, and Y. Wang, "Detect-slam: Making object detection and slam mutually beneficial," in *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, 2018, pp. 1001–1010.

[10] W. Chen, M. Fang, Y.-H. Liu, and L. Li, "Monocular semantic SLAM in dynamic street scene based on multiple object tracking," in *Proc. IEEE Int. Conf. Cybern. Intell. Syst., IEEE Conf. Robot., Automat. Mechatronics*, 2017, pp. 599–604.

[11] C. Yu *et al.*, "DS-SLAM: A semantic visual SLAM towards dynamic environments," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2018, pp. 1168–1174.

[12] B. Bescos, J. M. Fácil, J. Civera, and J. Neira, "DynaSLAM: Tracking, mapping, and inpainting in dynamic scenes," *IEEE Robot. Automat. Lett.*, vol. 3, no. 4, pp. 4076–4083, Oct. 2018.

[13] L. Cui and C. Ma, "Sof-SLAM: A semantic visual SLAM for dynamic environments," *IEEE Access*, vol. 7, pp. 166 528–166 539, 2019.

[14] Y. Bao, Y. Pan, Z. Yang, and R. Huan, "Utilization of semantic planes: Improved localization and dense semantic map for monocular SLAM in urban environment," *IEEE Robot. Automat. Lett.*, vol. 6, no. 3, pp. 6108–6115, Jul. 2021.

[15] P. Kim, B. Coltin, and H. J. Kim, "Low-drift visual odometry in structured environments by decoupling rotational and translational motion," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2018, pp. 7247–7253.

[16] Y. Zhang, J. Yang, H. Zhang, and J.-N. Hwang, "Bundle adjustment for monocular visual odometry based on detected traffic sign features," in *Proc. IEEE Int. Conf. Image Process.*, 2019, pp. 4350–4354.

[17] K.-N. Lianos, J. L. Schonberger, M. Pollefeys, and T. Sattler, "VSO: Visual semantic odometry," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 234–250.

[18] A. Hermans, G. Floros, and B. Leibe, "Dense 3D semantic mapping of indoor scenes from RGB-D images," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2014, pp. 2631–2638.

[19] X. Li and R. Belaroussi, "Semi-dense 3D semantic mapping from monocular SLAM," 2016, *arXiv:1611.04144*.

[20] S. Yang, Y. Huang, and S. Scherer, "Semantic 3D occupancy mapping through efficient high order CRFs," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2017, pp. 590–597.

[21] J. McCormac, A. Handa, A. Davison, and S. Leutenegger, "Semanticfusion: Dense 3D semantic mapping with convolutional neural networks," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2017, pp. 4628–4635.

[22] V. Vineet *et al.*, "Incremental dense semantic stereo fusion for large-scale semantic scene reconstruction," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2015, pp. 75–82.

[23] Y. Nakajima, K. Tateno, F. Tombari, and H. Saito, "Fast and accurate semantic mapping through geometric-based incremental segmentation," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2018, pp. 385–392.

[24] S. Yang and S. Scherer, "Monocular object and plane SLAM in structured environments," *IEEE Robot. Automat. Lett.*, vol. 4, no. 4, pp. 3145–3152, Oct. 2019.

[25] Z. Seymour, K. Sikka, H.-P. Chiu, S. Samarasekera, and R. Kumar, "Semantically-aware attentive neural embeddings for long-term 2D visual localization," in *Proc. Brit. Mach. Vis. Conf.*, 2019, pp. 1–15.

[26] Z. Chen, F. Maffra, I. Sa, and M. Chli, "Only look once, mining distinctive landmarks from convnet for visual place recognition," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2017, pp. 9–16.

[27] T. Naseer, G. L. Oliveira, T. Brox, and W. Burgard, "Semantics-aware visual localization under challenging perceptual conditions," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2017, pp. 2614–2620.

[28] S. Garg, N. Suenderhauf, and M. Milford, "Lost? Appearance-invariant place recognition for opposite viewpoints using visual semantics," in *Proc. Robot. Sci. Syst. XIV.*, 2018, pp. 1–10.

[29] S. Garg, N. Suenderhauf, and M. Milford, "Semantic-geometric visual place recognition: A new perspective for reconciling opposing views," *Int. J. Robot. Res.*, 2019, doi: 10.1177/0278364919839761.

[30] Z. Hong, Y. Petillot, D. Lane, Y. Miao, and S. Wang, "Textplace: Visual place recognition and topological localization through reading scene texts," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 2861–2870.

[31] E. Stenborg, C. Toft, and L. Hammarstrand, "Long-term visual localization using semantically segmented images," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2018, pp. 6484–6490.

[32] A. Cohen, T. Sattler, and M. Pollefeys, "Merging the unmatchable: Stitching visually disconnected SFM models," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 2129–2137.

[33] A. Taneja, L. Ballan, and M. Pollefeys, "Registration of spherical panoramic images with cadastral 3D models," in *Proc. 2nd Int. Conf. 3D Imag., Model., Process., Visual. Transmiss.*, 2012, pp. 479–486.

[34] C. Toft, C. Olsson, and F. Kahl, "Long-term 3D localization and pose from semantic labellings," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops*, 2017, pp. 650–659.

[35] C. Toft *et al.*, "Semantic match consistency for long-term visual localization," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 383–399.

[36] B. Xu, A. J. Davison, and S. Leutenegger, "Deep probabilistic feature-metric tracking," *IEEE Robot. Automat. Lett.*, vol. 6, no. 1, pp. 223–230, Jan. 2021.

[37] L. Von Stumberg, P. Wenzel, Q. Khan, and D. Cremers, "GN-Net: The Gauss-Newton loss for multi-weather relocalization," *IEEE Robot. Automat. Lett.*, vol. 5, no. 2, pp. 890–897, Apr. 2020.

[38] L. Von Stumberg, P. Wenzel, N. Yang, and D. Cremers, "LM-RELOC: Levenberg-marquardt based direct visual relocalization," in *Proc. Int. Conf. 3D Vis.*, 2020, pp. 968–977.

[39] C. Tang and P. Tan, "BA-Net: Dense bundle adjustment network," in *Proc. Int. Conf. Learn. Represent.*, 2019.

[40] W. Liu *et al.*, "SSD: Single shot multibox detector," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 21–37.

[41] J. Redmon and A. Farhadi, "Yolo9000: Better, faster, stronger," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 7263–7271.

[42] V. Badrinarayanan, A. Kendall, and R. Cipolla, "SegNet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 12, pp. 2481–2495, Dec. 2017.

[43] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 2961–2969.

[44] N. Brasch, A. Bozic, J. Lallemand, and F. Tombari, "Semantic monocular SLAM for highly dynamic environments," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2018, pp. 393–400.

[45] J. S. Bridle, "Probabilistic interpretation of feedforward classification network outputs, with relationships to statistical pattern recognition," in *Neurocomputing*, Berlin, Germany: Springer, 1990, pp. 227–236.

[46] H. Abu Alhaija, S. K. Mustikovela, L. Mescheder, A. Geiger, and C. Rother, "Augmented reality meets computer vision: Efficient data generation for urban driving scenes," *Int. J. Comput. Vis.*, vol. 126, no. 9, pp. 961–972, 2018.

[47] J. Shi *et al.*, "Good features to track," in *Proc. Proc. IEEE Conf. Comput. Vis. pattern Recognit.*, 1994, pp. 593–600.

[48] M. Cordts *et al.*, "The cityscapes dataset for semantic urban scene understanding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 3213–3223.