2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
Madrid, Spain, October 1-5, 2018

Semantic Monocular SLAM for Highly Dynamic Environments

Nikolas Brasch[1*], Aljaz Bozic[2*], Joe Lallemand[1], Federico Tombari[2]

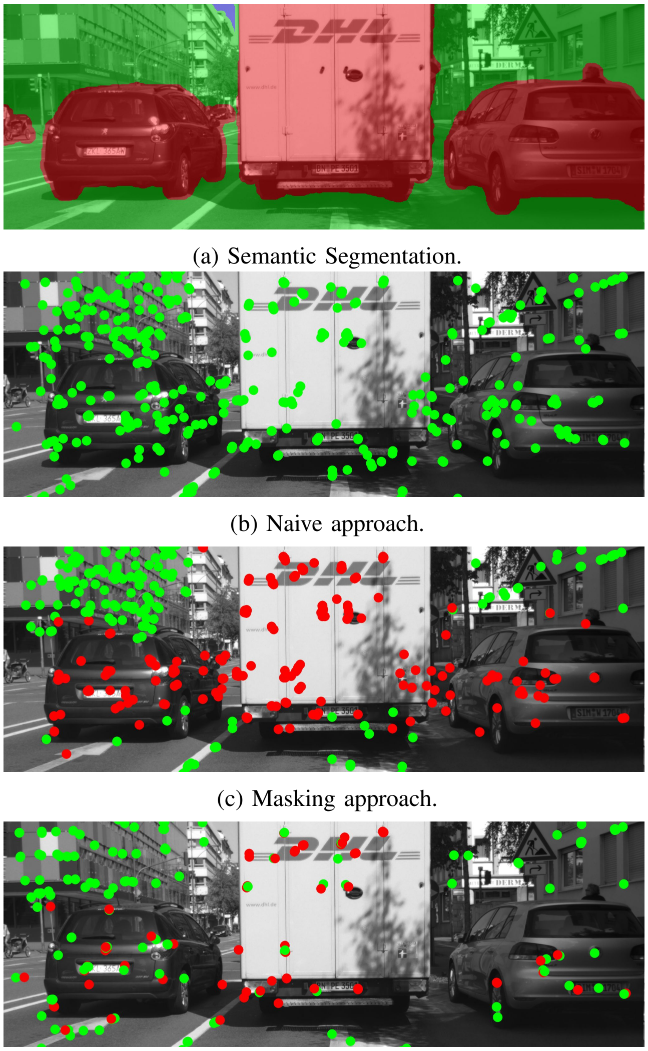
Abstract— **Recent advances in monocular SLAM have enabled real-time capable systems which run robustly under the assumption of a static environment, but fail in presence of dynamic scene changes and motion, since they lack an explicit dynamic outlier handling. We propose a semantic monocular SLAM framework designed to deal with highly dynamic environments, combining feature-based and direct approaches to achieve robustness under challenging conditions. The proposed approach exploits semantic information extracted from the scene within an explicit probabilistic model, which maximizes the probability for both tracking and mapping to rely on those scene parts that do not present a relative motion with respect to the camera. We show more stable pose estimation in dynamic environments and comparable performance to the state of the art on static sequences on the Virtual KITTI and Synthia datasets.**

I. INTRODUCTION

In the last years an intense research activity in the area of monocular Simultaneous Localization and Mapping (SLAM) allowed to achieve unseen accuracy, robustness and speed, enabling a variety of new applications in the areas of robotics and augmented reality. Compared to stereo- or RGB-D-based techniques, monocular SLAM algorithms [1], [2], [3], [4] rely on cheaper hardware, are simpler to calibrate and have no limitations in the depth range, making them particularly attractive for many mobile applications focused on both outdoor as well as indoor scenarios.

Monocular SLAM approaches can be divided into two groups. Descriptive methods [5], [1] use an explicit keypoint descriptor to find feature matches in different images and minimize the reprojection error between them. Differently, direct methods [6], [2], [4], [3] minimize the photometric error based on the projection of the pixel intensities from one image into the other. Both descriptive and direct approaches have their advantages and disadvantages, as analyzed in [3]. Specifically, descriptive methods are more robust against geometric noise, i.e. pixel position displacements, originating from incorrect camera intrinsic calibration or rolling shutter, while direct methods are better suited to cope with photometric noise, originating from motion blur.

Nevertheless, current monocular SLAM algorithms rely on the assumption that the surrounding environment is static, limiting their applicability to most real world scenarios. To deal with dynamic objects, they either employ M-estimators



(a) Semantic Segmentation.

(b) Naive approach.

(c) Masking approach.

(d) Our approach.

Fig. 1: An example from the CityScapes dataset of a difficult and highly dynamic scene, where a vehicle is standing in front of a traffic light. Most of the image area belongs to objects that are only temporarily static and will start moving slowly, thus causing the failure of approaches based on only motion clues for outlier detection (b). In (c) a semantic mask ignores all keypoints in potential dynamic areas and is thus unable to use parked vehicles for pose estimation. Our approach (d) uses point-wise outlier estimates using depth variance and fused semantic information (a). Red circles visualize an estimated outlier measurement, green circles are inliers and used for pose estimation.

[*] Authors contributed equally.

[1] Nikolas Brasch and Joe Lallemand are with BMW AG, 80788 Munich, Germany nikolas.brasch@bmw.de joe.lallemand@bmw.de

[2] Aljaz Bozic and Federico Tombari are with the Department of Computer Science, Technical University Munich, Boltzmannstr. 3, 85748 Garching, Germany aljaz.bozic@tum.de tombari@in.tum.de

978-1-5386-8093-3/18/$31.00 ©2018 IEEE

during optimization (Tukey [5], Huber [2], [1], [3]) or use RANSAC-based approaches to detect and filter out motion[7]. To properly work, both methods require the majority of points to be static with respect to the camera motion. Conversely, when dynamic objects cover a major part of the camera field of view, and especially when most visual features lie within these areas, current monocular SLAM approaches will fail. This is a particularly common condition for most outdoor driving-related scenarios: especially when dynamic objects move slowly or start moving from a standing still position (imagine the typical case of cars temporarily stopping at a traffic light, as shown in figure 1), the detection of outliers is extremely difficult. Without further knowledge about the observed scene, especially for monocular approaches, it is often not possible to differentiate between static and moving parts in the image. Due to recent advances in scene understanding and semantic segmentation based on Convolutional Neural Networks (CNNs), high level reasoning can be used to reduce the ambiguity between static and dynamic parts in the image. This is particularly interesting considering the development of new convolutional architectures and models capable of running efficiently and at a low memory footprint on mobile/embedded GPUs [8].

By relying on knowledge about the semantics of the scene, it is possible to detect potentially dynamic objects without the need to explicitly track them. Being able to segment the static parts of the scene such as buildings and lane markings, we can guide the feature extraction and matching towards these parts. Moreover, instead of relying only on frame-by-frame semantic information, we propose a probabilistic model, which takes into account semantic information of all frames to estimate the semantics of each map point. Beside the semantic information, we also use temporal motion information to argue whether a certain map point is dynamic or static. We update the probabilistic parameters of a map point when new observations are made. In order to enable a real-time SLAM system, we devised an efficient online probability update with a low constant memory consumption. In our evaluation we show more stable results in highly dynamic situations on synthetic and real datasets, while showing similar performance to state-of-the-art methods on static scenes.

II. RELATED WORK

Dynamic objects are treated as outliers by most SLAM algorithms. We propose to use semantic information to select a set of active features lying on static scene parts for a more robust pose estimation, in contrast to existing Semantic SLAM approaches, focusing on dense semantic 3D reconstruction. The semantic priors are generated by a deep model trained on RGB images.

A. Dynamic SLAM

In the past, different strategies have been proposed to handle dynamic outliers in visual SLAM. In [4] only active features, which have converged to a depth with a small variance after a certain number of observations, are used for tracking. Various modifications to [5] have been proposed to explicitly handle dynamic objects. In [7] an alternative RANSAC formulation is used, where the sampling is adjusted to distribute the sampled points. [9] use optical flow to find clusters in the flow orientation diagram of all features points and use the clusters to segment the dynamic objects from the static background. The use of RGB-D cameras or stereo cameras produces highly reliable and dense depth maps, in these cases freespace reasoning can be used to detect dynamic objects. Dynamic objects are detected if they move into areas which have been free before and labeled as outliers for the pose estimation [10]. When only sparse and noisy depth information is available, freespace reasoning is not possible. To handle dynamic scenes in monocular systems recent work focuses on multi-body structure-from-motion formulations. Here the scene is divided into multiple rigidly moving objects and the static world. Object instances are first detected via motion segmentation, then for each cluster a frame-to-frame transformation is computed and bundle adjustment is used to optimize the final trajectory [11]. Here the quality of the output depends on the motion segmentation. If the motion is small, the segmentation is poor and slowly moving objects will not be detected properly. The execution time is also far from real-time.

B. Semantic SLAM

Most of the existing approaches, combining classical S-LAM with a semantic segmentation of the scene, use the pose graph of the SLAM system to formulate a temporal or spatial consistent segmentation over an image sequence. The temporal or spatial consistency can be formulated as a CRF, over the images [12], a dense voxel grid [13] or a mesh [14]. The use of dense CRFs makes most of these approaches not suitable for large scale real-time applications in dynamic scenes, due to their low frame rate [12]. Other approaches use online updates for semantic fusion [15], which allows them to run in real-time. Most of the above approaches do not feed back the semantic information into the pose estimation pipeline. In [12] the semantic information is used to weight the measurements during the fusion in the 3D model. [16] remove points if the semantic class is different for multiple observations. To obtain dense 3D models stereo cameras are used in [12] and [17].

III. PROBABILISTIC SEMANTIC SLAM

The proposed SLAM system builds on top of the ORB-SLAM Framework [1], which consists of the three modules (1) Tracking, (2) Mapping and (3) Loop Closure. Figure 2 gives an overview of the framework. We propose an explicit model for dynamic and static map points and track the camera pose only on static points.

We start with the two first frames and initialize with ORB feature [1] based fundamental matrix estimation followed by global bundle adjustment, optimizing camera poses and map points jointly [1]. To compensate for pose estimation errors we use Lucas-Kanade optical flow [18] in combination with an epipolar constraint instead of a search along the epipolar

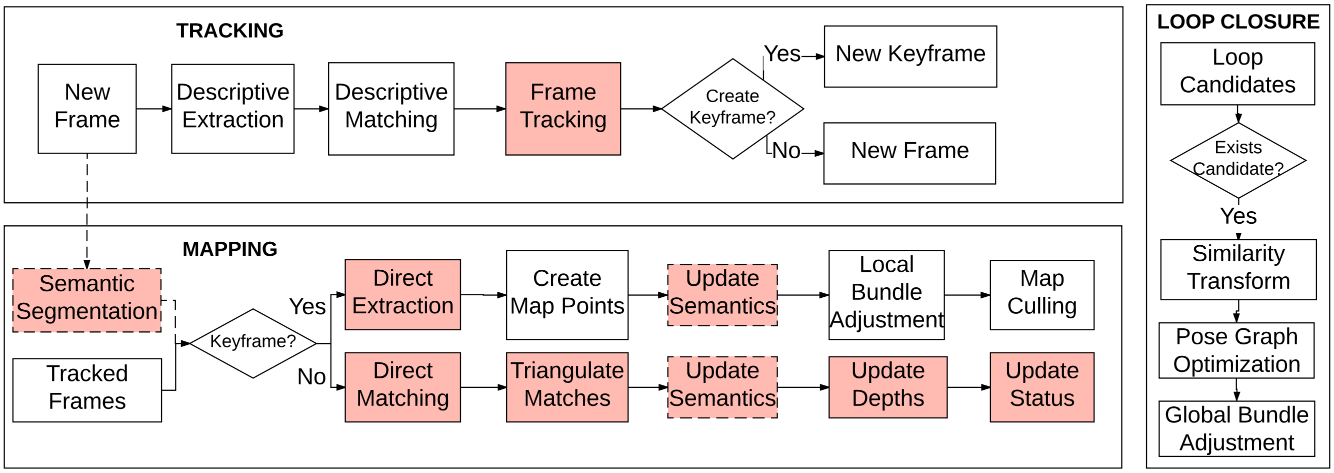


Fig. 2: Overview of our dynamic SLAM framework. The ORB-SLAM approach is extended with direct features extraction and matching in the mapping module, so they can be used in the tracking model for pose estimation. We also incorporate a probabilistic outlier model to update the status of each map point. Only active map points are used in the pose estimation. Added or modified modules to integrate the direct features and the probabilistic inlier model are shown in red.

line (as in [3]) to estimate an initial depth for the direct features.

For each new frame ORB features are extracted and correspondences are found via descriptor matching. The pose estimation is initialized based on a constant velocity motion model. We optimize the pose of the new frame based on both, descriptive and direct features, via multi-resolution, multi-step non-linear optimization. Depending on the number of descriptive and direct features per pyramid level we use multiple rounds, where we add or remove feature points based on their residual. We extract new direct features on keyframes only. If not enough correspondences with the current keyframe can be found, a new keyframe is created.

With the estimated pose of the new frame we can triangulate the descriptive and direct features to get a depth estimate. Map points and camera poses are refined jointly via local bundle adjustment on a fixed window of keyframes.

A. *Pose estimation and mapping*

Descriptive features like ORB are used in structure-from-motion approaches, providing fast and reliable matching, reducing the number of false-positive correspondences. Map points are in general parameterized as 3D points $X \in \mathbb{R}^3$ and the optimization is formulated as minimizing the reprojection error (see equation 2). Furthermore, descriptive features can be used to recognize loop closures and to relocalize in an existing map, which can be used for applications like localization in autonomous driving and tagging in augmented reality applications.

On the other hand, direct features avoid the computationally expensive extraction of keypoints and descriptors. On the other hand a good initial pose is required to allow the optimization to converge to the global minimum. This leads to the need for high frame rate or relatively slow camera motions. The image patch based matching accuracy is not as reliable as a descriptor comparison and false-positives or local minima might not be detected. Direct features are also sensitive to strong, abrupt or local illumination changes. Estimating an affine exposure model [3] jointly can only reduce these effects. Nevertheless direct features can be extracted also in low texture environments or if strong motion blur is present.

Due to these complementary properties we decided to use descriptive features whenever possible. In situations, where not enough features can be found, we use direct features in addition.

Instead of the 3D map point approach normally used for descriptive features, we decided to use an inverse depth formulation consistent with direct features. Hence, we can use the same probabilistic model for both kind of features, which simplifies the weighting in the joint optimization.

We decided to use ORB features as descriptive features due to their fast extraction (FAST-Corners + Orientation) and robust descriptors (BRIEF). We follow the implementation of [1] to get an equal distribution over the whole image using a grid extraction strategy and a multi-resolution pyramid. For the direct features we base our extraction on the work of [3] using a resolution pyramid and a grid based features selection technique.

During pose estimation and local bundle adjustment the weighted sum of reprojection error E_R and photometric error E_P is minimized.

$$E = \eta_R \sum_M E_R^T \Sigma_i^{-1} E_R + \eta_P \sum_N E_P^T \Sigma_i^{-1} E_P \qquad (1)$$

Here η_R, η_P balance the reprojection and photometric errors. Since we are reasoning on a per-point level, a dynamic adaptation of these two parameters is not needed.

We define T_n as a pose transformation in $SE(3)$ transforming a point $X \in \mathbb{R}^3$ expressed in world coordinate system to the coordinate system of frame n. Also, K is the camera intrinsic calibration matrix, π denotes the transformation from homogeneous to cartesian coordinates and d is the keypoint's estimated depth.

The reprojection error is given by the pixel distance of the observed keypoint $(x_j, y_j)^T$ and the projection of matched keypoint $\hat{x} = (x_i, y_i, 1)^T$ from frame i to frame j.

$$E_R = \begin{pmatrix} x_j \\ y_j \end{pmatrix} - \pi \left(K T_j \left(T_i^{-1} \left[d \left(K^{-1} \hat{x} \right] \right) \right) \right) \quad (2)$$

In the same way the photometric error is the pixel intensity difference between the patch around pixel (x_i, y_i) in image Φ_i and its projection into image Φ_j.

$$E_P = \Phi_i(x_i, y_i) - \Phi_j \left(\pi \left(K T_j \left(T_i^{-1} \left[d \left(K^{-1} \hat{x} \right) \right] \right) \right) \right) \quad (3)$$

We follow [3] and use an affine transformation model for the illumination change.

$$\Phi_i = \frac{1}{t_i} e^{\alpha_i} (I_i - \beta_i) \quad (4)$$

Where t_i is the shutter time, α_i and β_i are the affine transformation parameters that are estimated for each frame.

A weighted Gauss-Newton approach with robust Huber norm is applied to solve the non-linear least-squares problem. Similar to [2] we use covariance scaling, each term is weighted with its inverse covariance to reflect the uncertainty of each measurement.

The covariance propagation is performed after each new measurement by means of the update in equation 5.

$$\Sigma = \frac{\partial E^*}{\partial d} \sigma^2 \left(\frac{\partial E^*}{\partial d} \right)^T \quad (5)$$

B. Probabilistic outlier rejection

One core idea of SLAM is to refine the map of the 3D world with each new measurement by updating the 3D position of each map point after it has been observed again. Due to the fact that some measurements are more reliable than others, it is possible to do better than naively averaging by relying on a probabilistic approach that exploits the variance of the measurements as weight.

In a dynamic environment estimating only the positions of the map points is not enough. If we execute bundle adjustment using all points in the scene, including the dynamic ones, this will cause corrupted optimization estimates, since bundle adjustment assumes temporal consistency of point positions. Therefore, we also want to know which points are reliable enough for bundle adjustment.

We estimate an inlier ratio ϕ for each map point, describing how likely the map point is a reliable, stable point. The inlier ratio can be modelled in various ways, *e.g.* some approaches keep track of the number of successful and unsuccessful triangulations [1]. In [19], [4] a probabilistic model is used to model the depth jointly with the inlier ratio as a latent variable. In both cases the inlier ratio is updated by observing the map point's position through time, and deciding whether they are dynamic based on the estimated camera poses.

Motion estimation of map points to determine their inlier ratio can be ambigious in monocular SLAM. In the case of slowly moving objects or if a big dynamic object takes most of the camera view the object itself is assumed to be part of the static world. We include semantic information in the estimation of the inlier ratio to provide another independent source of information about how likely the map points are dynamic. Therefore, beside the depth d and inlier ratio ϕ, we also estimate a semantic class c for each map point.

When a map point is observed, we compute its current depth estimate d_i using triangulation, together with the estimated variance τ_i^2. The variance for the new measurement results from the triangulation, assuming that the keypoint's position in the image is only known with pixel accuracy [19]. We also estimate the matching accuracy $\alpha_i \in [0, 1]$, as described later, and retrieve the semantic class probabilities $CNN(c_k|I_i) \in [0, 1]$ for the keypoint from the neural network. Here $CNN(c_k|I_i)$ is the output of the network and can be understood as a probability that a keypoint belongs to the semantic class c_k, given current image frame I_i.

We define the depth measurement likelihood probability as in equation 6. It is based on [19], but we extend it with the use of matching accuracy. To simplify notation we use $\bar{x} = (1 - x)$.

$$p(d_i|d, \phi) := \alpha_i [\phi \mathcal{N}(d_i|d, \tau_i^2) + \bar{\phi} \mathcal{U}(d_i)] + \bar{\alpha}_i \mathcal{U}(d_i) \quad (6)$$

The intuition behind this definition is the following: if the current keypoint is correctly matched and the map point is static, then both the matching accuracy α_i and the inlier ratio ϕ will be close to 1. Therefore, the assumption is made, that the depth measurement d_i is distributed as a Gaussian ($\mathcal{N}(\mu, \sigma^2)$) around the mean μ, with variance τ_i^2. On the other hand, if the current matching is wrong, or the point is dynamic, then the current depth measurement is assumed to be uniformly distributed ($\mathcal{U}(a, b)$) and does not provide any useful information for the estimation of mean depth d.

Analogous to the case of depth, we model the semantics of a map point as a mixture of our network output $CNN(c_k|I_i)$ and a uniform distribution for the wrongly matched keypoints:

$$p(c_k|I_i) := \alpha_i CNN(c_k|I_i) + \bar{\alpha}_i \mathcal{U}(c_k) \quad (7)$$

This allows an efficient online update and a smooth transition between a dynamic and static state for map points.

Finally, we need to define the dependency of the inlier ratio on the semantic class. It turns out that we can derive efficient online parameter updates if we model the dependency as a Beta distribution, as given in equation 8.

$$p(\phi|c) = \prod_{k=1}^{K} \left(\frac{1}{B(A_k, B_k)} \phi^{A_k - 1} (1 - \phi)^{B_k - 1} \right)^{c_k} \quad (8)$$

Here the parameter c is a one-hot encoded semantic class and $A_k, B_k > 0$ are fixed constants, set for each semantic class. They represent the likelihood of a certain class being static or

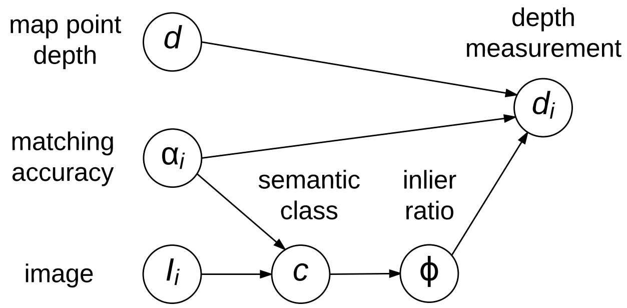


Fig. 3: Graph of the proposed joint probabilistic model, showing the relationship between depth measurement d_i, the matching accuracy α_i and the inlier ratio ϕ. The latter depends on the semantic class probability c predicted by a CNN from the current frame.

dynamic (e.g. a car class has a low A_k and high B_k, since it is more likely to be dynamic). The constants A_k, B_k can be scaled to put more or less weight on semantic measurements compared to depth measurements, *i.e.*, higher A_k and B_k would give more weight on the semantic prior rather than the motion prior for the estimation of the inlier ratio.

The dependency graph of the joint model for depth, inlier ratio and semantic class is given in figure 3. The measured depth d_i depends on the real depth d, the matching accuracy α_i and the inlier ratio ϕ, which depends on the semantics c.

Approximate inference leads to a posterior probability combining three terms. The first term includes the depth as a Gaussian distribution, the second term is the inlier ratio in the form of a Beta distribution based on depth measurements and the third is a Beta distribution modelling the dependency of the inlier ratio on the semantic class.

$$p(d, \phi | D, S) = \mathcal{N}(d|\mu, \sigma^2) Beta(\phi, a_{obs}, b_{obs}) \\ Beta(\phi, a_{sem}, b_{sem}) \quad (9)$$

Here $D = \{d_1, ..., d_N\}$ are all depth measurements, $S = \{s_1, ..., s_N\}$ are all observations with semantic information and $s_i = (CNN(c_1|I_i), ..., CNN(c_K|I_i))$ are the outputs for the K classes of the CNN as a probability density.

It can be shown that all depth measurements can be summarized with mean depth μ and depth variance σ^2. Similarly, the inlier ratio is distributed as a beta distribution with parameters $a_{obs} + a_{sem}$ and $b_{obs} + b_{sem}$. With some further algebraic manipulations, an efficient online update of these parameters can be derived, which enables fast updates of probabilistic models for map points.

For frames without semantic information the last term is not used. The parameters for the semantic beta distribution a_{sem} and b_{sem} are expressed by the following relationships:

$$a_{sem} = \sum_{k=1}^{K} A_k p(c_k|S) \quad (10)$$

$$b_{sem} = \sum_{k=1}^{K} B_k p(c_k|S) \quad (11)$$

The class posterior probability $p(c_k|S)$ is a probabilistic fusion of all semantic measurements, see equation 12. Compared to existing fusion approaches [13], the definition in equation 7 leads to a weighted semantic fusion, depending on the matching accuracy of each measurement α_i.

$$p(c_k|S) \propto \prod_{i=1}^{N} CNN(c_k|I_i)^{\alpha_i} \quad (12)$$

To estimate the matching accuracy of descriptive features, the Hamming distance is used to compare the binary descriptors.

$$\alpha^{descriptive} := 1 - \min\left(1, \frac{d(f_i, f_j)}{d_{max}}\right) \quad (13)$$

For the direct features we use the photometric difference between the two normalized image patches.

$$\alpha^{direct} := 1 - \min\left(1, \frac{\Delta\Phi(x_i, x_j)}{\Delta\Phi_{max}}\right) \quad (14)$$

In our implementation, we use the inverse depth [4], [20], to be able to model points at infinity. [20] have also shown that the inverse depth is more likely to be Gaussian distributed. It depends on the inlier ratio, whether we use the map point for pose estimation (active) or not (non-active). The current inlier ratio can be computed as in equation 15.

$$\phi = \frac{a_{obs} + a_{sem}}{a_{obs} + a_{sem} + b_{obs} + b_{sem}} \quad (15)$$

C. Real-time semantic segmentation

In highly dynamic scenes the image content can change rapidly. For a fast moving camera we need to extract new keypoints in every frame to keep enough active keypoints for reliable tracking. Therefore we extract keypoints in every new frame and not only in keyframes. To get a consistent semantic measurement for each new map point we run semantic segmentation on all new frames.

We use the 19-class pre-trained model of [8] for the CityScapes datase. We follow the proposed training procedure to finetune the model to the other datasets, adapting the final layer to the set of available classes. Training a model solely on static and dynamic classes instead of multi-class labels leads to slightly worse results, due to the lack of extra information about other classes, which are easy to recognize.

IV. EVALUATION

We split our evaluation two-fold. First, we evaluate on static scenes to compare the characteristics of descriptive vs. direct features and their combination.

Second, we evaluate on dynamic scenes to show the benefits of the semantic outlier model compared to the non-probabilistic model. The number of publicly available datasets with sequences showing highly dynamic scenes and providing ground truth pose is small. Therfore we make use of synthetic datasets providing highly accurate ground truth pose, semantic segmentation and depth maps enabling noise-free analysis of each component of the SLAM system.

Many off-the-shelf implementations of state-of-the-art approaches [3], [1], [4] rely on special initialization methods and handle tracking loss differently, comparing their results is challenging and might not reflect the actual effect of map point choice (descriptive or direct), their representation of the map as 3D points or inverse depth or the used optimization routine. Instead we extended the ORB-SLAM framework to work with direct features [3] and switch between our probabilistic formulation of depth [4] and a naive approach. We initialize with ground truth pose and scale the measurements of the first frame-to-frame pose to the correct ground truth scale. This way we can compare the accumulated drift without the need to scale and align the monocular trajectory afterwards, compensating for parts of the translation errors with the scaling.

We further run the evaluation in deterministic mode [21] alternating between tracking und mapping threads for each frame. This provides results independent of the used hardware and produces repeatable results for the evaluation of hyper parameters.

If ground truth trajectories are available we use the absolute trajectory error (ATE), the total accumulated translation error over a sequence, as well as the relative pose error (RPE), which includes the rotational error by computing the frame-to-frame pose error [22]. Furthermore, we show the ratio of points with correctly estimated depth (DR) as done in [15]. We set the threshold for correct depth to 10% of the ground truth depth and limit the evaluation to active points. All datasets come with different frame rates; for the deterministic setting we assume all frames can be used.

A. Static environment

The KITTI odometry dataset is used for evaluation in predominantly static environments [23], where ground truth 6D poses are available from a differential GPS and IMU system. We show the influence of the semantic probabilistic model in static scenes. Due to the fact that the semantic prior is not limited to moving objects but includes all objects of a potential dynamic class, features on these objects need longer to be included into the pose estimation. The KITTI dataset was recorded with 10 frames per second, which leads to relatively long distances between frames and a relatively small number of observations for each map point.

In table I the trajectory errors for different KITTI sequences are given, comparing the probabilistic model Ours with descriptive features only, to the non-probabilistic model ORB-SLAM*[1] and a naive masking approach Masked. It can be seen that the probabilistic model gives comparable results on static scenes, showing that the activation of map points, after their depth has converged, has no negative impact on the overall trajectory error. The naive masking approach results in an more unstable behaviour and tracking loss in unstructured environments or in the presence of many potential moving objects, leading to high errors for sequences 08 and 09. The majority of the KITTI Odometry sequences were captured in an urban area with buildings on both sides of the road, leading to a dense distribution of stable features. A part from the reoccurring tracking loss a semantic mask mainly removes unstable feature points on the reflective surface of parked vehicles, improving the pose estimation, if sufficient static structure is present.

Seq.	Length (m)	Configuration	ATE (m)	RPE (m)
00	713.17m	**Ours**	**37.63**	**49.38**
		ORB-SLAM*	45.28	58.18
		Masked	57.29	80.78
01	512.70m	Ours	30.05	39.98
		ORB-SLAM*	**20.01**	**23.19**
		Masked	43.75	60.57
02	5,065.70m	Ours	105.30	155.80
		ORB-SLAM*	101.06	**151.32**
		Masked	**99.29**	151.52
03	629.09m	Ours	33.00	39.85
		ORB-SLAM*	**26.59**	**30.89**
		Masked	28.29	35.98
04	552.43m	Ours	14.79	15.95
		ORB-SLAM*	14.08	14.59
		Masked	**7.51**	**9.12**
05	2,205.01m	Ours	31.88	41.12
		ORB-SLAM*	43.70	53.285
		Masked	**20.49**	**28.67**
06	2,119.37m	Ours	21.06	24.24
		ORB-SLAM*	**17.16**	**20.79**
		Masked	17.58	21.52
07	1,386.62m	**Ours**	**9.81**	**13.59**
		ORB-SLAM*	12.17	17.69
		Masked	13.32	16.37
08	3,221.97m	**Ours**	**50.92**	**76.06**
		ORB-SLAM*	87.06	114.77
		Masked	126.31	164.63
09	1,703.55m	**Ours**	**49.85**	**67.14**
		ORB-SLAM*	90.50	125.12
		Masked	113.18	157.87
10	919.08m	Ours	26.85	**33.19**
		ORB-SLAM*	32.14	40.57
		Masked	**26.30**	33.69
sum	19,028.69m	**Ours**	**411.14**	**556.30**
		ORB-SLAM*	489.75	650.40
		Masked	559.15	767.59

TABLE I: Evaluation on 11 KITTI sequences plus total ATE and RPE on all sequences.

Figure 5 shows the estimated trajectories together with the ground truth trajectories for sequence 07 in the KITTI dataset. The result of the probabilistic model on the top leads to similar drift as in the non-probabilistic case on the bottom.

To get a better understanding of the online update for the estimation of inverse depth and outlier ratio we show the evolution of the posterior probability given a new measurement for some example scenarios. For a map point on the static world, the semantic prior is set to increase the inlier ratio, to be able to use the potential static point as soon as possible. With a single triangulation of a map point it is not possible to determine, whether the map point belongs to a dynamic object or to the static world. Given semantic information and a high probability for a potential dynamic class, the map point is at first considered an outlier. Only when more observations become available it is possible to reason about the motion state of the map point.

If the map point belongs to a parked car, the inlier ratio

[1] We use a modified version of ORB-SLAM using an inverse depth formulation for map points instead of a full 3D point representation.

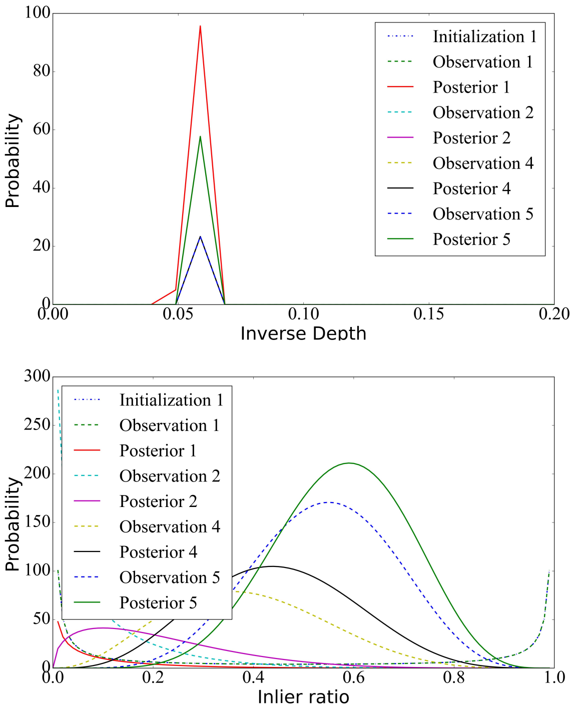


Fig. 4: Measurements and resulting posteriors for inverse depth and inlier ratio ϕ of a map point on a parked car in the KITTI 00 sequence. From an equal initialization the posterior converges to the correct inlier state.

increases after each successful triangulation. An example of this is shown in the bottom of figure 4. After the initialization of a new map point a and b take the same initial values leading to an inlier ratio of 0.5. The map point is equally likely inlier and outlier.

B. Dynamic environment

For the dynamic evaluation, we focus on showing the stability of each configuration for several shorter sequences, due to the fact that highly dynamic sequences are underrepresented in current datasets and that their impact on the metrics over a long sequence can be small. We use sequences from the CityScapes [24], Virtual KITTI [25] and Synthia [26] datasets showing different dynamic situations. The frame rates are 17 Hz for the CityScapes dataset, 10 Hz for the Virtual KITTI and around 5 Hz for the Synthia dataset. A lower frame rate leads to bigger motion of dynamic objects and the static world between frames.

Table II summarizes the trajectory errors for the challenging highway sequence 20 from the Virtual KITTI dataset, showing a traffic jam scene with slowly moving traffic. Here, ground truth segmentations have been used, to reduce the effect of the CNN output and a lack of training data for the dataset. Using the probabilistic model can reduce the trajectory error considerably. Where in general the extension with direct features can improve the pose estimation in static scenes, in some cases with few desriptive features and moving objects with hard shadow edges the pose estimation degrades. This is probably due to the fact that the synthetic scene of the highway does not give enough good ORB correspondences. Figure 6 shows the drift for the probabilistic model is much smaller than for the baseline model.

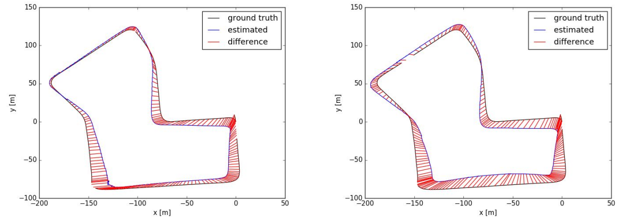


Fig. 5: Absolute trajectory error of `Ours` (left) and `ORB-SLAM`* (right) on the KITTI 07 sequence.

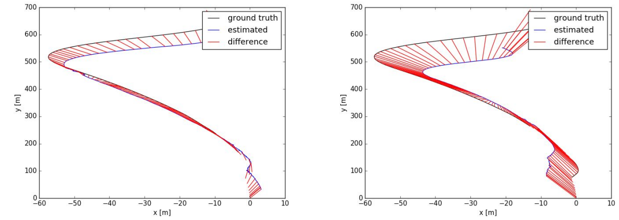


Fig. 6: Absolute trajectory error of `Ours` (left) and `ORB-SLAM`* (right) on the Virtual KITTI traffic jam sequence.

TABLE II: Evaluation on sequence 20 of the Virtual KITTI dataset. (989.96m)

Configuration	Approach	ATE [m]	RPE [m]	DR [%]
ORB	**`Ours`**	**49.34**	**60.98**	**38.32**
	`ORB-SLAM`*	69.80	87.52	23.58
ORB + Direct	**`Ours`**	**27.78**	**34.44**	**41.71**
	`ORB-SLAM`*	66.05	78.84	18.60

In table III, the results for sequence 01 from the Synthia dataset are shown. The sequence mostly covers driving on a motorway with multiple other vehicles. Due to the position and optics of the camera, considerable parts of the image can be covered by overtaking and preceding vehicles. The comparison shows that the probabilistic model gives only slightly better results than the non-probabilistic setting. It can be seen that for scenes with moderate traffic and a diverse environment, enough features can be extracted from the static world to reduce the effect of dynamic outliers.

In figure 7 we provide an example, where the camera approaches a row of stopped cars in front of a traffic light, when the vehicles start moving the camera initially moves backwards in the baseline approach. With semantics our approach recognizes features on the cars as outliers.

TABLE III: Trajectory evaluation on the sequence Synthia 01 (411.68 m).

Configuration	ATE (m)	RPE (m)	DR (%)
`Ours`	**2.99**	**6.19**	**73.95**
`ORB-SLAM`*	4.43	7.11	70.60

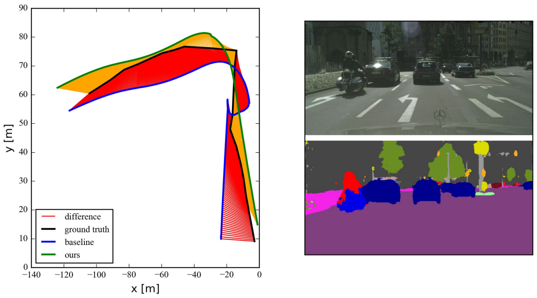


Fig. 7: CityScapes Frankfurt sequence frames 21,850-22,349 showing a backward moving camera in the baseline trajectory (blue) and a smooth trajectory for our approach (green). The ground truth (black) is given in form of low accuracy and low frequency gps positions, the only pose information provided in the dataset.

V. CONCLUSIONS

The here presented monocular SLAM approach for highly dynamic environments which models dynamic outliers with a joint probabilistic model based on semantic prior information predicted by a CNN. To find enough features with fast camera motions and in textureless environments we use a combination of descriptive and direct features. Compared to other approaches, the probabilistic outlier model allows smooth transitions between static and dynamic state, common in traffic scenes. An efficient online update obtained by approximate inference allows real-time applications. The semantic information for each pixel returns only the class of the object. In outdoor settings the intensities of static pixels, *e.g.* on the road, can be influenced by dynamic objects nearby. Especially for the direct methods, this can lead to unwanted features along the shadow borders of dynamic objects. Extending the CNN for semantic segmentation to directly predict the probability of a pixel belonging to a dynamic object based on the image context can reduce the number of observations necessary to get a reliable estimate of the inlier ratio. Using an efficient online update introduces a temporal dependency on the order in which the observations are made. With an efficient Expectation-Maximization (EM) approach this influence could be reduced in the future.

REFERENCES

[1] R. Mur-Artal, J. M. M. Montiel, and J. D. Tardos, "ORB-SLAM: a versatile and accurate monocular SLAM system," *Transactions on Robotics*, vol. 31, no. 5, 2015.

[2] J. Engel, T. Schöps, and D. Cremers, "LSD-SLAM: Large-scale direct monocular SLAM," in *European Conference on Computer Vision (ECCV)*, 2014.

[3] J. Engel, V. Koltun, and D. Cremers, "Direct sparse odometry," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 3, pp. 611–625, 2018.

[4] C. Forster, M. Pizzoli, and D. Scaramuzza, "SVO: Fast semi-direct monocular visual odometry," in *International Conference on Robotics and Automation (ICRA)*. IEEE, 2014.

[5] G. Klein and D. Murray, "Parallel tracking and mapping for small AR workspaces," in *International Symposium on Mixed and Augmented Reality (ISMAR)*. IEEE, 2007.

[6] R. A. Newcombe, S. J. Lovegrove, and A. J. Davison, "DTAM: Dense tracking and mapping in real-time," in *2011 International Conference on Computer Vision*. IEEE, nov 2011.

[7] W. Tan, H. Liu, Z. Dong, G. Zhang, and H. Bao, "Robust monocular slam in dynamic environments," in *International Symposium on Mixed and Augmented Reality (ISMAR)*. IEEE, 2013.

[8] H. Zhao, X. Qi, X. Shen, J. Shi, and J. Jia, "ICNet for Real-Time Semantic Segmentation on High-Resolution Images," *arXiv preprint arXiv:1704.08545*, 2017.

[9] J. Shimamura, M. Morimoto, and H. Koike, "Robust vSLAM for dynamic scenes," *Proceedings of the 12th IAPR Conference on Machine Vision Applications*, 2011.

[10] M. Keller, D. Lefloch, M. Lambers, S. Izadi, T. Weyrich, and A. Kolb, "Real-time 3d reconstruction in dynamic scenes using point-based fusion," in *International Conference on 3DTV*. IEEE, 2013.

[11] N. D. Reddy, P. Singhal, V. Chari, and K. M. Krishna, "Dynamic body vslam with semantic constraints," in *International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2015.

[12] V. Vineet, O. Miksik, M. Lidegaard, M. Nießner, S. Golodetz, V. A. Prisacariu, O. Kähler, D. W. Murray, S. Izadi, P. Pérez, *et al.*, "Incremental dense semantic stereo fusion for large-scale semantic scene reconstruction," in *International Conference on Robotics and Automation (ICRA)*. IEEE, 2015.

[13] J. McCormac, A. Handa, A. Davison, and S. Leutenegger, "SemanticFusion: Dense 3d semantic mapping with convolutional neural networks," in *International Conference on Robotics and Automation (ICRA)*. IEEE, 2017.

[14] J. P. Valentin, S. Sengupta, J. Warrell, A. Shahrokni, and P. H. Torr, "Mesh based semantic modelling for indoor and outdoor scenes," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2013.

[15] K. Tateno, F. Tombari, I. Laina, and N. Navab, "CNN-SLAM: Real-time dense monocular SLAM with learned depth prediction," in *Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2017.

[16] A. Kundu, Y. Li, F. Dellaert, F. Li, and J. M. Rehg, "Joint semantic segmentation and 3d reconstruction from monocular video," in *Computer Vision – ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part VI*. Springer International Publishing, 2014.

[17] S. Sengupta, E. Greveson, A. Shahrokni, and P. H. Torr, "Urban 3D semantic modelling using stereo vision," *Proceedings - IEEE International Conference on Robotics and Automation*, 2013.

[18] B. D. Lucas and T. Kanade, "An iterative image registration technique with an application to stereo vision," in *Proceedings of the 7th International Joint Conference on Artificial Intelligence - Volume 2*, ser. IJCAI'81. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1981.

[19] G. Vogiatzis and C. Hernández, "Video-based, real-time multi-view stereo," *Image and Vision Computing*, vol. 29, no. 7, 2011.

[20] J. Civera, A. J. Davison, and J. M. M. Montiel, "Inverse Depth Parameterization for Monocular {SLAM}," *IEEE Transactions on Robotics*, vol. 24, no. 5, 2008.

[21] M. Z. Zia, L. Nardi, A. Jack, E. Vespa, B. Bodin, P. H. Kelly, and A. J. Davison, "Comparative design space exploration of dense and semi-dense SLAM," in *2016 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, may 2016.

[22] J. Sturm, N. Engelhard, F. Endres, W. Burgard, and D. Cremers, "A benchmark for the evaluation of RGB-D SLAM systems," *IEEE International Conference on Intelligent Robots and Systems*, 2012.

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

[24] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," *CoRR*, vol. abs/1604.01685, 2016.

[25] A. Gaidon, Q. Wang, Y. Cabon, and E. Vig, "Virtual worlds as proxy for multi-object tracking analysis," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2016.

[26] G. Ros, L. Sellart, J. Materzynska, D. Vazquez, and A. M. Lopez, "The SYNTHIA Dataset: A Large Collection of Synthetic Images for Semantic Segmentation of Urban Scenes," *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.