VSO: Visual Semantic Odometry

Konstantinos-Nektarios Lianos [1,⋆], Johannes L. Schönberger [2],
Marc Pollefeys [2,3], Torsten Sattler [2]

[1] Geomagical Labs, Inc., USA [3] Microsoft, Switzerland
[2] Department of Computer Science, ETH Zürich, Switzerland
nelianos@geomagical.com {jsch,marc.pollefeys,sattlert}@inf.ethz.ch

Abstract. Robust data association is a core problem of visual odometry, where image-to-image correspondences provide constraints for camera pose and map estimation. Current state-of-the-art direct and indirect methods use short-term tracking to obtain continuous frame-to-frame constraints, while long-term constraints are established using loop closures. In this paper, we propose a novel visual semantic odometry (VSO) framework to enable medium-term continuous tracking of points using semantics. Our proposed framework can be easily integrated into existing direct and indirect visual odometry pipelines. Experiments on challenging real-world datasets demonstrate a significant improvement over state-of-the-art baselines in the context of autonomous driving simply by integrating our semantic constraints.

Keywords: visual odometry, SLAM, semantic segmentation

1 Introduction

Visual Odometry (VO) algorithms track the movement of one or multiple cameras using visual measurements. Their ability to determine the current position based on a camera feed forms a key component of any type of embodied artificial intelligence, *e.g.*, self-driving cars or other autonomous robots, and of any type of intelligent augmentation system, *e.g.*, Augmented or Mixed Reality devices.

At its core, VO is a data association problem, as it establishes pixel-level associations between images. These correspondences are simultaneously used to build a 3D map of the scene and to track the pose of the current camera frame relative to the map. Naturally, such a local tracking and mapping approach introduces small errors in each frame. Accumulating these errors over time leads to drift in the pose and map estimates. In order to reduce this drift, constraints between corresponding image observations are used to jointly optimize poses and map, *e.g.*, using an extended Kalman Filter [33] or bundle adjustment [26, 46].

In general, there are two orthogonal approaches to reduce drift in VO. The first uses short-term correspondences between images to enable temporal drift correction by transitively establishing constraints between subsequent camera frames. This is especially useful in automotive scenarios where a car drives along

⋆ This work was done while Konstantinos-Nektarios Lianos was at ETH Zürich.

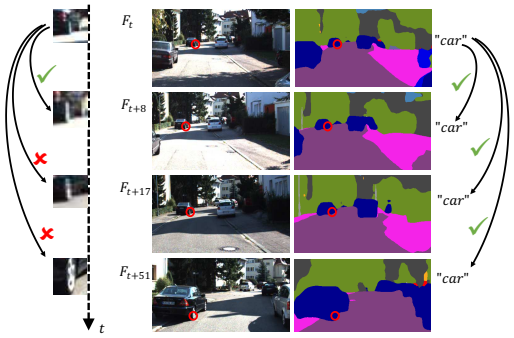

Fig. 1. Patch trackability under severe scale variation. Each row shows an image at time $t + \tau$. The point in red is tracked. Its patch appearance changes drastically but its semantic identity remains the same. While appearance-based tracking fails, image semantics can be used to establish medium-term constraints.

a straight path for a significant amount of time. The second approach establishes long-term constraints between temporally far-away frames using loop closure detection. The latter is only effective if the camera intersects its previous trajectory multiple times or in the case of localization against a pre-built map [30].

In this paper, we propose an approach to improve upon the first drift correction strategy using semantics for *medium-term continuous tracking* of points. The main limitation of the existing state of the art in this scenario is a lack of invariant representations: Both feature-based approaches, *e.g.*, ORB-SLAM [34, 35], and direct methods based on minimizing a photometric error, *e.g.*, LSD-SLAM [13] or DSO [12], are not able to continuously track a point over long distances as both representations are not fully invariant to viewpoint and illumination changes. An example for such a scenario is shown in Fig. 1, where the missing invariance of patch representations to scale changes prevents us from establishing medium-term correspondences while a car drives down a road.

The main idea of this paper is to use semantics as an invariant scene representation. The underlying intuition is that changes in viewpoint, scale, illumination, *etc.* only affect the low-level appearance of objects but not their semantic meaning. As illustrated in Fig. 1, scene semantics thus enable us to establish longer-term constraints, enabling us to significantly reduce drift in VO systems. Based on this idea, this paper derives a novel visual *semantic* odometry (VSO) approach that integrates semantic constraints into pose and map optimization.

In detail, this paper makes the following **contributions**: **1**) We derive a novel cost function for minimizing *semantic* reprojection errors and show that it can be minimized using an expectation maximization (EM) scheme. Our approach is flexible in the sense that it can be combined with any semantic segmentation algorithm. **2**) We demonstrate that including our semantic cost term into VO algorithms significantly reduces translational drift in the context of autonomous driving. Our approach can be readily integrated into existing VO approaches, independently of whether they rely on direct or indirect methods for data as-

sociation. **3**) We experimentally analyze the behavior of our approach, explain under which conditions it offers improvements, and discuss current restrictions.

2 Related Work

The large body of existing visual odometry systems can be categorized based on the employed optimization approach [46] (filtering or non-linear optimization), the sampling of observations (sparse or dense), and the data association approach (direct or indirect). In this paper, we aim at improving data association by introducing a semantic error term. As shown in Sec. 4, this allows us to reduce drift for both direct and indirect methods. As such, our proposed approach is orthogonal to the existing VO methods we review below. Most related to our work are methods that use semantics for VO or for image-to-model alignment.

Direct methods minimize the photometric error of corresponding pixels in consecutive camera frames [2, 12–14, 36, 48, 53]. The optimization objective is to align the camera poses such that the reprojected scene optimally explains the observed image intensities. The underlying energy functional is based on image gradients and thus typically requires a good initialization of the camera pose and scene structure to converge. In contrast, our proposed system aims to increase the convergence radius of the energy by incorporating longer-term constraints derived from semantic segmentation. In addition, photometric error metrics are generally not robust to even small viewpoint or illumination changes [37]. As a consequence, most direct methods track points only over a short time window. Our semantic constraints complement the accurate, short-term photometric constraints by increasing the trackability of points to larger time windows.

Indirect methods minimize the reprojection error between 3D map points and their observed projections in the image [10, 20, 21, 34, 35]. Indirect visual odometry methods typically use a sparse sampling of observations in the image by detecting and matching local features. As a result, (sparse) indirect methods are typically more robust to viewpoint and illumination changes [37]. Due to their local nature, feature detectors and descriptors are not fully invariant against such changes [31, 32]. Thus, indirect methods are still subject to the same principal limitation of direct methods and fail to track points over longer time frames. In contrast, we incorporate semantic information that is derived globally from the entire image. Sec. 4 shows that incorporating such global information into a state-of-the-art indirect visual odometry system [35] significantly reduces drift due to adding medium-term constraints between images.

Semantic mapping approaches focus on constructing semantic 3D maps from images and their known poses [6, 18, 24, 44, 47, 52]. The maps are built by jointly reasoning about semantics and geometry using fixed camera positions. As a by-product, our approach also generates a semantically annotated 3D map. However, we focus on jointly optimizing semantics, geometry, and camera poses.

Semantic visual odometry methods use higher-level features, such as lines [22], planes [4, 19, 28], or objects [3–5, 7, 16, 43] to improve the robustness of VO

or to obtain richer map representations [39, 42]. Conversely, VO can be used to improve object detection [11,15,25,38]. Most similar to our approach are object-based SLAM [3, 5, 7, 15, 40, 43] and Structure-from-Motion [4, 16] approaches. They use object detections as higher-level semantic features to improve camera pose tracking [4,7,15,16] and / or to detect and handle loop closures [3,5,15,43]. While some approaches rely on a database of specific objects that are detected online [7,15,43], others use generic object detectors [3,5,16]. The former require that all objects are known and mapped beforehand. The latter need to solve a data association problem to resolve the ambiguities arising from detecting the same object class multiple times in an image. Bowman *et al*. were the first to jointly optimize over continuous camera poses, 3D point landmarks, and object landmarks (represented by bounding volumes [5, 16]) as well as over discrete data associations [5]. They use a probabilistic association model to avoid the need for hard decisions. In contrast, our approach does not need a discrete data association by considering continuous distances to object boundaries rather than individual object detections. By focusing on the boundaries of semantic objects, we are able to handle a larger corpus of semantic object classes. Specifically, we are able to use both convex objects as well as semantic classes that cannot be described by bounding boxes, such as street, sky, and building. Compared to [5], who focus on handling loop closures, our approach aims at reducing drift through medium-term continuous data associations.

Semantic image-to-model alignment methods use semantics to align images with 3D models [8, 45, 50, 51]. Cohen *et al*. stitch visually disconnected models by measuring the quality of an alignment using 3D point projections into a semantically segmented image. Taneja *et al*. estimate an initial alignment between a panorama and a 3D model based on semantic segmentation [50]. They then alternate between improving the segmentation and the alignment. Most closely related to our approach is concurrent work by Toft *et al*. [51], who project semantically labeled 3D points into semantically segmented images. Similar to us, they construct error maps for each class via distance fields. Given an initial guess for the camera pose, the errors associated with the 3D points are then used to refine the pose. They apply their approach to visual localization and thus assume a pre-built and pre-labeled 3D model. In contrast, our approach is designed for VO and optimizes camera poses via a semantic error term while simultaneously constructing a labeled 3D point cloud. Toft *et al*. incrementally include more classes in the optimization and fix parts of the pose at some point. In contrast, our approach directly considers all classes.

3 Visual Semantic Odometry

The goal of this paper is to reduce drift in visual odometry by establishing continuous medium-term correspondences. Since both direct and indirect VO approaches are often not able to track a point over a long period of time continuously, we use scene semantics to establish such correspondences.

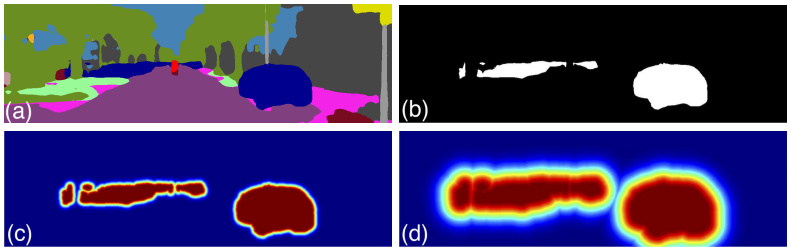

Fig. 2. Illustration of the semantic likelihood derivation. The example regards the *car* class (blue) in the input segmentation in (a) and its binary image $\mathbb{I}_{S=car}$ in (b). Semantic likelihoods $p(S|X, T, z = car)$ are shown for $\sigma = 10$ in (c) and for $\sigma = 40$ in (d), where red corresponds to value 1 and blue to value 0.

The idea behind this approach is illustrated in Fig. 1: Consider a 3D point (marked by the red circle) situated on the wheel of a parking car. As we move closer, the appearance of the patch surrounding the point changes so drastically that we are soon unable to associate it with the point's first observation. As a result, we cannot establish sufficient constraints between frame F_t and later frames to effectively prevent drift in the estimated trajectory. While the image-level appearance of the point changes, its semantic identity, *i.e.*, being part of a car, remains the same. Associating the point with a semantic label and enforcing consistency, *i.e.*, that the point labeled as *car* projects into an image region labeled as *car*, thus enables the creation of medium-term constraints. The scenario shown in Fig. 1 is prevalent in the case of forward motion in the automotive domain, where points are often visible for a long time. As our experiments show, the illustrated problem affects both direct and indirect methods.

In the following, we formalize our semantic constraints: Sec. 3.1 proposes our visual semantic odometry framework. Sec. 3.2 and 3.3 derive our semantic cost function and its optimization. Finally, Sec. 3.4 describes how the semantic cost function can be integrated into existing VO pipelines.

3.1 Visual Semantic Odometry Framework

In general, we can integrate our proposed system into any standard window-based visual odometry system, which we denote as the *base* system in the following. Given a set of input images $\mathcal{I} = \{I\}_{k=1}^{K}$, visual odometry tackles the problem of jointly optimizing the set of camera poses $\mathcal{T} = \{T\}_{k=1}^{K}$, with $T_k \in SE(3)$, and map points $\{P\}_{i=1}^{N}$ using a given set of corresponding observations $z_{i,k}$. Each map point is typically represented by its location $X_i \in \mathbb{R}^3$. An observation is either defined as a keypoint location in an image (in the case of indirect methods) or an image intensity (in the case of direct methods). To make real-time optimization feasible, point cross-correlations are typically ignored and the odometry objective functional is thus formulated as

$$E_{base} = \sum_k \sum_i e_{base}(k, i) \quad . \tag{1}$$

Here, the function $e_{base}(k, i)$ is the cost of the i-th point induced in the k-th camera. This function is either defined as a photometric (direct methods) or geometric error (indirect methods).

We now describe our proposed semantic cost function that can be readily combined with E_{base}. For each input image I_k, we require a dense pixel-wise semantic segmentation $S_k : \mathbb{R}^2 \rightarrow \mathcal{C}$, where each pixel is labeled as one of $|\mathcal{C}|$ classes from the set $\mathcal{C}$. In addition to its location X_i, each map point is thus also associated with a categorical variable $Z_i \in \mathcal{C}$. $p(Z_i = c|X_i)$ is the probability of a point P_i at position X_i to be of class c. We denote the label probability vector for each point P_i as $w_i \in \mathbb{R}^C$, where $w_i^{(c)} = p(Z_i = c|X_i)$ is the probability that point P_i belongs to class c. This probability vector is estimated online from semantic segmentations. Intuitively, the objective of our proposed semantic energy encourages point projections to be both semantically and photometrically/geometrically consistent.

To incorporate our semantic constraints into the odometry optimization functional, we define the *semantic cost function*

$$E_{sem} = \sum_k \sum_i e_{sem}(k, i) \ , \tag{2}$$

where each term associates the camera pose T_k and point P_i, represented by its label Z_i and location X_i, with the semantic image observation S_k. We optimize the base and semantic costs in the joint functional

$$\{\hat{X}\}, \{\hat{T}\} = \arg\min \ E_{base} + \lambda E_{sem} \ , \tag{3}$$

where λ weights the different terms, as explained in detail in the following section.

3.2 Semantic Cost Function

We follow a probabilistic approach and first define the observation likelihood model $p(S_k|T_k, X_i, Z_i = c)$, associating the semantic observations S_k with the camera pose T_k and point P_i. The intuition behind our observation model is that a semantic point observation $p(S_k|T_k, X_i, Z_i = c)$ should be likely if the pixel corresponding to X_i's projection $\pi(T_k, X_i)$ into S_k is labeled with c. The likelihood should decrease with the distance of $\pi(T_k, X_i)$ to the nearest region labeled as c. To implement this concept, we make use of the distance transform $DT_B(p) : \mathbb{R}^2 \rightarrow \mathbb{R}$, where $p \in \mathbb{R}^2$ is the pixel location and B the binary image on which the distance transform is defined (*c.f.* Fig. 2). More precisely, we compute a binary image $\mathbb{I}_{S_k=c}$ for each semantic class c such that pixels with label c in S_k have a value of 1 and all other pixels have value 0 (*c.f.* Fig. 2(b)). We then define a distance transform $DT_k^{(c)}(p) = DT_{\mathbb{I}_{S_k=c}}(p)$ based on this binary image (*c.f.* Fig. 2(c)). Using $DT_k^{(c)}(p)$, we define the observation likelihood as

$$p(S_k|T_k, X_i, Z_i = c) \propto e^{-\frac{1}{2\sigma^2} DT_k^{(c)}(\pi(T_k, X_i))^2} \ , \tag{4}$$

where π again is the projection operator from world to image space and σ models the uncertainty in the semantic image classification. For brevity, we omit the normalization factor that ensures that the sum over the probability space is 1. For a detailed derivation of Eq. 4, including its underlying assumptions, we refer to the supplementary material [1]. Fig. 2 illustrates the semantic likelihood for an example image. For a point with label c, the likelihood decreases proportionally to the distance from the image area labeled as c. Intuitively, maximizing the likelihood thus corresponds to adjusting the camera pose and point position such that the point projection moves towards the correctly labeled image area.

Using the observation likelihood (Eq. 4), we define a semantic cost term as

$$\begin{aligned} e_{sem}(k,i) &= \sum_{c \in \mathcal{C}} w_i^{(c)} \log(p(S_k|T_k, X_i, Z_i = c)) \\ &= -\sum_{c \in \mathcal{C}} w_i^{(c)} \cdot \tfrac{1}{\sigma^2} DT_k^{(c)}(\pi(T_k, X_i))^2 \ , \end{aligned} \tag{5}$$

where $w_i^{(c)}$ again is the probability that P_i is of class $c \in \mathcal{C}$. Intuitively, the semantic cost $e_{sem}(k,i)$ for a given the semantic image S_k and point P_i is a weighted average of 2D distances. Each distance $DT_k^{(c)}(\pi(T_k, X_i))$ of the point projection $\pi(T_k, X_i)$ to the nearest area of class c is weighted by the probability w_i that P_i is of class c. For example, if P_i has label *car* with high certainty, then its cost is the distance of the point projection to the closest area labelled as *car* in S_k. If P_i has labels *sidewalk* and *road* with equal probability, its cost is lowest on the boundary between the two classes.

The label probability vector w_i for point P_i is computed by jointly considering all of its observations. Concretely, if P_i is observed by a set of cameras $\mathcal{T}_i$, then

$$w_i^{(c)} = \tfrac{1}{\alpha} \prod_{k \in \mathcal{T}_i} p(S_k|T_k, X_i, Z_i = c) \ . \tag{6}$$

The constant α ensures that $\sum_c w_i^{(c)} = 1$. This rule allows for incremental refinement of the label vector w_i by accumulating semantic observations. If the observations have the same mode, *i.e.*, they have their maximum value for the same class, then the element-wise multiplication and normalization will cause the vector w_i to converge to a single mode corresponding to the true label.

The uncertainty parameter σ in Eq. 5 plays an important role as it defines the level of trust in the segmentation and thus in the label assignment for point P_i. For an example, consider a point projecting outside the *car* class in Fig. 2. We consider only two existing classes, *car* and *no_car*. The point is of class Z_i = *car*, with $w_i^{(car)} = 1$. Then the weight of the residual in Eq. 5 is inversely proportional to the value of σ. What is more, if σ is high, then from Eq. 4, the class likelihood for a point will be almost uniform, canceling out the semantic residual for a point for competing classes (*i.e.*, close to the boundary of objects).

[1] The supplementary material for this work is available online at http://cvg.ethz.ch/research/visual-semantic-odometry/.

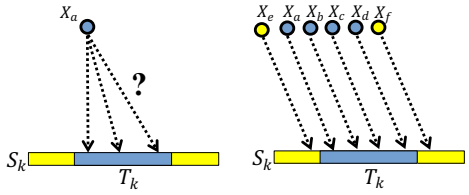

Fig. 3. *Left:* The optimization of a camera pose and a single point using semantics has infinitely many solutions. *Right:* Using multiple fixed points to optimize the camera pose constrains the solution and semantic optimization becomes feasible.

3.3 Optimization

Eqs. 5 and 6 result in a coupled structure for the optimization as both depend on the 3D point positions, the associated categorical variables, and the camera poses. For tractability, we use expectation maximization (EM) to minimize the functional E_{joint} in an alternating fashion: The E-step computes the weight vector w_i for each point P_i based on Eq. 6 while keeping the point positions and camera poses fixed. The M-step in turn optimizes the point positions and camera poses while fixing the weights. Since E_{sem} has a sparse structure, the M-step is implemented efficiently by using the Levenberg-Marquardt algorithm [27, 29] inside a sparse non-linear solver, *e.g.*, Ceres [1] or G^{2}o [23].

The presented optimization framework can be formally derived by modeling the point labels Z_i as latent variables and performing maximum likelihood inference using EM. Due to space constraints, we skip this derivation and directly explain the resulting optimization strategy. We refer the interested reader to the supplementary material for the derivation.

Using our proposed semantic formulation, we benefit from invariance but lack structural information. Optimizing the map points and camera poses with only the semantic term thus makes the problem under-constrained, as the likelihood (Eq. 4) is uniform inside object boundaries. This is illustrated in Fig. 3(left), where any projection of the 3D point into the blue image area will result in the same cost. To avoid this problem, the optimization of E_{sem} is performed as follows: 1) The semantic optimization is performed jointly with the base visual odometry functional. 2) Multiple points and semantic constraints are used to optimize a single camera pose. 3) As mentioned above, our semantic cost is under-constrained by itself. Points providing only semantic constraints and no base constraints, *i.e.*, points that are no longer optimized by the base system, are thus fixed and we optimize only their corresponding camera poses to reduce drift. This approach not only keeps the number of optimizable variables bounded, but introduces structural correlation between points as well, thus constraining the pose solution (see Fig. 3). 4) By frequent semantic optimization, we reduce the probability of associating a point to a wrong object instance. Since the optimization is based on the gradient of the distance transform, we ensure that a point stays close to a correctly labelled area and is thus pulled towards it.

3.4 Obtaining Semantic Constraints & System Integration

The semantic objective of Eq. 5 assumes the creation of constraints between a camera k with semantic image S_k and a point P_i. To establish these correspondences, we follow a standard approach: The base VO system creates a set of visible points $V(k)$ for each frame k. For each point in $V(k)$, an optimizable camera-point constraint is created. To update this list, direct VO methods compare the intensity of a candidate point P_i with the intensity of the image I_k at the projected location, while indirect methods use feature matching. Analogously, we also maintain a semantic visibility list $V_{sem}(k)$ for each frame k. A candidate point P_i is inserted into $V_{sem}(k)$ if the projection of the point i into the image k is sufficiently close to an area with the semantic class of the point. Here, allowing a certain amount of semantic reprojection error when establishing semantic constraints is necessary to be able to handle drift.

VO approaches typically maintain an *active window* (AW) of a few (key-)frames that are used to optimize the trajectory based on photometric/geometric constraints. Similarly, we define an *active semantic window* (ASW) of (key-)frames. Once a frame leaves the AW, we consider adding it to the ASW. We thereby try to limit the number of frames in the ASW while trying to cover as large a part of the trajectory as possible. The poses of frames in the ASW are not optimized anymore as they usually lack photometric/geometric constraints with the current frame. A more detailed description on how to obtain semantic correspondences from existing VO pipelines and on integrating our approach into existing systems can be found in the supplementary material.

4 Experimental Evaluation

In this section, we experimentally demonstrate that integrating our semantic medium-term constraints into state-of-the-art VO approaches significantly reduces translational drift. The focus of our experiments is an autonomous driving scenario with long periods without turns as this is the case where our semantic constraints can be most useful. We use the KITTI [17] and PlayingForBenchmarks (P4B) [41] datasets. A laptop with a quad-core 2.50GHz CPU, 8GB RAM, and a NVIDIA 1080 GPU was used for our experimental evaluation.

4.1 Experimental Setup

For semantic segmentation, we use the raw output of an off-the-shelf semantic classifier [54] pre-trained offline on the Cityscapes dataset [9]. While this semantic classifier achieves state-of-the-art performance on Cityscapes, it often introduces severe errors on the KITTI dataset (see Fig. 5). To account for the classification errors, the uncertainty is modeled by the parameters λ (weighting the relative importance of the semantic cost term) and σ (modeling the uncertainty of the classifier) in Eqs. 3 and 4, which we choose empirically per sequence. σ is measured in pixels and depends on the quality of the segmentation and the

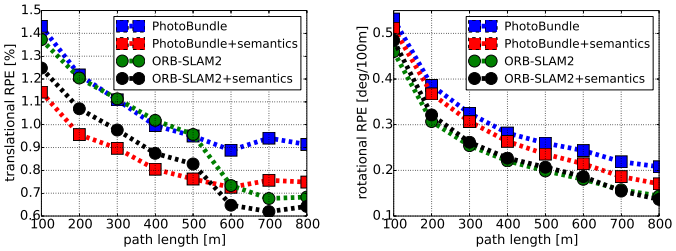

Fig. 4. Translation (left) and rotation (right) RPE as a function of the trajectory length, averaged over the sequences of the KITTI dataset. We report results for stereo ORB-SLAM2 and PhotoBundle with and without our semantic constraints. As can be seen, semantic constraints reduce the translational drift for both methods. The average was computed without sequences 01, 04, 06, 10.

image resolution. λ depends on the type of base cost (reprojection/photometric) and the classifier performance. For the P4B experiments, we show results using the ground truth semantic labels, thus establishing an upper bound on the potential of our method. To outline the versatility of our proposed framework, we implemented it on top of two VO pipelines, as detailed in the following.

We chose **ORB-SLAM2** [34] as a state-of-the-art representative of indirect VO methods. As shown in [53], stereo ORB-SLAM2 is state-of-the-art in terms of translation accuracy. We run the system with the default real-time settings and de-activate loop closing and global bundle adjustment, as we focus on showing the benefits of our semantic constraints in a pure VO setting. Notice that our constraints are complimentary to loop closure constraints as the latter only reduce drift in the case a place is re-visited. We experiment both with stereo and monocular ORB-SLAM2, denoting the latter as "mono-ORB-SLAM2".

We chose **PhotoBundle** [2] as a direct VO method. In contrast to LSD-SLAM [13] and DSO [12], which use custom-made optimizers, PhotoBundle uses Ceres [1] as its backend, allowing for an easy integration of our semantic constraints. For comparability with [2], we also use $3\times$ downsampled KITTI images. Equivalent to the original PhotoBundle approach, we use stereo depth maps to initialize 3D points, but do not enforce stereo constraints during the optimization. Depth maps are solely computed using static and not temporal stereo, resulting in failures in scenarios where horizontal lines are dominant. As such, the PhotoBundle base framework fails on one of the KITTI sequences and performs slightly worse than ORB-SLAM2. However, we are primarily interested in demonstrating the relative improvement by integrating our method. In contrast to ORB-SLAM2, PhotoBundle executes the tracking and mapping threads serially and is missing the SIMD parallelization of other direct approaches, *e.g.*, [13]. Thus, it does not operate in real-time.

Evaluation metrics. Following standard practice [49], we measure the RMSE of the Relative Pose Error (RPE) and the Absolute Trajectory Error (ATE) for each method. RPE measures the average deviation of the estimated from

Table 1. KITTI RPE results for translation t_{rel} (%) and rotation r_{rel} (deg./100m), averaged over 100m to 800m intervals (lower is better). We also report the relative improvements $t_{rel}^{(\%)}$ and $r_{rel}^{(\%)}$ (in %, higher is better) obtained with semantic constraints. The mean RPEs exclude sequences 01, 04 and 10

	ORB-SLAM2		+semantics				PhotoBundle		+semantics			
Seq.	t_{rel}	r_{rel}	t_{rel}	r_{rel}	$t_{rel}^{(\%)}$	$r_{rel}^{(\%)}$	t_{rel}	r_{rel}	t_{rel}	r_{rel}	$t_{rel}^{(\%)}$	$r_{rel}^{(\%)}$
00	1.00	**0.42**	**0.97**	**0.42**	2.51	-0.13	1.17	0.46	1.03	0.45	11.94	1.94
01	**1.79**	**0.31**	1.95	**0.31**	-8.76	0.62	-	-	-	-	-	-
02	1.05	**0.32**	**1.02**	0.34	2.10	-6.49	1.47	0.39	1.36	0.40	7.15	-1.54
03	1.93	**0.24**	**1.86**	0.27	3.52	-12.84	3.67	0.35	3.25	0.35	11.37	0.75
04	1.19	0.13	1.30	**0.10**	-9.52	17.31	0.81	0.28	**0.80**	0.27	1.60	3.36
05	0.87	**0.30**	0.82	**0.30**	6.64	1.92	0.94	0.41	**0.81**	0.41	13.48	-0.16
06	1.10	0.29	**0.98**	**0.26**	10.55	8.96	1.87	0.33	1.03	0.30	44.79	11.17
07	0.81	**0.38**	0.75	0.41	6.96	-7.03	0.70	0.42	**0.65**	0.40	7.38	5.05
08	1.33	**0.35**	1.26	**0.35**	5.11	-1.08	1.25	0.43	**1.18**	0.44	5.75	-0.84
09	1.10	0.29	1.07	**0.28**	3.09	4.22	1.04	0.35	**0.93**	0.34	10.97	3.02
10	1.25	**0.37**	1.28	0.38	-2.24	-1.88	**1.15**	**0.37**	1.17	**0.37**	-2.00	-0.62
mean	1.15	**0.33**	**1.09**	**0.33**	5.06	-1.56	1.51	0.39	1.28	0.39	14.10	2.42

the ground truth trajectory over intervals of fixed length. ATE measures the absolute difference between points on the two trajectories. For the monocular experiments, we follow the literature [14, 34] and calculate the ATE after performing 7-DoF alignment. For the stereo experiments, we measure the *Relative RPE* in % in order to quantify the relative reduction in drift obtained using semantic constraints. For the translational error, the *Relative RPE* $t_{rel}^{(\%)}$ is defined as $t_{rel}^{(\%)} = 100 \cdot (t_{rel}^{base} - t_{rel}^{joint})/t_{rel}^{base}$, where t_{rel}^{base} and t_{rel}^{joint} are the translation RPE values obtained without and with our constraints, respectively. The relative rotational RPE $r_{rel}^{(\%)}$ is defined accordingly.

4.2 Results

We quantitatively measure the impact of integrating our semantic constraints on drift accumulated during VO using relative and absolute error metrics.

KITTI Dataset. Fig. 4 shows the RPE of ORB-SLAM2 and PhotoBundle as a function of the trajectory length for the KITTI dataset. The plots were obtained by averaging over sub-trajectories of different lengths over different KITTI sequences. Using semantic constraints significantly reduces the translational drift for both direct and indirect VO methods.

We observe that our semantic constraints have limited impact on the rotational RPE. This is not surprising as we observe little rotational drift when the car travels along a straight path. Rotational drift mainly occurs during turns, *i.e.*, in situations in which semantics cannot provide medium-term constraints as the 3D points quickly leave the field-of-view of the cameras.

Tab. 1 shows the RPEs for the individual sequences of the KITTI benchmark. For most of the scenes, we observe a consistent improvement of up to 45% for the translational errors compared to the baselines. This improvement is consistent

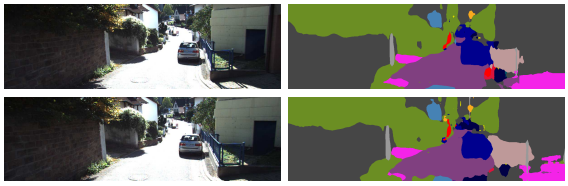

Fig. 5. Consecutive frames from KITTI sequence 10 with inconsistent semantic segmentations. Increasing the uncertainty σ in the semantic segmentation places less weight on our semantic constraints and allows us to handle such scenes. In general, small values for σ are preferable if shape details are consistent.

for both ORB-SLAM2 and PhotoBundle. For the few scenes that perform worse in terms of translational error, the negative impact is comparatively small, with the exception of sequences 01 and 04, as discussed in detail further below.

Tab. 2 shows the ATE metric for PhotoBundle, ORB-SLAM2, and mono-ORB-SLAM2. Especially for the latter, major improvement can be observed. Monocular VO is particularly challenging in the automotive domain, due to the forward motion which lacks significant parallax. Scale drift is usually a major source of error, as map points leave the field of view and scale information is discarded. Semantics help to preserve camera-point associations and thus the scale for longer intervals. The absolute improvement is further visualized in Fig. 6 for sequences 00 and 09.

P4B Dataset. This dataset consists of monocular synthetic images in an urban environment with ground-truth semantic segmentations. Due to high speed and sudden rotations, this benchmark is particularly challenging. In this evaluation, we select a subset of sequences to showcase the improvement obtained by incorporating semantics into mono-ORB-SLAM2. For all experiments, we ignore the *unlabeled* and *void* class and exclude any moving objects with labels 20-31. Fig. 6 shows trajectories of representative sequences, where monocular tracking using the base framework was feasible, while Tab. 3 shows numeric ATE results.

Failure cases. While our method leads to a large overall improvement in relative and absolute error metrics, we observed a few interesting failure cases in the KITTI sequences that we analyze in the following. In sequence 01, the classification in the highway segment is erroneous. Thus, outliers located in the background are introduced, which remain in the field of view for a long time. The resulting incorrect medium-term constraints lead to an increase in translational drift for ORB-SLAM2. Furthermore, PhotoBundle fails for sequence 01 due to perceptual aliasing caused by horizontal lines in the highway segment. In sequence 04, another car drives in front of the camera over the whole trajectory. Semantic constraints are successfully created on this moving object that remain in the field of view for a long period of time. As our approach implicitly assumes that the scene remains static, moving objects naturally lead to wrong semantic associations and thus an increase in drift. Excluding the "car" class from the semantic optimization for ORB-SLAM2+semantics solves this problem and

Table 2. ATE, in meters, on KITTI without / with semantics. The mean reductions in ATE are 0.55m, 0.51m, and 9.06m for ORB-SLAM2, PhotoBundle, and mono-ORB-SLAM2, respectively. We skip sequence 01 as PhotoBundle fails to handle it

	00	**02**	**03**	**04**	**05**
ORB-SLAM2	3.99 / 3.11	9.71 / 7.90	3.20 / 3.15	1.21 / 1.36	2.36 / 2.20
PhotoBundle	4.67 / 4.45	14.10 / 13.41	6.32 / 5.40	0.62 / 0.80	3.52 / 3.35
mono-ORB-SLAM2	56 / 45	25 / 23	2.0 / 2.1	1.4 / 1.9	27 / 19
	06	**07**	**08**	**09**	**10**
ORB-SLAM2	2.64 / 2.14	1.11 / 1.06	4.04 / 3.74	4.22 / 3.34	1.99 / 2.10
PhotoBundle	4.81 / 2.72	0.94 / 0.84	6.38 / 6.26	6.78 / 5.80	1.45 / 1.45
mono-ORB-SLAM2	47.1 / 40.5	13.6 / 12.5	50 / 42	43 / 43	6.8 / 7.7

Table 3. ATE error, in meters, on the Playing For Benchmarks (P4B) dataset for mono-ORB-SLAM2 without / with semantics. Each sequence is run 5 times. The ATE is calculated after 7-DoF alignment with the ground truth trajectory. Only the day sequences of P4B dataset for which at least 80% of the sequence can be tracked are considered. The results are obtained using the ground truth semantic labels

001	**002**	**003**	**005**	**065**	**067**
1.48 / 1.12	13 / 12	22 / 17	1.07 / 0.97	4 / 3.5	51 / 38
006	**044**	**045**	**051**	**069**	
14 / 8.5	6.0 / 3.0	68 / 57	25 / 16	57 / 51	

reduces the translational and rotational RPE to 1.23% and 0.13 deg./100m, respectively. As shown qualitatively in the supplementary material, this slightly larger drift compared to pure ORB-SLAM2 is caused by inconsistent semantic segmentations. We observed that stationary cars typically provide excellent semantic constraints as they are typically well-segmented and visible for a long time. Rather than excluding entire semantic classes, instance-level segmentation should be used in practice to distinguish between moving and stationary objects. In sequence 10, the semantic classifier performs particularly bad (see Fig. 5), resulting in an increase of the RPEs. In such cases, the reason for the decrease in performance is an over-estimation of the classification accuracy by setting a low uncertainty σ in Eq. 4. Choosing a larger value for σ down-weights the influence of the semantic cost, resulting in the same performance as the baselines.

Runtime results. The runtime of our system directly depends on the number of semantic constraints in the optimization. In general, the number of semantic constraints is scene and motion dependent, as shown in the supplementary material. For ORB-SLAM2, we use an average of 35 semantic constraints for KITTI, leading to a negligible computational overhead over the baseline. As a result, we achieve real-time performance when integrating semantic constraints into ORB-SLAM2. In contrast to the sparse measurements in ORB-SLAM2, PhotoBundle uses dense intensity-based measurements, resulting in 944 semantic constraints on average. In this setting, the joint optimization is on average (over all frames and sequences) 1.5× slower than base PhotoBundle. Executing the semantic op-

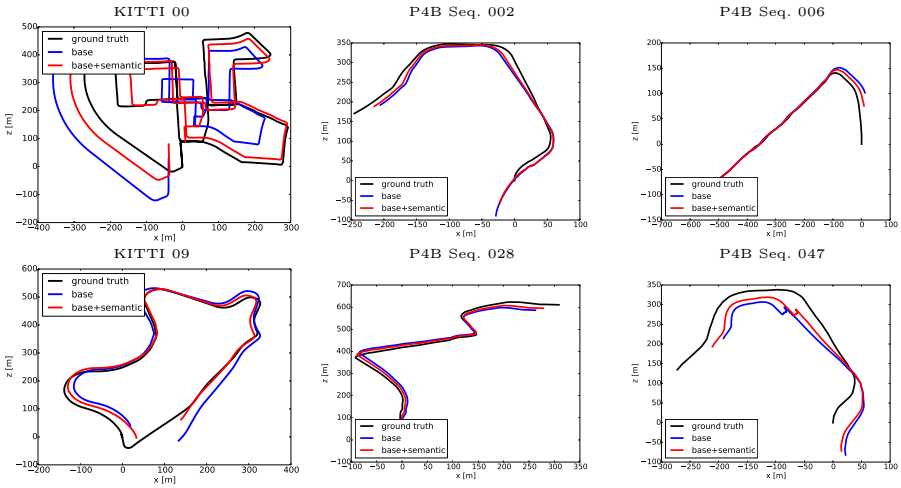

Fig. 6. Trajectory plots for mono-ORB-SLAM2 on sequences from KITTI and PlayingForBenchmarks (P4B). All sequences are 7-DoF aligned with the ground truth.

timization every 4th frame reduces the overhead to 1.125× at negligible loss in accuracy, allowing for real-time execution using PhotoBundle.

5 Conclusion

In this paper, we have proposed a novel visual semantic odometry (VSO) framework that can be readily integrated into existing VO systems. Our method harnesses the invariance of semantic object representations to incorporate medium-term constraints into the odometry objective. By appropriately handling the lack of structure of the semantic identity, we are able to effectively and significantly reduce translational drift. We have demonstrated consistent performance improvements for both direct and indirect systems in a challenging real-world scenario. The bottleneck of our method is the accuracy of the semantic segmentation, especially along object boundaries. In the future, we plan to experiment with multi-camera systems, for which we expect an even bigger improvement using semantics, since objects are continuously trackable for longer duration. In addition, class-specific uncertainty modeling could improve the performance and solve some of the current failure cases, *e.g.*, due to dynamic objects. Finally, we envision a system where geometry not only benefits from semantics but where both modalities are tightly coupled, thereby facilitating end-to-end learning of a geometric and semantic understanding of the world in real-time.

Acknowledgements. This project received funding from the European Union's Horizon 2020 research and innovation program under grant No. 688007 (TrimBot2020).

References

1. Agarwal, S., Mierle, K., Others: Ceres solver. http://ceres-solver.org
2. Alismail, H., Browning, B., Lucey, S.: Photometric Bundle Adjustment for Vision-Based SLAM. In: Asian Conference on Computer Vision (ACCV) (2016)
3. Atanasov, N., Zhu, M., Daniilidis, K., Pappas, G.J.: Semantic Localization Via the Matrix Permanent. In: Robotics: Science and Systems (RSS) (2014)
4. Bao, S.Y., Savarese, S.: Semantic structure from motion. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2011)
5. Bowman, S.L., Atanasov, N., Daniilidis, K., Pappas, G.J.: Probabilistic data association for semantic SLAM. In: IEEE International Conference on Robotics and Automation (ICRA) (2017)
6. Cherabier, I., Schönberger, J.L., Oswald, M., Pollefeys, M., Geiger, A.: Learning Priors for Semantic 3D Reconstruction. In: European Conference on Computer Vision (ECCV) (2018)
7. Civera, J., Gálvez-López, D., Riazuelo, L., Tardós, J.D., Montiel, J.: Towards semantic SLAM using a monocular camera. In: IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) (2011)
8. Cohen, A., Sattler, T., Pollefeys, M.: Merging the Unmatchable: Stitching Visually Disconnected SfM Models. In: IEEE International Conference on Computer Vision (ICCV) (2015)
9. Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B.: The Cityscapes Dataset for Semantic Urban Scene Understanding. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016)
10. Davison, A.J., Reid, I.D., Molton, N.D., Stasse, O.: MonoSLAM: Real-time single camera SLAM. IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI) **29**(6), 1052–1067 (2007)
11. Dong, J., Fei, X., Soatto, S.: Visual Inertial Semantic Scene Representation for 3D Object Detection. IEEE Conference on Computer Vision and Pattern Recognition (CVPR (2017)
12. Engel, J., Koltun, V., Cremers, D.: Direct Sparse Odometry. IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI) **40**(3), 611–625 (2018)
13. Engel, J., Schöps, T., Cremers, D.: LSD-SLAM: Large-scale direct monocular SLAM. In: European Conference on Computer Vision (ECCV) (2014)
14. Forster, C., Zhang, Z., Gassner, M., Werlberger, M., Scaramuzza, D.: SVO: Semidirect Visual Odometry for Monocular and Multicamera Systems. IEEE Transactions on Robotics (T-RO) **33**(2), 249–265 (2017)
15. Gálvez-López, D., Salas, M., Tardós, J.D., Montiel, J.: Real-time monocular object slam. Robotics and Autonomous Systems **75**, 435–449 (2016)
16. Gay, P., Rubino, C., Bansal, V., Del Bue, A.: Probabilistic Structure From Motion With Objects (PSfMO). In: IEEE International Conference on Computer Vision (ICCV) (2017)
17. Geiger, A., Lenz, P., Urtasun, R.: Are we ready for Autonomous Driving? The KITTI Vision Benchmark Suite. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2012)
18. Häne, C., Zach, C., Cohen, A., Pollefeys, M.: Dense Semantic 3D Reconstruction. IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI) **39**(9), 1730–1743 (2017)

19. Henry, P., Krainin, M., Herbst, E., Ren, X., Fox, D.: RGB-D mapping: Using Kinect-style depth cameras for dense 3D modeling of indoor environments. International Journal of Robotics Research (IJRR) **31**(5), 647–663 (2012)
20. Jin, H., Favaro, P., Soatto, S.: Real-time 3D motion and structure of point features: a front-end system for vision-based control and interaction. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2000)
21. Klein, G., Murray, D.: Parallel tracking and mapping for small AR workspaces. In: IEEE and ACM International Symposium on Mixed and Augmented Reality (ISMAR) (2007)
22. Kottas, D.G., Roumeliotis, S.I.: Efficient and consistent vision-aided inertial navigation using line observations. In: IEEE International Conference on Robotics and Automation (ICRA) (2013)
23. Kümmerle, R., Grisetti, G., Strasdat, H., Konolige, K., Burgard, W.: G^2o: A General Framework for Graph Optimization. In: IEEE International Conference on Robotics and Automation (ICRA) (2011)
24. Kundu, A., Li, Y., Dellaert, F., Li, F., Rehg, J.M.: Joint semantic segmentation and 3d reconstruction from monocular video. In: European Conference on Computer Vision (ECCV) (2014)
25. Leibe, B., Cornelis, N., Cornelis, K., Van Gool, L.: Dynamic 3d scene analysis from a moving vehicle. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2007)
26. Leutenegger, S., Lynen, S., Bosse, M., Siegwart, R., Furgale, P.: Keyframe-based visual-inertial odometry using nonlinear optimization. International Journal of Robotics Research (IJRR) **34**(3), 314–334 (2015)
27. Levenberg, K.: A method for the solution of certain non-linear problems in least squares. Quarterly of Applied Mathematics **2**(2), 164–168 (1944)
28. Ma, L., Kerl, C., Stückler, J., Cremers, D.: CPA-SLAM: Consistent Plane-Model Alignment for Direct RGB-D SLAM. In: IEEE International Conference on Robotics and Automation (ICRA) (2016)
29. Marquardt, D.W.: An algorithm for least-squares estimation of nonlinear parameters. Journal of the Society for Industrial and Applied Mathematics **11**(2), 431–441 (1963)
30. Middelberg, S., Sattler, T., Untzelmann, O., Kobbelt, L.: Scalable 6-DOF Localization on Mobile Devices. In: European Conference on Computer Vision (ECCV) (2014)
31. Mikolajczyk, K., Schmid, C.: A performance evaluation of local descriptors. IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI) **27**(10), 1615–1630 (2005)
32. Mikolajczyk, K., Tuytelaars, T., Schmid, C., Zisserman, A., Matas, J., Schaffalitzky, F., Kadir, T., Van Gool, L.: A comparison of affine region detectors. International Journal of Computer Vision (IJCV) **65**(1), 43–72 (2005)
33. Mourikis, A.I., Trawny, N., Roumeliotis, S.I., Johnson, A.E., Ansar, A., Matthies, L.: Vision-Aided Inertial Navigation for Spacecraft Entry, Descent, and Landing. IEEE Transactions on Robotics (T-RO) **25**(2), 264–280 (2009)
34. Mur-Artal, R., Montiel, J.M.M., Tardos, J.D.: ORB-SLAM: a Versatile and Accurate Monocular SLAM System. IEEE Transactions on Robotics (T-RO) **31**(5), 1147–1163 (2015)
35. Mur-Artal, R., Tardós, J.D.: ORB-SLAM2: an Open-Source SLAM System for Monocular, Stereo and RGB-D Cameras. IEEE Transactions on Robotics (T-RO) **33**(5), 1255–1262 (2017)

36. Newcombe, R.A., Lovegrove, S.J., Davison, A.J.: DTAM: Dense tracking and mapping in real-time. In: IEEE International Conference on Computer Vision (ICCV). pp. 2320–2327 (2011)
37. Park, S., Schöps, T., Pollefeys, M.: Illumination change robustness in direct visual SLAM. In: IEEE International Conference on Robotics and Automation (ICRA) (2017)
38. Pillai, S., Leonard, J.: Monocular SLAM Supported Object Recognition. In: Robotics: Science and Systems (RSS) (2015)
39. Pronobis, A.: Semantic Mapping with Mobile Robots. Ph.D. thesis, KTH Royal Institute of Technology, Stockholm, Sweden (2011)
40. Reid, I.: Towards semantic visual SLAM. In: International Conference on Control Automation Robotics & Vision (ICARCV) (2014)
41. Richter, S.R., Hayder, Z., Koltun, V.: Playing for Benchmarks. In: IEEE International Conference on Computer Vision (ICCV) (2017)
42. Rusu, R.B., Marton, Z.C., Blodow, N., Dolha, M., Beetz, M.: Towards 3D point cloud based object maps for household environments. Robotics and Autonomous Systems **56**(11), 927–941 (2008)
43. Salas-Moreno, R.F., Newcombe, R.A., Strasdat, H., Kelly, P.H., Davison, A.J.: Slam++: Simultaneous localisation and mapping at the level of objects. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2013)
44. Savinov, N., Häne, C., Ladický, L., Pollefeys, M.: Semantic 3D Reconstruction with Continuous Regularization and Ray Potentials Using a Visibility Consistency Constraint. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016)
45. Schönberger, J.L., Pollefeys, M., Geiger, A., Sattler, T.: Semantic Visual Localization. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2018)
46. Strasdat, H., Montiel, J.M.M., Davison, A.J.: Visual slam: Why filter? Image Vision Computing **30**(2), 65–77 (2012)
47. Stückler, J., Waldvogel, B., Schulz, H., Behnke, S.: Dense real-time mapping of object-class semantics from RGB-D video. Journal of Real-Time Image Processing **10**(4), 599–609 (2015)
48. Stühmer, J., Gumhold, S., Cremers, D.: Real-time dense geometry from a handheld camera. In: Joint Pattern Recognition Symposium (2010)
49. Sturm, J., Engelhard, N., Endres, F., Burgard, W., Cremers, D.: A Benchmark for the Evaluation of RGB-D SLAM Systems. In: IEEE/RSJ International Conference on Intelligent Robot Systems (IROS) (2012)
50. Taneja, A., Ballan, L., Pollefeys, M.: Registration of Spherical Panoramic Images with Cadastral 3D Models. In: International Conference on 3D Imaging, Modeling, Processing, Visualization Transmission (3DIMPVT) (2012)
51. Toft, C., Olsson, C., Kahl, F.: Long-Term 3D Localization and Pose from Semantic Labellings. In: IEEE International Conference on Computer Vision (ICCV) Workshops (2017)
52. Vineet, V., Miksik, O., Lidegaard, M., Nießner, M., Golodetz, S., Prisacariu, V.A., Kähler, O., Murray, D.W., Izadi, S., Pérez, P., Torr, P.H.S.: Incremental dense semantic stereo fusion for large-scale semantic scene reconstruction. In: IEEE International Conference on Robotics and Automation (ICRA) (2015)
53. Wang, R., Schwörer, M., Cremers, D.: Stereo DSO: Large-Scale Direct Sparse Visual Odometry with Stereo Cameras. In: IEEE International Conference on Computer Vision (ICCV) (2017)
54. Yu, F., Koltun, V.: Multi-Scale Context Aggregation by Dilated Convolutions. In: International Conference on Learning Representations (ICLR) (2016)