Scene Flow Propagation for Semantic Mapping and Object Discovery in Dynamic Street Scenes

Deyvid Kochanov, Aljoša Ošep, Jörg Stückler and Bastian Leibe

***Abstract*— Scene understanding is an important prerequisite for vehicles and robots that operate autonomously in dynamic urban street scenes. For navigation and high-level behavior planning, the robots not only require a persistent 3D model of the static surroundings—equally important, they need to perceive and keep track of dynamic objects. In this paper, we propose a method that incrementally fuses stereo frame observations into temporally consistent semantic 3D maps. In contrast to previous work, our approach uses scene flow to propagate dynamic objects within the map. Our method provides a persistent 3D occupancy as well as semantic belief on static as well as moving objects. This allows for advanced reasoning on objects despite noisy single-frame observations and occlusions. We develop a novel approach to discover object instances based on the temporally consistent shape, appearance, motion, and semantic cues in our maps. We evaluate our approaches to dynamic semantic mapping and object discovery on the popular KITTI benchmark and demonstrate improved results compared to single-frame methods.**

I. INTRODUCTION

Great progress has recently been achieved in the development of vehicles that operate autonomously in urban street scenes. Such systems need in-depth scene understanding to navigate safely in complex everyday traffic scenarios. For motion planning and navigation, the vehicle not only requires a 3D map of its static surrounding, but it should also keep track of moving objects in the scene. It should be able to parse task-relevant semantics in the scene and observe object instances for which pre-trained detectors are not available or would be difficult to obtain.

In this paper, we propose a novel approach to 3D semantic mapping with stereo cameras that explicitly takes the motion in the scene into account (see Fig. 1): our method maps 3D occupancy of the static scene parts, and – importantly – propagates and updates moving objects in the dynamic map using stereo depth and scene flow. We probabilistically filter image-based semantic segmentations within these maps in order to obtain temporally and spatially consistent semantic 3D segmentations. Based on this persistent 3D semantic representation of the dynamic environment, we propose an object discovery approach that finds object instances based on the shape, appearance, semantics, and motion cues maintained in our maps.

In contrast to previous approaches to semantic mapping in urban street scenes (e.g. [1], [2]), our approach filters a probabilistic belief on occupancy and semantics of static as well as dynamic objects. This not only has the potential to improve semantic segmentation alone. By discovering objects in the persistent dynamic map, our proposal-generating method is less susceptible to noise in the depth or scene flow observed in single stereo frames. Object discovery can make use of semantic segmentation from a series of views, such that objects become observable even through occlusions.

All authors are with the Computer Vision Group, Visual Computing Institute, RWTH Aachen University, Aachen, Germany `deyvid.kochanov@rwth-aachen.de`, `{osep,stueckler,leibe}@vision.rwth-aachen.de`

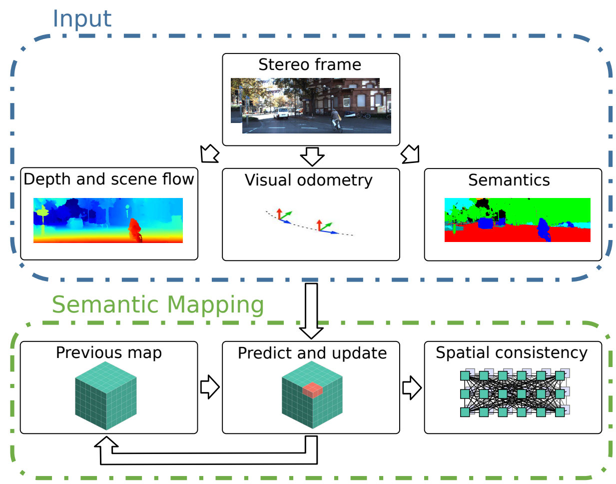

Fig. 1. We map dynamic environments in 3D occupancy grid maps using stereo visual odometry, depth, and scene flow. Image-based semantic segmentation is additionally temporally filtered in the semantic maps and made spatially consistent in a dense CRF. Scene motion is compensated for by warping the mapped dynamic objects with the estimated scene flow.

In experiments, we demonstrate the performance of our approach to semantic 3D mapping and object discovery in dynamic urban traffic scenes. We evaluate our method on sequences of the popular KITTI benchmark suite [3]. We compare our approach to single-frame methods and demonstrate improvements in terms of the accuracy of semantic segmentation and the quality of object proposals.

The main contributions of our work are summarized as follows: (1) We propose a method for 3D occupancy mapping in dynamic environments based on state-of-the-art methods for stereo depth, scene flow, and semantic segmentation. (2) We fuse the depth measurements and the semantic segmentation of individual stereo frames within our dynamic maps to induce temporal consistency. To this end, we propose a probabilistic filtering technique that warps the occupancy and semantics belief in the map using the observed scene flow. In addition, we use a dense CRF on the 3D voxel grid in order to enforce spatial consistency. (3) We propose an approach to object discovery based on the aggregated shape, appearance, motion and semantic cues in the maps.

II. RELATED WORK

Recent trends in the development of autonomous vehicles and robots have fuelled research in semantic mapping and scene understanding. Simultaneous localization and mapping in urban street scenes with vision sensors has attracted much attention in recent years. Current state-of-the-art methods such as ORB-SLAM [4] or LSD-SLAM [5], [6] demonstrate consistent large-scale trajectory estimation and 3D reconstruction. While these methods can cope with a limited number of moving objects as outliers to the SLAM process, they are inherently designed for static environments. Only few SLAM methods have been proposed that explicitly distinguish between the static parts of the environment and dynamic objects (e.g. SLAMMOT [7], [8], or [9]). In the indoor RGB-D SLAM domain, KinectFusion [10] and point-based fusion techniques [11] have been proposed that can separate dynamic parts from the static background and exclude them from tracking and mapping. Recently, Newcombe et al. [12] propose DynamicFusion, a SLAM method that takes non-rigid motion in a small-scale scene into account to map a canonical shape model of the deforming object.

None of the aforementioned methods incorporates scene semantics such as the categorization of surfaces into car, building or road, which is often argued to be an important aspect in scene understanding. Approaches in this line of research can be distinguished by the way the semantic segmentation is obtained and how this information is integrated and made temporally and spatially consistent in a global map. Koppula et al. [13] investigate semantic mapping using RGB-D sensors in indoor scenes. They label the 3D points in an aggregated point cloud map and impose a Markov random field model on the points with appearance and geometric features. Hermans et al. [14] also semantically label a 3D point cloud map which they obtain using RGB-D visual odometry. Their approach, however, first segments individual RGB-D images with a random forest classifier and probabilistically filters the soft labelling of the random forest in the 3D points. A dense 3D conditional random field (CRF) enforces spatial consistency on the semantic labels of the point cloud map. The semantic SLAM approach by Stückler et al. [15] filters the semantic segmentation from an RF classifier in multi-resolution surfel maps, while concurrently performing keyframe-based SLAM based on the map representation. Differently to our method, the aforementioned approaches assume the environment to remain static during the semantic mapping process.

Semantic mapping in outdoor street scenes is considered by Sengupta et al. [16]. In this work, stereo images are first segmented into object classes with a CRF approach. In a second CRF stage, the image-based semantic segmentation is fused on a ground plane projection. In [1], Sengupta et al. fuse a CRF semantic labeling of stereo images in a 3D point cloud map. Floros and Leibe [17] use a higher-order CRF in order to enforce a consistent semantic labelling of 3D points that backproject into individual image-based semantic segmentations. Valentin et al. [18] use a triangle mesh to represent the map. They segment the mesh with a CRF using image-based appearance and mesh-based geometry features.

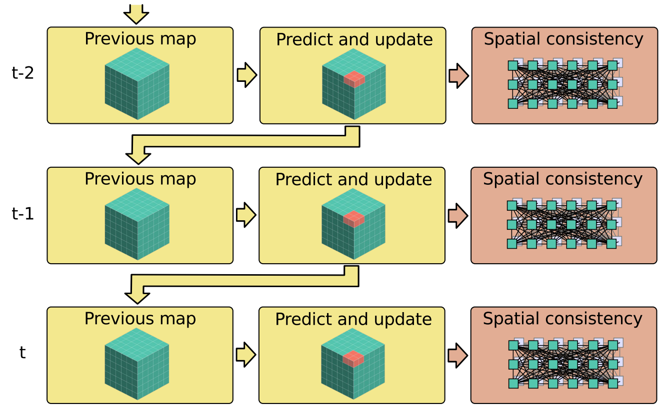

Fig. 2. We enforce temporal consistency through recursive Bayesian filtering (yellow shaded parts) and spatial consistency with a dense CRF (brown shaded parts) in our semantic maps.

In contrast to our approach, these semantic mapping methods assume the scene to remain static during mapping. Very recently, Vineet et al. [2] addressed this shortcoming. They apply random forest segmentation to individual stereo images and fuse the labelled stereo depth in a 3D truncated signed distance function (TSDF) representation using memory-efficient voxel hashing. For voxels annotated with a potentially movable object class, they allow for a faster decay of the integrated TSDF values, such that their measurements are removed more easily from the map from conflicting observations towards the static background. We explicitly take the estimated motion on such objects into account and propagate occupancy belief in a 3D map with scene flow. By doing this, we can probabilistically filter the semantic labelling of static parts as well as moving objects.

We also demonstrate the utility of these combined features in a temporally integrated map representation for object discovery. Several previous approaches to object discovery use single-image cues such as geometric or appearance-based saliency (e.g. [19], [20], [21]). Some methods also discover objects from motion cues over multiple frames [22], [23]. Our approach filters geometry, appearance, semantics and motion cues in a consistent map over subsequent frames. We apply clustering in this temporally integrated multi-cue map in order to discover objects.

III. OVERVIEW

Our approach takes as an input a sequence of stereo image pairs and incrementally reconstructs a 3D semantic map (see Fig. 1 for an illustration of the overall approach). More precisely, the map is represented as a 3D voxel grid which stores information about occupancy and semantic object-class categories in the contained volume. For 3D mapping, we determine depth from the stereo images and estimate the camera pose using visual odometry in order to obtain the camera trajectory from which we integrate the measured depth in a global map reference frame.

A major challenge in typical street scenes are moving objects. If we simply update the voxels and do not take motion into account, moving objects would leave trails of erroneous occupancy belief in the map. We thus compute scene flow, i.e., the 3D motion of each stereo image pixel, in order to take motion into account. We accumulate the flow in the voxels and use it to propagate the occupancy and semantic belief on dynamic objects within the map. To this end, we propose an efficient grid warping procedure which also takes the uncertainty of the flow measurements into account.

At every time step, we also extract semantic information from the stereo images and filter it in the map through time. We compute an image-level semantic segmentation which results in a probabilistic per-pixel label assignment. Using the label distribution, we perform a Bayesian update on the semantic category of the corresponding voxel.

This method of updating the map allows us to integrate information from multiple frames (see Fig. 2). A main advantage of the integration over time is the ability to enforce temporal consistency in the semantic labeling and 3D reconstruction. We additionally employ a probabilistic 3D Voxel-CRF model to enforce spatial consistency of the semantic labels. Note that we filter the distribution over semantic labels in the temporal domain and only apply a CRF on top of the accumulated voxel grid to obtain the most likely explanation of accumulated measurements in each frame. This explanation may change in future frames in the presence of new evidence.

Finally, we describe our object proposal generation method that builds on the reconstructed 3D semantic map. Our method uses a clustering algorithm on the aggregated maps that groups voxels into objects proposals based on shape, appearance, motion and semantic cues.

IV. Semantic Mapping in Dynamic Street Scenes

A. Stereo Depth and Motion Estimation

Based on the stereo frames we estimate visual odometry, image depth and scene flow, which we will further use in our semantic mapping pipeline. We build on the stereo visual odometry proposed by Geiger et al. [24]. It is a sparse keypoint-based method which is specifically designed for the stereo setup and the street scenes typical to automotive scenarios.

Given two stereo pairs, captured at consecutive time steps, scene flow methods estimate the 3D motion at each pixel in the scene. In our experiments we used the state-of-the-art scene flow method by Vogel et al. [25]. For each pixel in the images the method computes both depth and 3D flow concurrently. It approximates the scene with a set of planar segments in superpixels for which a rigid motion towards a reference stereo frame is determined. Note that we can easily subtract the scene-flow induced by the ego-motion of the camera using the visual odometry estimate.

B. Semantic Stereo Image Segmentation

Semantic segmentation aims at mapping the pixels in the stereo images to one of several category labels $l \in \{1, \ldots, L\}$ such as car, road or building. We apply the semantic segmentation approach proposed in [26] which is based on classifying supervoxels using a random forest classifier. Instead of making a hard decision on each supervoxel, it provides a probabilistic classification output in the form of a label distribution. This way, the uncertain decisions can be filtered temporally and be incorporated as per-voxel label evidence in a probabilistic spatial CRF model. We use the image and point cloud to compute the following 150-dim. features for each supervoxel within the random forest classifier as in [26].

1) Appearance Features: These features capture the color and texture statistics of the supervoxels. For the color, we compute a 10-bin color histogram in the CIELab color space for each channel. In addition, we determine the mean and covariance of the gradients in each channel. Finally, we add a histogram of textons computed as described in [27].

2) Density Features: This shape feature describes the density of 3D points around the supervoxel They are computed by estimating a ground plane and discretizing the point cloud into a grid with 3 height bins, using 3 different grid resolutions to capture context at different scales. We project each centroid of the supervoxels to the grid and examine its 4-neighbourhood at its height.

3) Spectral Features: We compute this second type of shape features from the eigenvectors and the eigenvalues of the covariance matrix of the points in the supervoxel. Looking at eigenvalues, we can quantify pointness, linearness, surfaceness and curvature [28] of the segment. We compute the surface normal from the eigenvectors and measure the orientation of the supervoxel as the angle between the surface normal and the ground plane normal.

4) Locational Features: The relative location of a supervoxel in the scene is encoded by the distance of its centroid to the ground plane, its distance from the camera, and the horizontal angle between the optical axis and the ray from the focal point to the supervoxel's centroid.

C. Mapping of Dynamic Scenes

We now arrive at our algorithm for 3D mapping of occupancy and semantics from a moving stereo camera in a dynamic (street) scene. The main challenge in this setting is to construct a mapping method which can effectively account for the motion of objects. In each time step t, our algorithm receives the current stereo frame I_t. From this image, we extract a depth map d_t and a semantic segmentation s_t which we summarize in the observation $x_t = \{d_t, s_t\}$. Semantic segmentation yields a probability distribution $p(l_u \mid I_t)$ on the labelling of each image pixel u.

In addition to these single-frame measurements, visual odometry provides us with the pose p_t of the camera in the world frame at time t. Piece-wise rigid scene flow f_t is estimated from the last frame I_{t-1} to the current frame I_t, which we compensate for the ego-motion of the camera with the visual odometry estimate. We summarize the visual odometry and scene flow estimates up to time t by $\mathcal{P}^t$ and $\mathcal{F}^t$, respectively. In analogy, we write $\mathcal{X}^t$ to denote the series of observations $x_0, \ldots, x_t$.

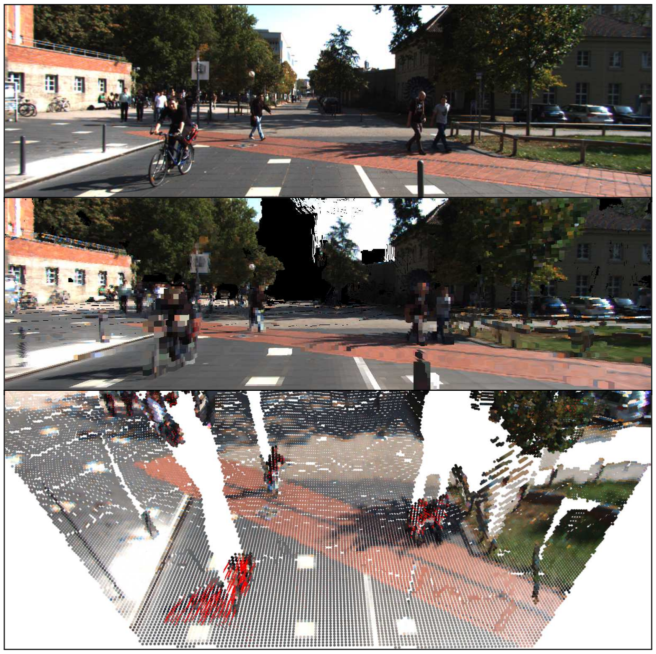

Fig. 3. Voxel map representation. Top: Corresponding image from the KITTI benchmark. Middle: Voxel map for the stereo frame, colorized with the average color of image pixels in the voxel. Bottom: Average voxel flow (red lines, voxel centers depicted as colored disks).

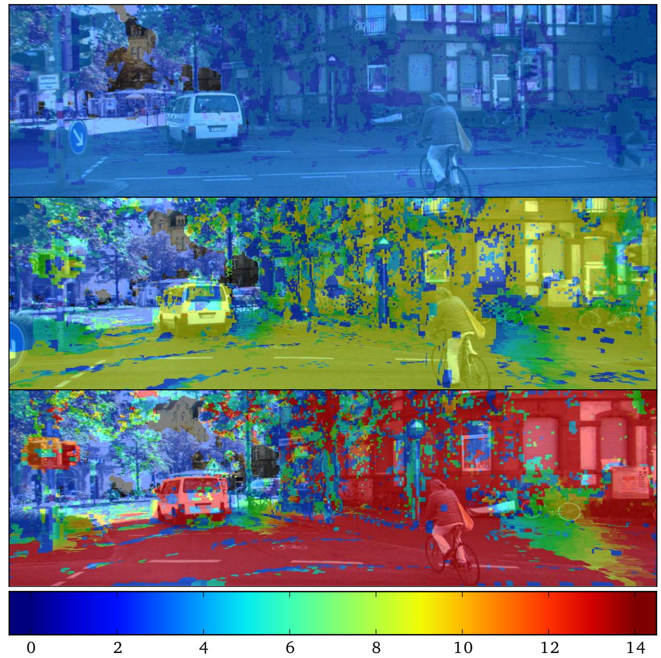

Fig. 4. Voxel-age (color-coded) in frames 1, 10, and 15 of the KITTI tracking sequence 00. Voxel-age corresponds to the number of frames in which the voxel was updated. Remarkably, due to scene flow propagation, voxels on dynamic objects exhibit similar ageing like the static parts *(best viewed in color)*.

1) Map Representation: We represent our map with voxels v_j using sparse and memory-efficient voxel hashing [29]. Each voxel of size $0.1m \times 0.1m \times 0.1m$ maintains a distributions on occupancy $p(o_j \mid \mathcal{X}^t, \mathcal{P}^t, \mathcal{F}^t)$ and semantic labelling $p(l_j \mid \mathcal{X}^t, \mathcal{P}^t, \mathcal{F}^t)$.

2) Probabilistic Mapping: The stereo images provide observations of voxel occupancy and semantic labelling, which we transform from the camera frame to the world frame using the visual odometry estimate. Scene dynamics is observed by ego-motion-compensated scene flow. Under this model, the occupancy and semantic belief in each voxel $y_{t,j}$ is updated in a recursive Bayesian filtering scheme,

$$p(y_{t,j} \mid \mathcal{X}^t, \mathcal{F}^t, \mathcal{P}^t) = \eta \, p(x_t \mid y_{t,j}, p_t) \, p(y_{t,j} \mid \mathcal{X}^{t-1}, \mathcal{F}^t, \mathcal{P}^t), \quad (1)$$

where η is a normalization factor. In the following, we will drop the dependency on the visual odometry estimates $\mathcal{P}^t$ for brevity.

The filter decomposes into a prediction and a correction step. The prediction step applies a state transition model and warps the voxel map based on the scene flow. A subsequent correction update step incorporates the image-based depth and semantic observations into the map. For efficient integration of the stereo image-based observations, we first accumulate the occupancy, semantics, and scene flow measurements in a local map. This local map is aligned with the grid of the temporally integrated map. Fig. 3 shows an example of a local measurement map m_t generated from a stereo frame.

3) Dynamic Map Prediction: In the prediction step, we determine the distribution

$$p(y_{t,j} \mid \mathcal{X}^{t-1}, \mathcal{F}^t) = \sum_k \sum_y p(y_{t,j} \mid y_{t-1,k} = y, f_{t,k}) \, p(y_{t-1,k} = y \mid \mathcal{X}^{t-1}, \mathcal{F}^{t-1}), \quad (2)$$

by propagating the occupancy and label beliefs in the voxels from the last time step based on the current scene flow estimate. Note that we model occupancy and label belief as stochastically independent, such that we can process both modalities in separate Bayesian filters.

Clearly, the estimated scene flow and the imposed model assumptions are not fully satisfied in a real setting. Hence, the state transition should also induce additional uncertainty on the occupancy and label belief. We incorporate this by approximating the state transition model with two model terms

$$p(y_{t,j} \mid \mathcal{X}^{t-1}, \mathcal{F}^t) = \sum_y p(y_{t,j} \mid \widetilde{y}_{t,j} = y) \, p(\widetilde{y}_{t,j} = y \mid \mathcal{X}^{t-1}, \mathcal{F}^t), \quad (3)$$

where $p(y_{t,j} \mid \widetilde{y}_{t,j})$ now smoothes the occupancy and label distribution in a voxel. The propagation then splits into two separate processes: In a first step, we propagate the belief from the previous frame with the scene flow to obtain an intermediate distribution on $\widetilde{y}_{t,j}$,

$$p(\widetilde{y}_{t,j} = y \mid \mathcal{X}^{t-1}, \mathcal{F}^t) = \sum_k p(\widetilde{y}_{t,j} = y \mid y_{t-1,k} = y, f_{t,k}) \, p(y_{t-1,k} = y \mid \mathcal{X}^{t-1}, \mathcal{F}^{t-1}). \quad (4)$$

The second step applies the smoothing model,

$$p(y_{t,k} = y \mid \widetilde{y}_{t-1,k} = y') = \begin{cases} \delta & \text{if } y = y' \\ \frac{1-\delta}{N-1} & \text{if } y \neq y', \end{cases} \quad (5)$$

where N is the number of state variables and δ is a parameter that controls the degree of smoothing. We apply the smoothing separately to the occupancy and semantic states.

By the separation into history- and smoothing-based transitions, we can approximate scene flow propagation using a particle propagation scheme,

$$p(\widetilde{y}_{t,j} = y \mid \mathcal{X}^{t-1}, \mathcal{F}^t) \propto \sum_{i \in \mathcal{P}_k^j(f_{t,k})} w_{t-1,k}^{[i]} \, p(y_{t-1,k} = y \mid \mathcal{X}^{t-1}, \mathcal{F}^{t-1}). \quad (6)$$

For each voxel k, we generate a set of particles $\mathcal{S} = \left\{ s_{t-1,k}^{[i]} \right\}$. The position of the particles is sampled according to the distribution of the current flow measurement f_t of the voxel. With $\mathcal{P}_k^j(f_{t,k})$ we denote the set of particles originating from voxel k that end up in voxel j through the flow $f_{t,k}$. Each particle is associated a weight $w_{t-1,k}^{[i]} = \frac{1}{N_k}$, where N_k is the number of particles sampled in the originating voxel.

Since the scene flow estimate itself is affected by noise, we resample the particle positions in the 3D voxel map under the distribution of the scene flow. Unfortunately, piece-wise rigid scene flow does not provide this distribution. Instead, we approximate it with a normal distribution centered at the scene flow and with the covariance of the difference vector between the stereo depth estimates of the corresponding pixels in the subsequent stereo frames, $\Sigma_f \approx \Sigma_{u,t-1} + \Sigma_{u,t} \approx 2\,\Sigma_{u,t}$, where $\Sigma_{u,t}$ is an approximation of the covariance of the stereo depth estimate assuming constant pixel and disparity noise. Consequently, the distribution of the scene flow is approximated with the normal distribution $\mathcal{N}(f, \Sigma_f)$. The particles are sampled from their initial voxel positions into their new voxels according to this normal distribution.

Special care needs to be taken for voxels in the free-space that gets occupied by a moving object. If particles are sampled into such a voxel, we reset the occupancy and semantic belief with the belief of the particles. Since we only use a low number of samples in each voxel, our particle scheme can be more efficient than 3D Gaussian convolutions, while it still approximates the belief propagation well.

4) Occupancy and Semantics Measurement Update: We integrate measurements into the global voxel map using Bayesian updates,

$$p(y_{t,j} \mid \mathcal{X}^t, \mathcal{F}^t) = \eta \, \frac{p(y_{t,j}^m \mid x_t)}{p(y_{t,j})} \, p(y_{t,j} \mid \mathcal{X}^{t-1}, \mathcal{F}^t), \quad (7)$$

where η is a normalization factor. Note that we use uniform priors $p(y_{t,j})$.

Instead of raycasting in the map, we determine voxels as measured free by projecting them into the stereo frame and determining if they lie in front of the measurements. In such cases, we apply a constant occupancy likelihood $p(o_t \mid x_t) = 0.2$. Measurements that fall into voxels are first accumulated in a local measurements map. For the occupancy update, we employ a method similar to the counting model of Haehnel et al. [30]. We approximate the occupancy belief in a voxel of the local measurement map directly from the point count,

$$p(o_{k,t}^m \mid x_t) = \frac{\min(\alpha, N_{k,t}^m)}{\beta} + \gamma, \quad (8)$$

where $N_{k,t}^m$ is the number of measurements in voxel k within the local measurement map m_t, and α, β, and γ are parameters of the inverse sensor model. Consequently, all voxels in the global map that contain measurements in the local measurement map are updated with the occupancy likelihood computed in eq. (8). Occluded voxels are neglected for the update.

We accumulate the semantic image segmentation in the local measurement map by averaging the label distribution of the pixels within a voxel,

$$p(l_{k,t}^m \mid x_t) = \frac{1}{|\mathcal{U}(k)|} \sum_{u \in \mathcal{U}(k)} p(l_u \mid x_t), \quad (9)$$

where $\mathcal{U}(k)$ is the set of pixels that fall into voxel k, and update the semantic belief in the integrated map accordingly.

D. Spatially Consistent Semantics

Due to the spatial coherence of objects, neighboring pixels or voxels in the map are very likely to have the same semantic category. However, the image-based classifier as well as the Bayesian mapping approach treat pixels and voxels independently. We hence enforce spatial consistency on the voxel map using a dense conditional random field (CRF) [31].

We use the belief on the label distribution in the voxels of our semantic maps as unary potentials for the CRF. The pairwise potentials model spatial and appearance-based smoothness using the kernel

$$k(f_j, f_k) = w_1 e^{\left(-\frac{\|p_j - p_k\|_2^2}{2\theta_\alpha^2} - \frac{\|c_j - c_k\|_2^2}{2\theta_\beta^2}\right)} + w_2 e^{\left(-\frac{\|p_j - p_k\|^2}{2\theta_\gamma^2}\right)},$$

where f_j and f_k are features consisting of the voxel center positions p_j, p_k and the average CIELab colors c_j, c_k in the voxels. The weight parameters w_1 and w_2 control the importance of the two kernels in the pairwise potential both in relation to each other and the unary potentials. The standard deviations θ_α, θ_β and θ_γ set their range.

E. Large-Scale Mapping

In our online mapping system, we only maintain a volume close to the current camera frustum in the temporally integrated map. In order to obtain large-scale reconstructions of the scene, we fuse the integrated map periodically into a large-scale global map. Each voxel in the global map is set to its most recent belief from the integrated map. We only transfer the voxels with an average scene flow below a threshold in order to include only the static parts. For robustness against outliers, we discard voxels below a specific voxel-age (we require at least 2 frames in our experiments).

V. OBJECT DISCOVERY IN DYNAMIC SEMANTIC MAPS

Generating object proposals from semantic maps instead of individual images has several potential advantages. Integrating semantic segmentations over time leads to a more accurate, stable labeling. The integration of occupancy information over time significantly improves the noisy stereo depth information present in a single frame. Finally, integration over time in a map helps us to generate proposals for objects which may be strongly occluded in individual frames.

We propose to employ density-based spatial clustering (DBSCAN [32]) on features extracted from the semantic map (see. Fig. 8). DBSCAN clusters points in a feature-space based on their distance in a bottom-up way. It starts clustering at core points with at least $N_{\min}$ number of points in an ε-neighborhood. It expands from these points recursively to core points within the ε-neighborhood and includes all points within the neighborhood.

As features for each occupied voxel labelled with the 'object' class label, we use a concatenation of its center position, its average color in CIELab color space, and its average scene flow. We extract proposals at multiple scales by varying the ε-neighborhood in a discrete range of values ($\varepsilon \in \{1.7, 1.9, 2.1, 2.3, 2.4, 2.6, 2.8, 3.0\}$ in our experiments). For each radius, we additionally vary the occupancy thresholds θ_o at which voxels are considered for clustering (we use the values $\theta_o \in \{0.6, 0.7, 0.8, 0.85, 0.9, 0.95, 0.98\}$ in our experiments). After computing proposals across scales, we merge them according to their bounding box overlap in the image domain and rank the merged proposals by the number of the matchings.

VI. EXPERIMENTS

We evaluate our approaches to semantic mapping and object discovery on datasets from the popular KITTI benchmark [3]. On the KITTI odometry dataset, we evaluate semantic mapping based on custom dense object-class annotations of 200 of the images. We used the same split into training and test set to train the random forest classifier and to evaluate our segmentation results as in [26]. Our method for object discovery is assessed on the tracking dataset which comes with ground truth annotations for bounding boxes on objects such as cars, pedestrians, and bicyclists.

Besides training the RF classifier, we also determined the remaining parameters of our method empirically on the training split. To find good values for the CRF parameters we performed grid search. We ran the MAP inference using the RF unaries and evaluated the semantic segmentation on 15 frames from our training set. This yielded the settings $w^{(1)} = 2.5$, $w^{(2)} = 1$, $\theta_\alpha = 2.5$, $\theta_\beta = 7$ and $\theta_\gamma = 0.3$.

A. Semantic Dynamic Mapping

We evaluate the quality of the semantic labelling in our 3D maps using the image-based ground truth annotations on the KITTI odometry dataset. To this end, we generate a semantic labelling of the stereo frames from the belief contained in our semantic maps. We directly lookup the semantic labelling of each image pixel using its depth measurement at its corresponding voxel in the map. Only if no depth is available at a pixel, we ray-cast for the label.

	RF + CRF		semantic map	
	recall	IoU	recall	IoU
object	84.02	67.39	83.02	**70.39**
road	**94.04**	91.50	93.41	**92.11**
building	83.29	75.27	**86.91**	**76.68**
tree/bush	**70.36**	**64.59**	68.97	64.35
sign/pole	**4.09**	**3.83**	1.96	1.88
sky	35.84	34.43	**40.91**	**39.75**
grass/dirt	60.63	**26.77**	**79.46**	25.23
pixel avg. (< 25 m)	87.71	78.25	**87.99**	**78.55**
class avg. (< 25 m)	68.54	57.47	**70.46**	**57.81**
pixel avg. (all depths)	78.08	70.46	**79.13**	**71.05**
class avg. (all depths)	61.75	51.97	**64.95**	**52.91**

TABLE I

SEMANTIC SEGMENTATION RESULTS ON THE KITTI ODOMETRY DATASET.

Fig. 5 shows a qualitative result of our mapping approach in a high-traffic road scene. It can be seen that the dynamic objects are well segmented from the static parts and not included in the large-scale map. The semantic segmentation finds a consistent labelling of the cars and traffic signs as objects. It also segments the larger surface categories such as road and vegetation well.

The segmentation quality is assessed using the Pascal VOC intersection-over-union measure (IoU [33]). We compare our approach with a purely image-based semantic segmentation method that applies spatial smoothing on the RF output using the dense CRF [26]. Note that for pixels without a valid depth, the pixel is assigned to a void class which is accounted as false positive if it should label one of the object classes. In Table I we report recall and IoU over all pixels, class-wise, and by class-averages on the test set. From the results we observe that on average the semantic information contained in our maps clearly outperforms the single-frame-based semantic segmentation baseline (RF+CRF). Our method also demonstrates improvements in recall and IoU over the single-frame-based segmentation on several object classes with medium-sized and larger structures. Notably, on classes which contain finer structures, image-based segmentation can perform slightly better. This is likely due to the highly noisy stereo depth that is unreliable in such thin structures. Especially at very far distances, this renders consistent integration of fine structures in the map difficult. Interestingly, averaged over all classes, the improvements by our persistent maps are stronger if we consider all depths compared to a limited range of up to 25 m (chosen as in [2]).

B. Object Discovery

For evaluating our object-instance proposal method on the temporally integrated semantic maps, we follow the evaluation protocol in [26]. We use the KITTI tracking training set [3] due to the public availability of the ground-truth annotations of object bounding boxes. Note that semantic segmentation was trained on a non-overlapping set of semantic labels and that we set the parameters of the object

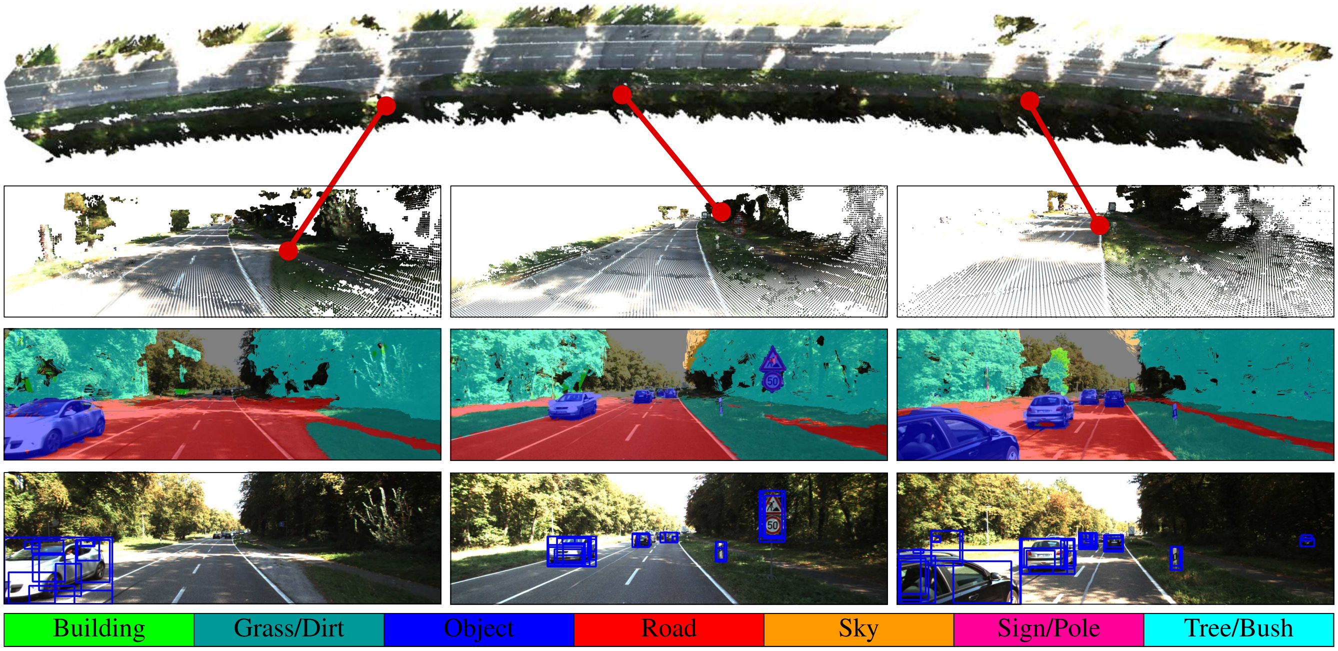

Fig. 5. Large-scale semantic mapping result on KITTI tracking seq. 18. Top: birds-eye view on large-scale map. Second row: views into large-scale map. Third row: semantics looked up in the temporally integrated map. Bottom row: object instances discovered in the temporally integrated map.

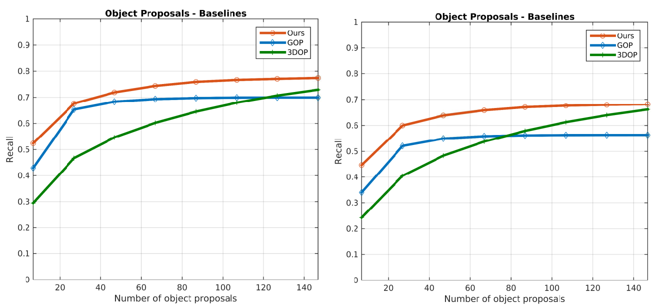

Fig. 6. Object discovery results over all categories (left: 30 m camera range; right: 50 m camera range).

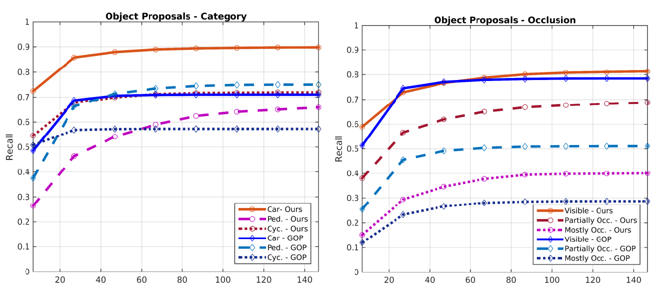

Fig. 7. Object discovery results for the three most frequently annotated categories (left) and three levels of occlusions (right) at 30 m depth range.

discovery algorithm empirically. We accept object proposals as matching a ground-truth bounding box, if they achieve an intersection-over-union value of at least 0.5.

We compare our method with the density-based multi-scale approach (GOP) [26] and state-of-the-art 3D object proposal generation method (3DOP) [34]. In order to make a fair comparison of the two methods, we use the depth maps obtained with piece-wise rigid scene flow in the baseline methods as well. Both our method and GOP [26] method use the same semantic segmentation of the stereo frames while

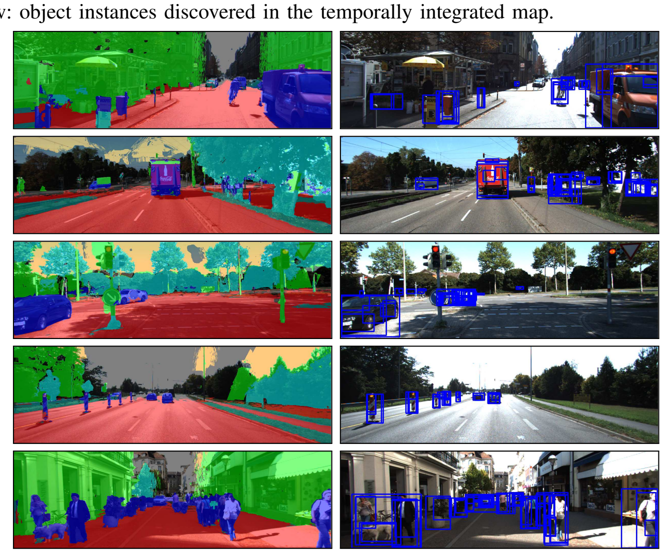

Fig. 8. Object proposal examples from the KITTI tracking sequences. Left: semantic segmentation looked up in maps temporally integrated over ten previous frames. Right: object proposals discovered on maps temporally integrated over ten previous frames.

[34] does not make use of semantic information. However, they make use of object category size statistics.

In Fig. 6, we show recall vs. number of highest ranked proposals (see Sec. V) for 30 m and 50 m depth range. The results demonstrate significant improvements compared to previous work [26], especially in the camera far-range. We assume that the reason for the improvement is two-fold: First, by performing temporal integration we are able to bridge short occlusions. Second, the method by [26] only relies on the depth and semantic measurements, while our approach also takes motion and appearance into account.

In Fig. 7 (left) we show the performance of our method on the three most frequently annotated categories on the KITTI tracking dataset, i.e. car, pedestrian, and cyclist. While our method clearly outperforms GOP on the car and cyclist categories, this baseline seems better suitable for detection of individual pedestrians. This is due to the fact that pedestrians in KITTI mostly appear in groups. Due to inaccuracies in the scene flow estimation, occupancy beliefs in these cells become blurred and groups are perceived as single objects.

Fig. 7 (right) compares our method with the baseline w.r.t. the amount of occlusion on the objects. The results clearly show that the main advantage of our method over GOP is due to the temporal integration. Finally, Fig. 8 shows example results obtained with our approach. It can be seen that our method provides proposals on a wide range of generic objects (*e.g*. truck, dog, traffic sign/poles, post-box etc.) and finds them even in difficult occlusion situations (*e.g*. the car behind the traffic sign in 3rd row).

VII. CONCLUSIONS

In this paper, we have proposed a novel approach to 3D semantic mapping and object discovery in dynamic street scenes. We use scene flow to propagate occupancy and semantic belief in the map. In this way, our maps maintain a temporally consistent semantic belief not only on the static parts of the environment as in previous approaches, but also on dynamic objects. Based on our map representation, we develop an object discovery approach that is less susceptible to occlusions and noisy observations in single stereo frames.

We develop our method as an important building-block for our future research in detailed 3D scene understanding in the close camera range. Potential next steps will be the tracking of discovered objects over time and to reason about their occupancy in separate, object-centric voxel grids.

ACKNOWLEDGEMENTS

This work was funded by ERC Starting Grant project CV-SUPER (ERC-2012-StG-307432). We would like to thank Alexander Hermans for helpful discussions.

REFERENCES

[1] S. Sengupta, E. Greveson, A. Shahrokni, and P. Torr, "Urban 3D semantic modelling using stereo vision," in *Proc. of IEEE ICRA*, 2013.
[2] V. Vineet, O. Miksik, M. Lidegaard, M. Niessner, S. Golodetz, V. Prisacariu, O. Kahler, D. Murray, S. Izadi, P. Peerez, and P. Torr, "Incremental dense semantic stereo fusion for large-scale semantic scene reconstruction," in *Proc. of IEEE ICRA*, 2015, pp. 75–82.
[3] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for Autonomous Driving? The KITTI Vision Benchmark Suite," in *Proc. of IEEE CVPR*, 2012.
[4] R. Mur-Artal, J. Montiel, and J. Tardos, "ORB-SLAM: A versatile and accurate monocular SLAM system," *IEEE Trans. on Robotics*, vol. 31, no. 5, pp. 1147–1163, 2015.
[5] J. Engel, T. Schöps, and D. Cremers, "LSD-SLAM: Large-scale direct monocular SLAM," in *Proc. of ECCV*, 2014.
[6] J. Engel, J. Stückler, and D. Cremers, "Large-scale direct SLAM with stereo cameras," in *Proc. of IEEE/RSJ IROS*, 2015.
[7] C.-C. Wang, C. Thorpe, and S. Thrun, "Online simultaneous localization and mapping with detection and tracking of moving objects: Theory and results from a ground vehicle in crowded urban areas," in *Proc. of IEEE ICRA*, 2003.
[8] E. Einhorn and H.-M. Gross, "Generic NDT mapping in dynamic environments and its application for lifelong SLAM," *Robotics and Autonomous Systems*, vol. 69, pp. 28 – 39, 2015.
[9] R. Danescu, C. Pantilie, F. Oniga, and S. Nedevschi, "Particle grid tracking system stereovision based obstacle perception in driving environments," *IEEE Intelligent Transportation Systems Magazine*, vol. 4, no. 1, pp. 6–20, 2012.
[10] R. A. Newcombe, S. Izadi, O. Hilliges, D. Molyneaux, D. Kim, A. J. Davison, P. Kohli, J. Shotton, S. Hodges, and A. W. Fitzgibbon, "KinectFusion: Real-time dense surface mapping and tracking." in *ISMAR*, 2011, pp. 127–136.
[11] 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 *Proc of Int. Conf. on 3D Vision (3DV)*, 2013, pp. 1–8.
[12] R. A. Newcombe, D. Fox, and S. M. Seitz, "DynamicFusion: Reconstruction and tracking of non-rigid scenes in real-time," in *Proc. of IEEE CVPR*, 2015.
[13] H. S. Koppula, A. Anand, T. Joachims, and A. Saxena, "Semantic labeling of 3D point clouds for indoor scenes," in *Proc. of NIPS*, 2011.
[14] A. Hermans, G. Floros, and B. Leibe, "Dense 3D semantic mapping of indoor scenes from RGB-D images," in *Proc. of IEEE ICRA*, 2014.
[15] J. Stückler, B. Waldvogel, H. Schulz, and S. Behnke, "Dense real-time mapping of object-class semantics from RGB-D video," *J. of Real-Time Image Processing*, vol. 10, no. 4, pp. 599–609, 2013.
[16] S. Sengupta, P. Sturgess, L. Ladicky, and P. Torr, "Automatic dense visual semantic mapping from street-level imagery," in *Proc. of IEEE/RSJ IROS*, 2012.
[17] G. Floros and B. Leibe, "Joint 2d-3d temporally consistent semantic segmentation of street scenes," in *Proc. of IEEE CVPR*, 2012.
[18] J. Valentin, S. Sengupta, J. Warrell, A. Shahrokni, and P. Torr, "Mesh based semantic modelling for indoor and outdoor scenes," in *Proc. of IEEE CVPR*, 2013, pp. 2067–2074.
[19] D. Mitzel and B. Leibe, "Taking Mobile Multi-Object Tracking to the Next Level: People, Unknown Objects, and Carried Items," in *Proc. of ECCV*, 2012.
[20] A. Karpathy, S. Miller, and L. Fei-Fei, "Object discovery in 3D scenes via shape analysis," in *Proc. of IEEE ICRA*, 2013.
[21] G. M. Garcia, E. Potapova, T. Werner, M. Zillich, M. Vincze, and S. Frintrop, "Saliency-based object discovery on RGB-D data with a late-fusion approach," in *Proc. of IEEE ICRA*, 2015.
[22] A. Bewley, V. Guizilini, F. Ramos, and B. Upcroft, "Online Self-Supervised Multi-Instance Segmentation of Dynamic Objects," in *Proc. of IEEE ICRA*, 2014.
[23] T. Scharwächter, M. Enzweiler, U. Franke, and S. Roth, "Stixmantics: A Medium-Level Model for Real-Time Semantic Scene Understanding," in *Proc. of ECCV*, 2014.
[24] A. Geiger, J. Ziegler, and C. Stiller, "StereoScan: Dense 3d Reconstruction in Real-time," in *Intel. Vehicles Symp.'11*, 2011.
[25] C. Vogel, K. Schindler, and S. Roth, "Piecewise rigid scene flow," in *Proc. of IEEE Int. Conf. on Computer Vision (ICCV)*, 2013.
[26] A. Ošep, A. Hermans, F. Engelmann, D. Klostermann, , M. Mathias, and B. Leibe, "Multi-scale object candidates for generic object tracking in street scenes," in *Proc. of IEEE ICRA*, 2016.
[27] J. Shotton, J. M. Winn, C. Rother, and A. Criminisi, "TextonBoost for Image Understanding: Multi-Class Object Recognition and Segmentation by Jointly Modeling Texture, Layout, and Context," *IJCV*, vol. 81, no. 1, pp. 2–23, 2009.
[28] D. Munoz, N. Vandapel, and M. Hebert, "Onboard Contextual Classification of 3-D Point Clouds with Learned High-order Markov Random Fields," in *Proc. of IEEE ICRA*, 2009.
[29] M. Nießner, M. Zollhöfer, S. Izadi, and M. Stamminger, "Real-time 3D reconstruction at scale using voxel hashing," *ACM Trans. Graphics*, 2013.
[30] D. Hähnel, "Mapping with mobile robots," Ph.D. dissertation, University of Freiburg, 2005.
[31] P. Krähenbühl and V. Koltun, "Efficient Inference in Fully Connected CRFs with Gaussian Edge Potentials," in *Proc. of NIPS*, 2011.
[32] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," in *Proc. of Int. Conf. on Knowledge Discovery and Data Mining (KDD)*, 1996.
[33] M. Everingham, L. V. Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (voc) challenge," *IJCV*, vol. 88, no. 2, pp. 303–308, 2009.
[34] X. Chen, K. Kundu, Z. Zhang, H. Ma, S. Fidler, and R. Urtasun, "Monocular 3d object detection for autonomous driving," in *Proc. of IEEE CVPR*, 2016.