2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
October 25-29, 2020, Las Vegas, NV, USA (Virtual)

Robust and Efficient Object Change Detection by Combining Global Semantic Information and Local Geometric Verification

Edith Langer, Timothy Patten and Markus Vincze

Abstract— **Identifying new, moved or missing objects is an important capability for robot tasks such as surveillance or maintaining order in homes, offices and industrial settings. However, current approaches do not distinguish between novel objects or simple scene readjustments nor do they sufficiently deal with localization error and sensor noise. To overcome these limitations, we combine the strengths of global and local methods for efficient detection of novel objects in 3D reconstructions of indoor environments. Global structure, determined from 3D semantic information, is exploited to establish object candidates. These are then locally verified by comparing isolated geometry to a reference reconstruction provided by the task. We evaluate our approach on a novel dataset containing different types of rooms with 31 scenes and 260 annotated objects. Experiments show that our proposed approach significantly outperforms baseline methods.**

I. Introduction

The ability to detect new, moved or missing objects in large environments is key for enabling many robot tasks such as surveillance, tidying up, or maintaining order in homes or workplaces. These tasks share the commonality of operating in the same environment every day. As such, revisiting a particular environment enables robots to utilize domain knowledge and to exploit their memory from previous visits. By storing a reference map of the environment, a robot can check for scene consistency and therefore detect changes on the object level. A household robot, for example, uses the cleaned-up version of the environment as a reference map to discover objects it should tidy-up (see Figure 1). While a surveillance robot knows which objects are expected in its environment and triggers an alarm when the comparison to the current state reveals a missing object. In both cases, the robot is only interested in new or removed objects, but not in objects that have a permanent place, such as a lamp or computer keyboard, which may move only slightly.

The standard approach to detect inconsistencies in the scene is to compute the difference between a reference and the current situation. This has the advantage over recognition methods, e.g. [1]–[3], since no object models are required and it is therefore suitable for open-set conditions. Some methods apply scene differencing on single frames and specify waypoints to guide the robot to regions of interest [4], [5]. This, however, restricts the search space, which leads to objects being missed if the viewpoints do not cover the whole environment. Therefore, more recent approaches compute a global scene difference on reconstructions of entire environments [6]–[11]. A disadvantage, however, is that change detection applied at a large scale is sensitive to sensor noise and localization error. Furthermore, readjustment of uninteresting objects such as furniture or decorations cannot be distinguished from new objects.

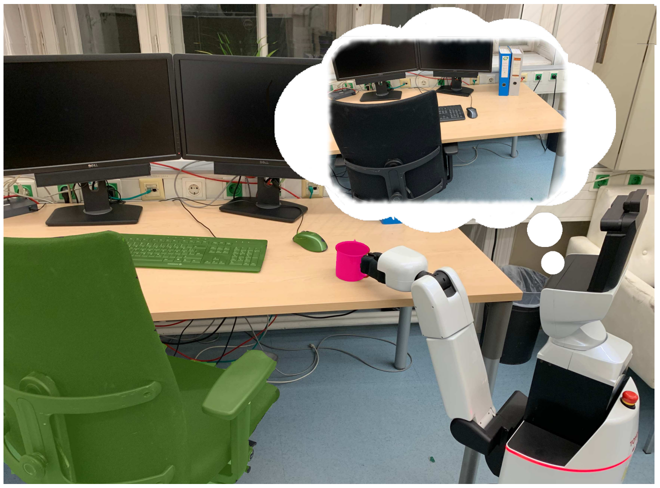

Fig. 1: A household robot tidying-up a room. It compares a previously acquired reference map to the current state of the environment. Although the chair and other permanent objects moved slightly (colored in green), only the mug (colored in pink) should be detected as novel and therefore tidied-up.

This paper presents a new approach to detect objects in real-world indoor environments based on reconstructions and overcomes the limitations of existing global scene differencing methods. Our idea is to exploit the strength of different approaches by combing full knowledge about scene context with local geometry. At a global level, semantic segmentation reveals structures where objects are likely to be located, such as on a table, couch, or floor. In agreement with the real-world fact that objects are mainly placed on surfaces [5], [12], we use relevant structures to identify horizontal planes. The set of object candidates that are extracted from the planes are processed at a local level through geometrical verification against the reference map. In contrast to global scene differencing, local alignment is robust to the effects of sensor noise and localization error.

For the quantitative evaluation of our method, we present a new annotated dataset for novel object detection. While

The research leading to these results has received funding from the Vienna Science and Technology Fund (WWTF) under grant agreement No. ICT15-045 RALLI and the Austrian Science Foundation (FWF) under grant agreement I3967-N30 BURG and I3969-N30 InDex.

All authors are with the Vision4Robotics group, Automation and Control Institute, TU Wien, 1040 Vienna, Austria {langer,patten,vincze} @acin.tuwien.ac.at.

978-1-7281-6211-9/20/$31.00 ©2020 IEEE

datasets exist for related problems, e.g. [8], [10], none simultaneously fulfill the requirements of comprising different environments, containing many objects (especially small ones), and are recorded by a mobile robot. Our new dataset consists of differently sized scene reconstructions with each scene consisting of an object-free setup (i.e. reference map) and with various additional setups containing novel objects and rearranged furniture and permanent items. Overall, we consider five distinct environments and a total of 31 reconstructions are provided including 260 annotated objects. Experiments with this new dataset show that our approach significantly outperforms the baseline methods.

In summary, the contributions of this paper include:

- Exploitation of knowledge from the task domain such as previous visits and the structure of the environment,
- Proposal of a unified approach for open-set 3D novel object change detection that combines semantic information, surface extraction, and local verification,
- A new robotic dataset for this problem comprising rooms with varying complexity and rearranged furniture and permanent objects, and
- Significantly improved detection rate of novel objects using our combination of knowledge and perception in comparison to competing approaches.

The remainder of the paper is as follows. Section II reviews relevant literature. Section III outlines our proposed approach. Section IV describes the new robotic dataset. Section V presents experimental results. Section VI concludes and discusses future work.

II. Related Work

Change detection is widely used to discover objects in 3D environments. An advantage of change detection is that it requires no training or a priori object information. Instead, the principle of scene differencing identifies the change between multiple observations for data structures such as point clouds [5], [8], [13] or voxel grids [4], [7], [9]–[11], [14]. For long-term operation, where a robot observes the same environment multiple times, the static map can be retrieved in order to apply change detection during revisits to the same scene [8], [10]. These methods have shown the capability of the principle for object discovery, however, they are confounded by sensor noise and mapping errors. As a result, post-processing steps are applied such as removing planar [5] or small clusters [8], enforcing spatial consistency with a Markov random field [6], or morphological operations like opening [10].

An alternative approach is to learn geometric descriptors from local 3D patches as in [15], [16]. Large datasets are used to learn descriptors by feeding a deep network with matching and non-matching pairs of small 3D volumes. These approaches do not require 3D models and can operate efficiently on full reconstructions. However, they detect objects by finding correspondences across scenes. In other words, they only re-localize objects. The methods are incapable of identifying newly introduced objects that have never been seen.

A final approach to find objects in 3D scenes is to establish correspondences between the scene and known 3D models. A variety of 3D descriptors have been developed for this purpose and are applied not just for detection but also for instance recognition, e.g. SHOT [1], and pose estimation, e.g. PPF [2], [17]. While these methods report accurate detection results, they rely on given 3D models. As such, they are unsuitable for detecting unknown objects.

In contrast to the existing change detection methods, we derive object candidates from semantic information and apply change detection only in local regions. This is more robust to sensor noise, map misalignment and map warping. While it is a requirement to have a reference map, this is easily created by combining observations from an environment at different times. The methods presented in [8] or [10] can be applied. The main advantage of our, and related change detection approaches, is that no specific object information is necessary. In particular, no knowledge such as shape, geometry or learned descriptors are needed. Therefore, we can identify completely novel objects, which is more general than model-based or learned local descriptor matching methods.

III. Method

This work addresses the problem of detecting novel objects in 3D environments. Novel or new objects refer to those that are introduced into the scene. These differ from permanent objects that were already present in the scene but might have moved slightly. Our approach, as outlined in Figure 2, combines multiple sources of information. Semantic information with horizontal plane detection is used at the scene level to generate an initial set of object candidates. The candidates are then verified through local geometric alignment. The local verification step overcomes the inaccuracies of global differencing because smaller regions suffer less from noise and warping in the reconstruction. The global detection stage is necessary to determine where to apply local verification, which would be time consuming if performed exhaustively.

This section describes the proposed approach. We first explain the procedure for extracting object candidates from the global scene using semantic information. We then outline the verification procedure using local geometry.

A. Object Detection from Global Semantic Context

We consider 3D reconstructions of entire rooms to be independent of single camera perspectives and robot trajectories. From the global reconstruction, semantic information is exploited to discover new objects. Semantic segmentation has received most attention in the computer vision community for pixel-wise classification of images and the rise of deep learning, in particular CNNs, has drastically improved results [18], [19]. The introduction of the ScanNet dataset [20] has enabled the transition to apply semantic segmentation to dense 3D reconstructions of indoor scenes. In this work, semantic segmentation generates class labels for all vertices in the 3D reconstruction. We use SparseConvNet [21] trained on ScanNet, however, other methods and other training

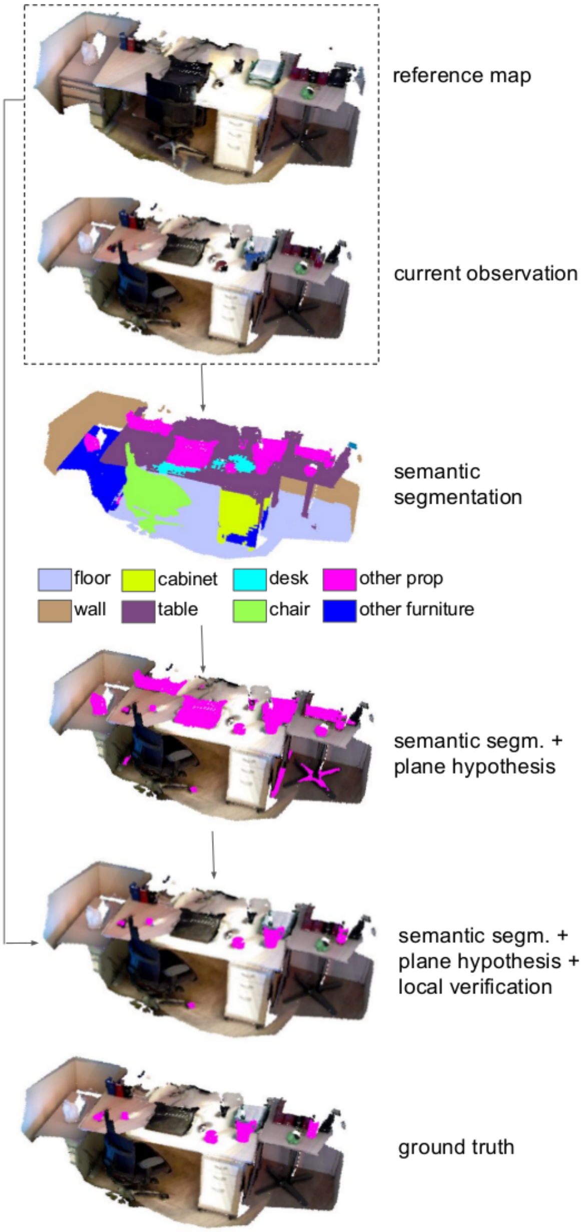

Fig. 2: Overview of our proposed object discovery method showing object detection results for each step. Detected objects are displayed in pink.

datasets could also be applied. Specific details of our implementation are given in Section V-A.2.

We identify objects or parts of objects by searching for protrusions on supporting planes, i.e. horizontal surfaces on which objects may lie. Instead of searching on all horizontal planes, the semantic information is leveraged by limiting the search to surfaces belonging to relevant classes[1]. The reconstruction vertices corresponding to these labels are clustered and for each cluster, horizontal planes are fit using RANSAC [22]. Considering only the semantically relevant subset of vertices not only reduces the number of points that need to be processed but also achieves more accurate plane estimates because there are fewer outliers. Candidate objects are found by segmenting the vertices that lie above the detected semantic planes using Euclidean clustering [12].

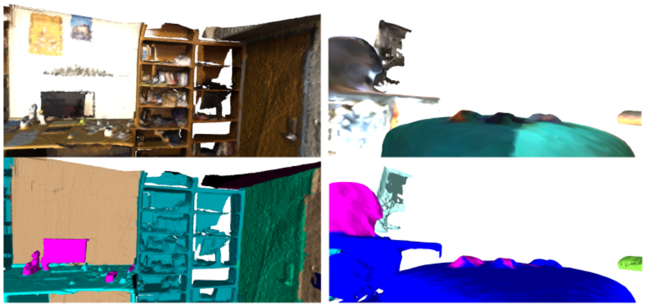

Fig. 3: Examples from ScanNet showing annotation inaccuracy. Small objects are incorrectly labeled, either under-segmented or not separated from the supporting structure. Top: original scan. Bottom: annotation with objects in pink.

Many semantic segmentation methods also predict the `otherprop` class (as defined in ScanNet), which is a general label typically associated to small items that do not belong to indoor structures. The vertices with the `otherprop` prediction could be used to directly identify objects, however, they are insufficient for discovering novel objects as we show in our evaluation (see Section V). The main reason is that the ground truth scenes in ScanNet are labeled on pre-segmented patches, therefore, trained models do not exhibit high precision around object boundaries. This is particularly problematic for small objects that either lack precise boundaries or are merged with the supporting structure (see Figure 3). Nonetheless, we also include all clusters of the `otherprop` class in the initial set of object candidates. These are then verified using local geometry as described next.

B. Object Verification with Local Geometry

Change detection is commonly used to identify novel objects in a scene as it requires no prior object information. Typically, a difference is computed between two spatially aligned observations. Points in the result set D are points from observation C that do not have a corresponding point in observation S within a distance d (adapted from [8]):

$$D = \{c | c \in C \wedge \nexists\; s \in S, \|c, s\| < d\}. \qquad (1)$$

This formulation is highly sensitive to the distance threshold used. Furthermore, the pervasiveness of noise from depth sensors and localization error reduce detection accuracy.

Our approach leverages the idea of change detection but applies the operation locally. Since object candidates are already generated from the full scene using semantic information, it is no longer necessary to perform global change detection. It is sufficient to apply the operation in local regions around the initial candidates. This is a two-stage process. First, for each detected object cluster, we

[1] Floor, cabinet, bed, chair, sofa, table, bookshelf, counter, desk, shelves, night stand, other structure, other furniture and other props.

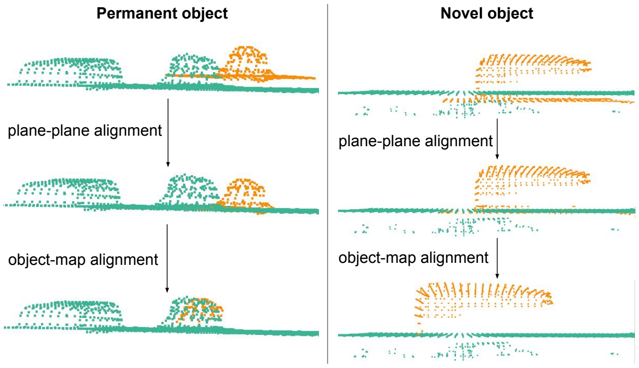

Fig. 4: The two steps of the local verification are shown for a permanent and a novel object. The reference crop is colored in turquoise, object and supporting plane in orange. Note in the novel object example, the two-step approach prevents ICP from aligning the object to the reference plane.

extract the supporting plane in its surrounding. It is aligned to the nearest horizontal plane in the reference map (plane-plane alignment) by applying the Iterative Closest Point (ICP) algorithm [23] with fixed rotation around the x- and y-axes. This initial step makes the verification of the actual object more efficient and robust since the second ICP step allows only transformations in the x- and y-direction and rotation around the z-axis to align the object cluster with the reference map (object-map alignment). For this operation, we use a crop of the reference map with some margin around the potential object location as input. Moreover, the plane points detected in the first step are removed to ensure better alignment. Figure 4 exemplifies the steps of local verification.

Given ICP convergence, we transform the object point cloud and use Eq. (1) to determine the object's overlap with the reference map. Objects that have a very small overlap are considered to be introduced into the scene. Details about the parameters used in the local verification stage are provided in Section V-A.3.

Applying local verification has advantages over global differencing. Firstly, it can adapt to subtle changes between the observations, thus accounting for moved objects that are not necessarily novel. Secondly, the inaccuracy of warped reconstructions can be absorbed. In contrast, differencing is anchored to the full scene, which means it is inflexible and requires precisely aligned observations.

In our formulation, permanent objects are identified if they only move within the dimension specified by the crop. We use a dimension of 20 cm to account for reasonable scene rearrangements. In reality, if a permanent object moves a greater distance, it is in fact out of place and should be detected. If a room is rearranged, a new reference map would need to be made.

IV. Dataset

To evaluate our proposed approach for novel object discovery we present a new robotic dataset.[2] The dataset consists of five diverse scenes: office, kitchen, living room, small room and a large room. The scenes range from partially viewed rooms (office, kitchen, living room) to complete rooms of different size (small and large room). The RGB and depth streams were recorded from the onboard RGBD sensor of a mobile robot that navigated through the scenes. The recorded data also includes the transformation matrices between coordinate frames and is used to generate reconstructions with [24]. For each environment, data from a reference setup was recorded in which no novel objects were present. Data from additional setups were recorded for each room consisting of 3–18 novel objects in various locations. Furniture and permanent objects such as decorations were also rearranged. In total we provide five different scenes with 31 observations (including the reference maps) and 260 annotated novel objects.

Objects introduced into the scenes were from the YCB dataset [25]. We selected objects of diverse size ranging from small such as a screwdriver to large such as a plastic water pitcher. The reconstructions were annotated by first aligning the 3D object models in the reconstruction and finding all points within a small distance threshold to the model points using a kd-tree. As a final step, single points were manually added to or removed from the object masks. This is necessary because most objects are usually not reconstructed precisely.

The characteristics of our dataset are summarized and compared to other datasets applicable for object discovery in Table I. The publicly available datasets are those in [5], [13], [10] and [16]. The dataset of [5] is captured over a long period of time using a mobile robot. Despite the large amount of data, only a single environment is considered. Also, objects are not annotated so it cannot be used for quantitative evaluation. The dataset of [13] consists of data from a robot in a regular indoor environment, so it includes a large variety of objects and even people. However, the dataset only considers one room and the base of the robot was not moved during acquisition. Only the RGB-D sensor was rotated on a pan/tilt unit. As a result, the room is always viewed from the same perspective. Unfortunately, no reference map is provided, therefore, the distinction between novel and permanent objects is unclear. Some specially selected objects are annotated as new objects while others that are physically new in a scene compared to a previous one are not. The dataset of [10] provides raw recordings from a handheld Google Tango for three different rooms. The main focus of the work was not on object detection but static room recovery. Therefore, the dataset mainly contains furniture that moved between successive observations. It does not consider small items. It also does not provide annotations of the objects. The recent dataset of [16] also provides recordings from a handheld Google Tango. While a massive number

[2] https://www.acin.tuwien.ac.at/object-change-detection-dataset-of-indoor-environments/

TABLE I: Comparison of object discovery and change detection datasets. Columns for novel objects, small objects and furniture indicate if these categories change between scenes.

	Scenes	Observations	Robot	Reference map	Novel objects	Small objects	Furniture	Annotated	Available
Finman et al. [7]	2	67	–	–	✓	–	–	?	–
Langer et al. [11]	1	4	✓	✓	✓	–	–	–	–
Katsura et al. [14]	2	10+?	✓	✓	✓	–	–	?	–
Herbst et al. [6]	4	24	–	–	✓	–	–	✓	–[3]
Mason et al. [5]	1	67	✓	–	✓	✓	✓	–	✓[4]
Ambrus et al. [13]	1	88	✓	–	✓	✓	✓	–[5]	✓
Fehr et al. [10]	3	23	–	✓	–	–	✓	–	✓
Wald et al. [16]	478	1482	–	–	?	–	✓	✓	✓
Ours	5	31	✓	✓	✓	✓	✓	✓	✓

of observations are captured, mainly large items such as furniture are changed in the setups. Different to [10], all instances are segmented, however, the annotation tool from ScanNet [20] is used, which means that small items are inaccurately labeled due to the pre-segmentation step.

V. Experimental Results

This section presents experimental results with our collected dataset. We first describe our implementation, in particular, for reconstruction and semantic segmentation. We then describe the comparison methods and outline the metrics used for evaluation. Lastly, we present quantitative and qualitative results as well as show the generality of our approach by applying it to different reconstruction methods.

A. Implementation Details

1) Reconstruction: The *Voxblox* framework [24] is used to generate reconstructions. It was initially developed for planning purposes, but is shown to be suitable for other robotic applications, such as incremental scene segmentation [26] and for extracting 3D object models [27]. Voxblox creates dense 3D maps based on the TSDF representation. We use the robot pose from the recordings instead of using camera tracking. However, the option to refine the pose by aligning the input data to the existing structure with ICP is employed. The Voxblox framework is very suitable for robotic applications not only because the pose of the robot can be used, but also because it runs directly on a CPU. Details about the parameters are given in Table II.

2) Semantic Segmentation of Dense 3D Maps: SparseConvNet [21] is used to perform semantic segmentation on the full 3D reconstructions. It accepts a set of colored points as input. The output from Voxblox is converted to this format by taking the centroid and average color of each voxel in the reconstruction. The network is trained on the ScanNet [20] dataset with the standard test, validation and training splits. The annotations use the second version of the dataset. Data is augmented using the provided tools from SparseConvNet. To train the model we used the default settings of SparseConvNet except for the following parameters: m=32, residual_blocks=True, scale=50, block_reps=2, batch_size=5. In addition to the 20 classes in the ScanNet benchmark, we included `otherprop` to have a total of 21 classes.

TABLE II: Parameters used for the experiments.

Method	Parameter	Value	Method	Parameter	Value
Voxblox	resolution [m]	0.01	LV	reference crop margin [m]	0.2
Voxblox	method	simple	LV	current crop margin [m]	0.05
Voxblox	max. ray length [m]	2.0	LV	ICP max. dist. plane [m]	0.05
RANSAC	dist. threshold [m]	0.01	LV	ICP max. dist. object [m]	0.15
RANSAC	angle threshold [deg]	5.0	LV	max. diff. dist. [m]	0.014
Clustering	dist. threshold [m]	0.02	LV	min. rejection overlap	0.7

3) Parameters: Table II lists all parameters used in our implementation. This comprises the parameters for creating the reconstructions with Voxblox, for detecting objects above the semantic planes using RANSAC and the object clustering threshold. The various parameters in the local verification (LV) stage are also given.

B. Comparison Methods

We select two baseline methods to compare our method against. Both perform scene differencing to detect dynamic objects. The method proposed by Ambrus et al. [8] (*Meta-room*) creates a reference map from several observations. This point-based volumetric representation is called meta-room and is further used for change detection. After aligning an observation to the meta-room, the difference between them is computed. The remaining points are clustered, then planar and small clusters are removed. Because the focus of our paper is object change detection, we use our object-free reference map created with Voxblox as the meta-room when evaluating the change detection scheme in [8]. To ensure a fair comparison, we adapted the original parameters to the characteristics of our dataset, which achieved better results.

As another baseline, we compare against the method in [11] that uses an octomap [28] (*Octomap*) for representation and differencing. The effect of noisy sensor input is slightly reduced due to the quantization of the points into voxels and also the dilation applied to the reference map. We modified the method to detect all objects, not only those on the floor as in the original approach.

In addition to the two baseline methods, we give results for the important modules of our proposed approach. We use the direct output of semantic segmentation on the object level by considering all vertices labeled with the `otherprop` class (*Semantic segm.*). *Ours (no planes)* states the result when applying semantic segmentation and locally verify the object candidates, but without including possible objects from supporting planes. We also evaluate our proposed method using only object detection from global semantic context and without the verification stage of ICP alignment and local differencing (*Ours (no LV)*) to demonstrate the improvements of the final stage of our pipeline (*Ours (full)*).

Note that our method without local verification (as well as semantic segmentation) can only propose potential objects,

[3] The URL provided in the publication no longer works.

[4] Rosbags available on request.

[5] Only partially and inconsistently annotated.

TABLE III: Comparison of different methods on the robotic dataset.
(Pr = precision, Re = recall, F1 = F1-score, M = missed objects, W = wrongly detected objects)

	Small Room					**Big Room**				
	Pr	Re	F1	M	W	Pr	Re	F1	M	W
Octomap [11]	0.11±0.05	0.61±0.18	0.19±0.08	15	176	0.07±0.04	0.42±0.15	0.12±0.07	42	434
Meta-room [8]	0.04±0.03	0.39±0.13	0.07±0.05	29	260	0.25±0.30	0.55±0.05	0.25±0.28	32	465
Semantic segm.	0.15±0.13	0.39±0.32	0.22±0.18	32	81	0.48±0.10	0.59±0.09	0.52±0.05	41	103
Ours (no planes)	0.30±0.25	0.39±0.31	0.34±0.27	42	**35**	**0.73±0.05**	0.58±0.09	**0.65±0.07**	41	**54**
Ours (no LV)	0.23±0.07	**0.72±0.11**	0.35±0.09	**5**	204	0.29±0.05	**0.74±0.05**	0.41±0.05	**7**	488
Ours (full)	**0.48±0.222**	0.70±0.13	**0.56±0.17**	8	84	0.59±0.16	0.73±0.04	**0.65±0.12**	8	173
	Living Room (partial)					**Office (partial)**				
	Pr	Re	F1	M	W	Pr	Re	F1	M	W
Octomap [11]	0.11±0.08	0.50±0.08	0.17±0.10	19	74	0.18±0.07	0.77±0.13	0.28±0.10	8	73
Meta-room [8]	0.13±0.19	0.45±0.12	0.13±0.15	13	115	0.15±0.20	0.41±0.21	0.16±0.18	12	149
Semantic segm.	0.62±0.27	0.39±0.08	0.45±0.13	24	29	0.18±0.08	0.62±0.04	0.27±0.10	8	61
Ours (no planes)	**0.95±0.05**	0.38±0.08	0.54±0.09	17	**25**	**0.71±0.35**	0.60±0.03	0.60±0.19	12	**10**
Ours (no LV)	0.45±0.16	**0.63±0.07**	0.51±0.11	**5**	81	0.16±0.06	**0.81±0.03**	0.26±0.08	**0**	148
Ours (full)	0.75±0.23	0.61±0.07	**0.65±0.13**	8	37	0.61±0.28	0.76±0.06	**0.63±0.17**	2	28
	Kitchen (partial)					**Average**				
	Pr	Re	F1	M	W	Pr	Re	F1	M	W
Octomap [11]	0.43±0.08	0.41±0.08	0.41±0.07	9	**40**	0.18±0.14	0.54±0.18	0.23±0.13	18.6	159.4
Meta-room [8]	0.66±0.17	0.35±0.12	0.46±0.14	10	58	0.25±0.28	0.43±0.14	0.22±0.21	19.2	209.4
Semantic segm.	0.30±0.08	0.73±0.20	0.40±0.09	6	143	0.35±0.22	0.55±0.21	0.38±0.15	22.2	83.4
Ours (no planes)	0.67±0.16	0.68±0.18	0.66±0.12	7	58	**0.67±0.28**	0.53±0.19	0.56±0.16	23.8	**36.4**
Ours (no LV)	0.27±0.06	**0.75±0.16**	0.39±0.07	**1**	149	0.28±0.12	**0.73±0.10**	0.39±0.11	**3.6**	214.0
Ours (full)	**0.75±0.18**	0.68±0.16	**0.69±0.09**	5	58	0.63±0.22	0.70±0.10	**0.64±0.13**	6.2	76.0

but can not exclude objects because they do not incorporate knowledge from a reference map.

C. Metrics

A number of different metrics are considered for the evaluation. The commonly used metrics of precision and recall are applied at the point level. These measure the accuracy of the object detections by considering all detected points in the scene and all points from the ground truth annotation. Precision measures the proportion of detected points that correspond to the ground truth ($TP/(TP+FP)$) and recall measures the proportion of ground truth points that are in the detection set ($TP/(TP+FN)$). The F1-score is also reported as it provides the harmonic mean of the two quantities.

Since we are concerned about detection performance of objects, we also report two additional metrics. We measure the number of missed objects by comparing the overlap of the clustered detections with the ground truth objects. If no point of a ground truth object is detected then it is considered missing. To measure overestimation (i.e. false positives), we sum the number of detected clusters that do not overlap with a ground truth object for each setup. While this is not an accurate measure for false positive detections, it allows an additional comparison of approaches at the object level.

D. Results

Table III shows the performance of the evaluated methods for each room in the robotic dataset. The results are averaged over the different setups for precision, recall and F1-measure. We also provide the standard deviation for these three metrics. Missed objects and wrong detections objects are summed. The total average for all rooms is also given. Qualitative results for some example scenes using our method are shown in Figure 5 and further examples are given in the supplementary video.

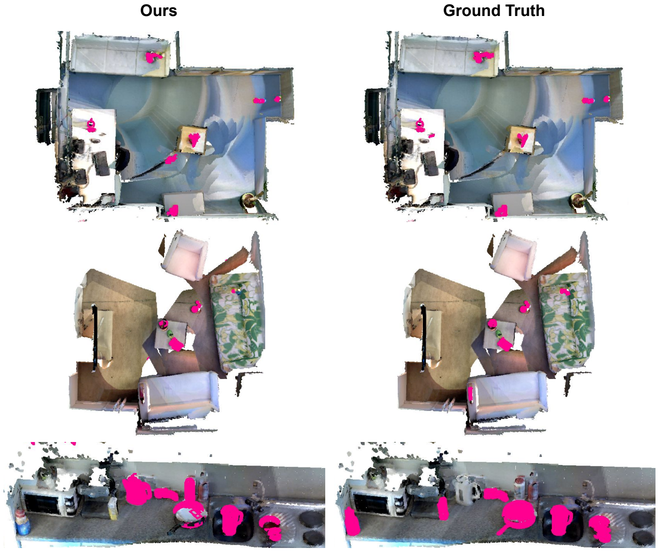

Fig. 5: Qualitative examples from three scenes. Detected/-ground truth objects colored in pink. Top: small room, middle: living room, bottom: kitchen.

The quantitative results show that our approach clearly outperforms the baseline methods. Precision increases drastically for the methods that apply local verification. This shows the benefit of performing differencing only in restricted local areas. Both *Meta-room* and *Octomap* report lower recall than our approach because of their post-filtering step, which is used to address the limitations of global differencing.

Semantic segmentation (*Semantic segm.*) performs surprisingly well in terms of recall, given the fact it was trained on a completely different dataset. However, the high

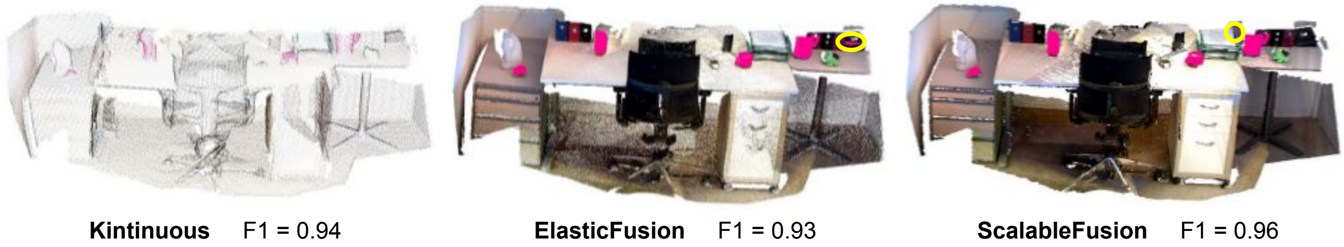

Fig. 6: Qualitative examples of our approach applied on the same scene for different reconstruction methods. Detected objects are visualized in pink, wrong detections are highlighted with a yellow ellipse. All reconstructions are displayed with original point size.

number of missed objects indicates that especially smaller objects can not be detected. Comparing the object candidates proposed by semantic segmentation to the reference map by applying our local verification step (*Ours (no planes)*) results in higher precision and it decreases the number of wrong objects. However, objects are still missing and the recall remains low. This is improved by exploiting the semantic information to explore horizontal planes (*Ours (no LV)*) and find small objects. However, without utilizing the knowledge from a reference map the number of wrongly detected objects explodes. Incorporating both proposed steps, semantic plane detection and local verification (*Ours (full)*), results in the best trade-off. The full method achieved both good precision as well as recall and therefore the highest F1-score overall. On average, our method identifies 97.6 % of all novel objects in the dataset while detecting only 3 false positives per scene. This shows the importance of combining global and local procedures for accurate novel object detection.

The most common failure case of our method occurs when spatial clustering is not able to separate new and permanent objects that are touching. An example of this can be seen in Figure 5 at the bottom where the yellow-white sugar box is clustered together with the coffee machine. During local verification, the whole cluster is removed because the alignment of the coffee machine leads to a high overlap. In order to deal with this case, a segmentation algorithm could be applied to separate objects. This is left for future work.

E. Generality to Different Reconstruction Methods

Our approach is applied to the output of different reconstructions to show its generality. We consider Kintinuous [29], ElasticFusion [30] and ScalableFusion [31]. Reconstructions of the reference map as well as observations are generated using the default implementations of each mentioned method. The outputs of these methods are converted to a point set by taking the centroids (and average color values) of the voxels [29] or surfels [30] or taking the mesh vertices from [31]. For ElasticFusion and ScalableFusion, the differencing threshold is reduced by half to 0.007 m (d in Eq. (1)) due to the higher point density.

Figure 6 shows example outputs from the reconstruction methods for a sample of the office room. The setup includes four new objects to be detected. All mentioned reconstruction methods are able to produce reasonable results, achieving high F1-scores. In all three cases, every novel object was identified. ElasticFusion and ScalableFusion detected one very small wrong object each.

VI. Conclusion

This work addressed the problem of detecting novel objects in 3D environments. We presented an approach that analyzes the scene globally and detects objects using semantic information. The semantic context is exploited to extract objects on horizontal planes from relevant structures. The detected objects are then verified by performing change detection with a reference map in a local region. Results with our new dataset show that our combined approach outperforms existing baselines.

Our approach correctly detects new objects while excluding moved furniture and decoration. It is even possible to detect only partially visible objects and also cluttered novel objects. Our approach, however, sometimes rejects small objects in the course of local verification because they can be aligned to the reference map with high overlap. For future work, this could be improved by using descriptive object features. This would also help to detect novel objects that are placed in locations that were previously occupied by permanent objects in the reference map. Another direction for future work is to incorporate more context in the form of an ontology. This would allow high-level semantic reasoning in the verification process. Finally, we expect our approach to be highly valuable for surveillance applications and we plan to investigate missing object detection in the future.

VII. ACKNOWLEDGEMENTS

Thanks to Christian Eder and Kevin Wolfram for their help in annotating the dataset.

References

[1] F. Tombari, S. Salti, and L. Di Stefano, "Unique signatures of histograms for local surface description," in *Proc. of ECCV*, 2010, pp. 356–369.

[2] B. Drost, M. Ulrich, N. Navab, and S. Ilic, "Model globally, match locally: Efficient and robust 3D object recognition," in *Proc. of IEEE CVPR*, 2010, pp. 998–1005.

[3] T. Fäulhammer, M. Zillich, J. Prankl, and M. Vincze, "A multi-modal rgb-d object recognizer," in *IEEE International Conference on Pattern Recognition (ICPR)*, 2016, pp. 733–738.

[4] P. Alimi, D. Meger, and J. J. Little, "Object persistence in 3D for home robots," in *Proc. of ICRA Workshop on Semantic Perception, Mapping and Exploration*, 2012.

[5] J. Mason and B. Marthi, "An object-based semantic world model for long-term change detection and semantic querying," in *Proc. of IEEE/RSJ IROS*, 2012, pp. 3851–3858.

[6] E. Herbst and D. Fox, "Toward object discovery and modeling via 3-D scene comparison," in *Proc. of IEEE ICRA*, 2011, pp. 2623–2629.

[7] R. Finman, T. Whelan, M. Kaess, and J. J. Leonard, "Toward lifelong object segmentation from change detection in dense RGB-D maps," in *Proc. of ECMR*, 2013, pp. 178–185.

[8] R. Ambrus, N. Bore, J. Folkesson, and P. Jensfelt, "Meta-rooms: Building and maintaining long term spatial models in a dynamic world," in *Proc. of IEEE/RSJ IROS*, 2014, pp. 1854–1861.

[9] E. Herbst, P. Henry, and D. Fox, "Toward online 3-D object segmentation and mapping," in *Proc. of IEEE ICRA*, 2014, pp. 3193–3200.

[10] M. Fehr, F. Furrer, I. Dryanovski, J. Sturm, I. Gilitschenski, R. Siegwart, and C. Cadena, "TSDF-based change detection for consistent long-term dense reconstruction and dynamic object discovery," in *Proc. of IEEE ICRA*, 2017, pp. 5237–5244.

[11] E. Langer, B. Ridder, M. Cashmore, D. Magazzeni, M. Zillich, and M. Vincze, "On-the-fly detection of novel objects in indoor environments," in *Proc. of IEEE ROBIO*, 2017, pp. 900–907.

[12] R. B. Rusu, N. Blodow, Z. C. Marton, and M. Beetz, "Close-range scene segmentation and reconstruction of 3D point cloud maps for mobile manipulation in domestic environments," in *Proc. of IEEE/RSJ IROS*, 2009, pp. 1–6.

[13] R. Ambrus, J. Folkesson, and P. Jensfelt, "Unsupervised object segmentation through change detection in a long term autonomy scenario," in *Proc. of IEEE-RAS HUMANOIDS*, 2016, pp. 1181–1187.

[14] U. Katsura, K. Matsumoto, A. Kawamura, T. Ishigami, T. Okada, and R. Kurazume, "Spatial change detection using voxel classification by normal distributions transform," in *Proc. of IEEE ICRA*, 2019, pp. 2953–2959.

[15] A. Zeng, S. Song, M. Nießner, M. Fisher, J. Xiao, and T. Funkhouser, "3DMatch: Learning local geometric descriptors from RGB-D reconstructions," in *Proc. of IEEE CVPR*, 2017.

[16] J. Wald, A. Avetisyan, N. Navab, F. Tombari, and M. Nießner, "RIO: 3D object instance re-localization in changing indoor environments," in *Proc. of IEEE ICCV*, 2019, pp. 7658–7667.

[17] J. Vidal, C.-Y. Lin, X. Lladó, and R. Marti, "A method for 6D pose estimation of free-form rigid objects using point pair features on range data," *Sensors*, vol. 18, no. 8, 2018.

[18] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proc. of IEEE CVPR*, 2017, pp. 2881–2890.

[19] A. Valada, R. Mohan, and W. Burgard, "Self-supervised model adaptation for multimodal semantic segmentation," *International Journal of Computer Vision*, 2019.

[20] A. Dai, A. X. Chang, M. Savva, M. Halber, T. Funkhouser, and M. Nießner, "Scannet: Richly-annotated 3D reconstructions of indoor scenes," in *Proc. of IEEE CVPR*, 2017, pp. 5828–5839.

[21] B. Graham, M. Engelcke, and L. van der Maaten, "3D semantic segmentation with submanifold sparse convolutional networks," in *Proc. of IEEE CVPR*, 2018, pp. 9224–9232.

[22] M. A. Fischler and R. C. Bolles, "Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography," *Communications of the ACM*, vol. 24, no. 6, pp. 381–395, 1981.

[23] P. J. Besl and N. D. McKay, "A method for registration of 3-D shapes," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 14, no. 2, pp. 239–256, 1992.

[24] H. Oleynikova, Z. Taylor, M. Fehr, R. Siegwart, and J. Nieto, "Voxblox: Incremental 3D Euclidean signed distance fields for on-board MAV planning," in *Proc. of IEEE/RSJ IROS*, 2017, pp. 1366–1373.

[25] B. Calli, A. Singh, J. Bruce, A. Walsman, K. Konolige, S. Srinivasa, P. Abbeel, and A. M. Dollar, "Yale-CMU-Berkeley dataset for robotic manipulation research," *The International Journal of Robotics Research*, vol. 36, no. 3, pp. 261–268, 2017.

[26] M. Grinvald, F. Furrer, T. Novkovic, J. J. Chung, C. Cadena, R. Siegwart, and J. Nieto, "Volumetric instance-aware semantic mapping and 3d object discovery," *IEEE Robotics and Automation Letters*, vol. 4, no. 3, pp. 3037–3044, 2019.

[27] F. Furrer, T. Novkovic, M. Fehr, A. Gawel, M. Grinvald, T. Sattler, R. Siegwart, and J. Nieto, "Incremental object database: Building 3d models from multiple partial observations," in *Proc. of IEEE IROS*, 2018, pp. 6835–6842.

[28] A. Hornung, K. M. Wurm, M. Bennewitz, C. Stachniss, and W. Burgard, "OctoMap: An efficient probabilistic 3D mapping framework based on octrees," *Autonomous Robots*, vol. 34, no. 3, pp. 189–206, 2013.

[29] T. Whelan, M. Kaess, M. Fallon, H. Johannsson, J. J. Leonard, and J. McDonald, "Kintinuous: Spatially extended Kinectfusion," in *Proc. of RSS Workshop on RGB-D: Advanced Reasoning with Depth Cameras*, 2012.

[30] T. Whelan, S. Leutenegger, R. Salas-Moreno, B. Glocker, and A. Davison, "ElasticFusion: Dense SLAM without a pose graph," in *Proc. of RSS*, 2015.

[31] S. Schreiberhuber, J. Prankl, T. Patten, and M. Vincze, "ScalableFusion: High-resolution mesh-based real-time 3D reconstruction," in *Proc. of IEEE ICRA*, 2019, pp. 140–146.