2023 IEEE International Conference on Robotics and Automation (ICRA 2023)
May 29 - June 2, 2023. London, UK

3D VSG: Long-term Semantic Scene Change Prediction through 3D Variable Scene Graphs

Samuel Looper[1], Javier Rodriguez-Puigvert[2], Roland Siegwart[1], Cesar Cadena[1], and Lukas Schmid[1,3]

***Abstract*—Numerous applications require robots to operate in environments shared with other agents, such as humans or other robots. However, such shared scenes are typically subject to different kinds of long-term semantic scene changes. The ability to model and predict such changes is thus crucial for robot autonomy. In this work, we formalize the task of *semantic scene variability estimation* and identify three main varieties of semantic scene change: changes in the position of an object, its semantic state, or the composition of a scene as a whole. To represent this variability, we propose the Variable Scene Graph (VSG), which augments existing 3D Scene Graph (SG) representations with the variability attribute, representing the likelihood of discrete long-term change events. We present a novel method, *DeltaVSG*, to estimate the variability of VSGs in a supervised fashion. We evaluate our method on the 3RScan long-term dataset, showing notable improvements in this novel task over existing approaches. Our method DeltaVSG achieves an accuracy of 77.1% and a recall of 72.3%, often mimicking human intuition about how indoor scenes change over time. We further show the utility of VSG prediction in the task of active robotic change detection, speeding up task completion by 66.0% compared to a scene-change-unaware planner. We make our code available as open-source.**

I. Introduction

Mobile robotics have the potential to impact numerous applications in healthcare, home robotics, service robotics, or delivery. These tasks require robots to operate in complex indoor environments that are shared with other agents, such as humans. However, shared environments are frequently subject to long-term scene changes, as agents interacting with the scene often cause the position of objects or other semantic scene attributes to change. This poses a major challenge, as most current methods assume that scenes are static. Thus, the capacity to capture, model, and predict such changes is essential to enable efficient operation in shared environments.

Current approaches for semantic scene understanding with autonomous robots typically rely on 3D reconstruction of the environment, where each surface element is labeled with the semantic class [1,2] or object instance [3]–[5]. However, scenes often change in a semantically consistent way, at the level of objects rather than individual surface elements [5]. Humans have an intuition about how our surroundings may change over time. On the one hand, it is unlikely that some objects, such as the fridge and coffee machine, will completely disappear. On the other hand, handheld objects such as mugs can be moved, and typically belong near objects such as tables or drying racks. Our understanding of how objects change depends significantly on local scene context and relationships between objects. Recently, 3D Scene Graphs (SG) [6]–[8] have emerged as a compact representation that encodes semantic, geometric, and relationship information.

Scene changes can be classified into two categories: *short-term* dynamics occur within view of the sensor, such as people moving, and *long-term* scene changes, denoting changes beyond the current view of a robot. This results in notably different change characteristics. While short-term changes are typically continuous time-series, long-term changes are characterized by discrete and abrupt changes between two observations. While some SGs model short-term dynamics [8], scene semantics are typically assumed static.

To model and predict such long-term changes, we propose the concept of Variable Scene Graphs (VSG). VSGs are an extension of traditional SGs, which additionally model the likelihood of changes occurring for individual objects, which we call *variability*. While the presented formalism is general and could also account for short-term changes by placing a high variability on moving objects, we focus on the modeling of discrete long-term changes in this work. In particular, we formalize *semantic scene variability estimation*, the task of estimating variability for all objects in a scene, to compose a VSG. We present DeltaVSG, a novel method to address this task using graph-based learning. The resulting VSG can predict which parts of a scene are likely to change, allowing robots repeatedly operating in the same environment to harness this predictive power for informed planning based on their previous map. We demonstrate the utility of VSGs in the task of robotic active change detection.

We make the following contributions:

- We propose 3D Variable Scene Graphs (VSG), a novel formulation to model long-term dynamic scenes and predict scene changes in SGs.
- We propose DeltaVSG, a method to estimate VSGs, i.e. long-term scene variability, from existing SGs.
- We extensively evaluate our method on real world data, showing that accurate VSGs can be generated from standard SGs and demonstrating the potential of VSGs in the task of robotic active change detection. We make our code available as open-source[1].

This project has received funding from the Microsoft Swiss Joint Research Center, the European Union's Horizon 2020 research and innovation programme under grant agreement No 101017008 (Harmony), and the Swiss National Science Foundation (SNSF).

[1] Autonomous Systems Lab, ETH Zürich, Zürich, Switzerland. {slooper, rsiegwart, cesarc, schmluk}@ethz.ch
[2] Universidad de Zaragoza, Zaragoza, Spain. jrp@unizar.es
[3] Massachusetts Institute of Technology, USA. lschmid@mit.edu

[1]Released at https://github.com/ethz-asl/3d_vsg.

II. Related Work

A. 3D Semantic Scene Representations

Understanding human-made environments is a central topic in robotics. A first level of semantic understanding can be gained by detecting and classifying object classes or instances in 3D scene reconstructions [1]–[5,9]. This has been shown to improve mapping [10] and task planning [11] for mobile robots. Semantic understanding can be further improved by learning object-wise semantic or geometric features [12], relationships between objects [13], or scene-specific object attributes [14]. Several models have been proposed to infer support relationships [15], human-scene interactions [16], and the likelihood of robot co-occurrence [17]. This enhanced semantic understanding has been shown to help in robotic tasks such as object re-localization [18].

Recently, Scene Graphs (SG) have emerged as a powerful abstract representation of indoor scene semantics. Initially developed for image understanding [19], they have been extended to 3D and combine various forms of object-level data, as first proposed by Armeni *et al.* [6]. Wald *et al.* [7] extend SGs with rich semantic attributes, affordances, and relationships. Rosinol *et al.* [8] introduce dynamic SGs, providing rich hierarchical abstractions of the SG and accounting for short-term moving agents, such as humans. Recently, Giuliari *et al.* [20] presented Spatial Commonsense Graphs (SCG), embedding additional nodes from knowledge graphs such as "used for reading", which they call *commonsense concepts*. Different approaches have been proposed to estimate SGs from images [21,22] or dense reconstructions [7,8]. Recently, methods to incrementally build SGs have been proposed, such as SGFusion [23] or Hydra [24], enabling application in online robotics or interaction [25]. Such semantically rich SGs have shown to improve numerous applications, ranging from task planning [26,27], object retrieval [28,29] or synthetic scene generation [30,31]. However, these methods assume that scene semantics are static and do not yet account for long-term changes. Our proposed VSGs thus represent a novel extension of SGs to account for semantic scene variability in long-term changing scenes.

B. Changes in Scene Semantics

Semantic scene changes have primarily been studied in the application of change detection. A first family of approaches focuses on detecting changes in images [32]–[34]. However, these methods rarely reason about semantic changes beyond highlighting differences in image maps. As a step towards this, several methods use natural language to classify and caption changes [35]. Ru *et al.* [36] classify semantic changes between two 2D images using an attention-based fusion component, while Kim *et al.* [37] classify scene change by operating on scene graphs. While they leverage scene graphs and/or semantics in a similar fashion to our work, they don't capture this knowledge on changes in semantic graphs explicitly to make future predictions. These last two methods also don't consider 3D geometry, as they work in a context where geometry can be easily inferred from 2D satellite images. Generating semantic-level change annotations is a significant bottleneck for these methods. This can be mitigated using simulated data, as in [38].

3D spatial information can provide useful geometrical context to semantic data. Qiu *et al.* use multiple camera views to improve semantic change captioning for 3D scenes [39], while Ku *et al.* [40] and Li *et al.* [41] directly use depth data for image-based change detection. Volumetric representations provide even greater context, allowing scene differencing for 3D change detection [42,43]. Such volumetric methods have been extended to utilize semantics for more complete change detection [44], and even during online operation [5]. Similar to our work, Liao *et al.* [45] demonstrate how semantic scene graphs can improve change captioning in simple 3D scenes.

However, while change detection involves highlighting semantic differences between a reference and target scene, change prediction is an altogether different task, requiring models to predict the likelihood of future change from a single existing reference scene. The ability to predict rather than just react is essential for efficient robot operation in changing environments. However, this has typically been addressed by predicting change frequencies on individual voxels [46,47]. While this can already improve planning and localization in changing scenes, such models lack semantic context and consistency provided by SGs. To this end, Rosinol *et al.* [8] recently proposed dynamic SGs considering moving agents, whereas Casas *et al.* [48] further build a SG to predict the movement of agents in a driving scene. However, while this allows for modeling of various short-term dynamic agents like humans and robots in a scene [49], they do not yet consider changes to the scene or objects. In contrast, our work focuses on predicting long-term semantic scene changes in both location, semantic attributes, and topology of the individual objects constituting an indoor scene.

III. 3D Variable Scene Graphs

Environments shared with other agents inevitably change over time. The ability to model and predict such changes is thus essential for long-term autonomy in these environments. We therefore define *Variable* SGs (VSG) as a representation to capture and predict changing environments, and define the task of *semantic scene variability estimation*.

Building on the definition of SGs from Wald *et al.* [7], we define a VSG $\mathcal{G}$ as a set of vertices $\mathcal{V}$ and a set of edges $\mathcal{E}$, where each entry $v_i \in \mathcal{V}$ represents an object instance in the scene. The semantic information of an instance is represented by the object's semantic class $o_i \in O$, and by a set of attributes $a_i \subseteq A$. The set of possible attributes A includes static properties (e.g. color, rigidity) that are unlikely to change, dynamic properties (e.g. full, open/closed, on/off) that can change as result of interaction, which are called *states*, and the possible interactions an object permits (e.g. sitting, opening) which are called affordances. Semantic relationships between objects are defined by the set $\mathcal{R}$ and a directional edge mapping $e(v_i, v_j) : (\mathcal{V} \times \mathcal{V}) \mapsto \mathcal{R}$. Such relationships include support relationships (e.g. standing, lying on), proximity relationships (e.g. next to, in front of),

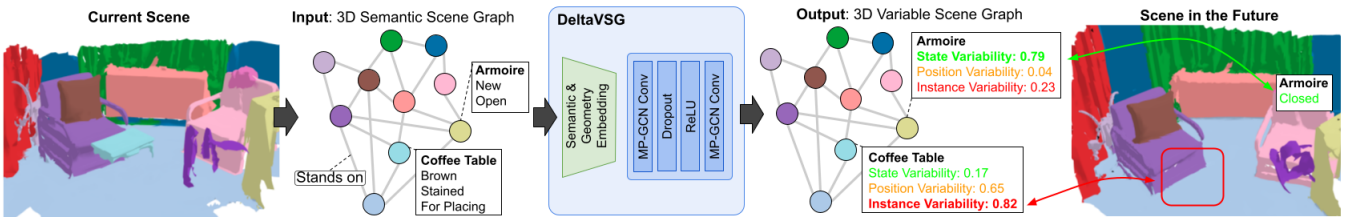

Fig. 1: Given a SG of objects $v_i^t \in \mathcal{V}^t$, attributes $a_i^t \in A^t$, and relations $\mathcal{E}$ of the current scene (left), DeltaVSG estimates a Variable Scene Graph (VSG) by predicting the variability y_i attributes for of each object (center). The VSG can be employed by a robot to predict semantic scene changes for long-term operation (right), or back-propagation of the variability if the future scene is known for training.

and comparative relationships (e.g. darker than, same shape). As scene geometry is an important cue, we additionally include the explicit relative position between two objects as a vector-valued relationship: $r(e_i) : \mathcal{E} \mapsto \mathbb{R}^3$.

To model changes, we define *variability* as the likelihood of a semantic scene change occurring before the next measurement. Since scenes typically change in a semantically consistent way, i.e. an entire object is changed rather than an independent piece of space [5], the variability y_i can be integrated into the VSG as an additional attribute of each vertex v_i. We identify three major ways a scene can change:

- *Position* variability y_P denotes objects moving in space (beyond a minimum threshold), i.e. $r(v_i^t, v_i^{t+1}) \geq \epsilon$.
- *State* variability y_S models changes in object states, such as a door changing from open to closed, i.e. $a_i^t \neq a_i^{t+1}$.
- *Instance* variability y_I denotes topological changes to the graph, i.e. $v_i^t \notin \mathcal{V}^{t+1} \vee v_i^{t+1} \notin \mathcal{V}^t$.

While this formulation could also account for transient short-term trajectories by assigning high variability, we focus on long-term changes in this work. Such long-term changes can result from interactions with the scene outside the current view of the robot, and are thus characterized by abrupt, discrete changes between two measurements. The task of long-term semantic scene variability estimation can thus be summarized as: estimate $y_i = \{y_P, y_S, y_I\}_i \ \forall n_i \in \mathcal{N}$. This is a challenging task, as many variations over time are in principle possible, whereas only a single realization occurs. Lastly, due to our definition of variability, the VSG can directly be employed by a robot for long-term operation, as its attributes are the change prediction.

IV. Approach

This section details our method, *DeltaVSG*, to predict long-term scene variability and generate VSGs. An overview of our approach is shown in Fig. 1. We discuss scene embeddings in Sec. IV-A, followed by the network architecture in Sec. IV-B and the data processing pipeline in Sec. IV-C.

A. Embedding Semantics and Geometry in Scene Graphs

Since semantic changes typically occur on the level of objects, we assume the abstract but rich representation of a SG has sufficient information for long-term scene variability prediction. In addition, since long-term changes typically occur abruptly, we assume a Markovian world and only consider the current state as input. Our method thus operates on a SG $\mathcal{G}$ as defined in Sec. III of the current scene to predict the variability y. This allows application of our approach directly on existing SGs, augmenting them to a VSG. Therefore, we design an embedding function that seeks to embed an input graph $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$ containing $N_v = |\mathcal{V}|$ objects and $N_e = |\mathcal{E}|$ edges into a *node embedding matrix* $M_v \in \mathbb{R}^{V_n \times d_v}$, where d_v is the dimension of node embedding vectors, an *edge relationship embedding matrix* $M_R \in \mathbb{R}^{N_e \times N_v}$, and an *edge indices matrix* $M_E \in \mathbb{N}^{N_e \times 2}$, whose rows $e_k = [i, j] \ \forall (i, j) \in \mathcal{E}$.

For node embeddings, semantic information must be compactly represented for neural network inference. From the input graph, this is done in a binary encoding vector over a set of discrete features, where each node is assigned a vector $u_i \in [0, 1]^{|A|}$ with each j^{th} element $u_i^{(j)}$ corresponds to an attribute in the defined taxonomy. Without loss of generality, we adopt the classification by Wald *et al.* [7] for this work, which proposes a taxonomy of 92 discrete object semantic attributes and 41 relationship attributes, as well as a hierarchical taxonomy of objects with 527 low-level classes. However, any such complex taxonomy results in embedding vectors that are extremely sparse, since objects only belong to a single class and have only a small number of attributes. We therefore apply principle component analysis (PCA) to reduce the dimensionality of our semantic embedding space. We empirically find that using $d_v = 120$, the original 619-dimensional binary embedding vectors u_i retain 90% of their information at up to $\sim 5\times$ compression.

Geometry and semantic relationships are combined to create the *edge relationships embedding matrix* R. For any two objects $v_i, v_j \in \mathcal{V}$, an embedding vector $q_{i,j} \in \mathbb{R}^{|\mathcal{R}|+3}$ is created by concatenating a binary encoding of any semantic relationship between the objects, and the relative position ${}_I\mathbf{r}_{(i,j)} = {}_I\mathbf{r}_j - {}_I\mathbf{r}_i$ where ${}_I\mathbf{r}_i \in \mathbb{R}^3$ represents the position of object v_i in a common reference frame I. To model spatially local scene context, an edge and associated embedding vector is included for any set of objects (v_i, v_j) where $||{}_I\mathbf{r}_{(i,j)}|| < \tau$ for a distance threshold τ. This can result in varying levels of connectivity, as illustrated Fig. 2. Naturally, once the set of object-to-object edges $\mathcal{E}$ and the associated relationship embedding matrix M_R is defined, the *edge indices matrix* M_E can be computed by referring to the index of objects in the node embedding matrix M_v.

B. Learning Scene Graph Variability

Given a pair of *current* and *future* scene graphs $(\mathcal{G}^c, \mathcal{G}^f)$, we can formulate long-term scene variability estimation as a supervised learning task. We first match all objects $V_i^{(c)} \in \mathcal{G}^c$

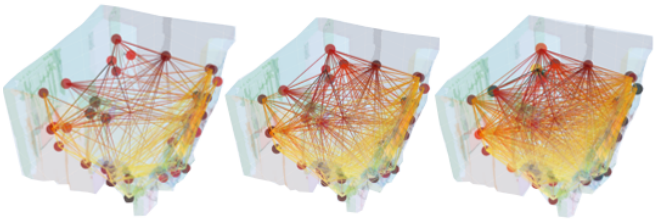

Fig. 2: SG relationships at different geometric edge distance thresholds τ, shown for $\tau = 1^{\text{st}}$, 2^{nd}, and 3^{rd} quartile (left to right).

to their counterparts $V_i^{(f)} \in \mathcal{G}^f$. The ground truth variability of $v_i^{(c)}$ can then be computed as $y_i = [y_P^{(i)}, y_S^{(i)}, y_I^{(i)}] \in [0,1]^3$, as per the definitions of variability in section III. These labels are aggregated for all objects in a scene $Y = [y_1, \ldots, y_{N_v}] \in \mathbb{R}^{N_v \times 3}$. Lastly, the input graph $\mathcal{G}^c$ is embedded to construct a training sample $X = (M_V^{(c)}, M_E^{(c)}, M_R^{(c)})$, Y. We thus formulate a supervised learning task with the objective of learning a model Φ that provides an estimated output vector $\tilde{y}_i = \Phi(M_V^{(i)}, M_E^{(i)}, M_R^{(i)})$.

We propose a Graph Neural Network (GNN) architecture to predict the object-wise variability attributes from the resulting embedded SG. Our proposed architecture to address this learning task, *DeltaVSG*, is based on graph convolutional networks [50]. The network utilizes message-passing graph convolution layers (MP-Conv), which computes a latent feature vector z_i^l for layer l and node i using a standard graph convolution with a nonlinear activation over edge embedding vectors $q_{i,j}$:

$$z_i^{l+1} = f_\theta(z_i^l) + \sum_{j \in \text{Neighbors}(i)} z_j^l \cdot h_\psi(q_{i,j}) \qquad (1)$$

In this case, f_θ and h_ψ are feed-forward neural networks parametrized by weights θ and ψ, respectively. As with the original image convolution, this offers a form of weight sharing which reduces the overall size of the network needed to learn useful representations. As scene changes are oftentimes relatively rare and sparse, we use only two layers of MP-Conv to keep the number of parameters low. We include a Rectified Linear Unit (ReLU) activation between MP-Conv layers, as well as dropout to prevent overfitting. An overview of the architecture is shown in Fig. 1.

C. Data Augmentation and Training Details

We formulate a dataset $\mathcal{D} = \{(X_1, Y_1), \ldots, (X_{N_D}, Y_{N_D})\}$ of samples for the described learning task. However, a central limitation in building datasets for semantic scene variability prediction is that collecting a single sample requires scanning and labeling an entire scene at two different times. In addition, much of the scene may be static. Therefore, positive samples of scene variability are comparably scarce. To overcome this, we hypothesize that through long-term object persistence, the temporal direction of long-term scene changes can be neglected, i.e., a cup is equally likely to move from the kitchen to the table as vice-versa. Thus, we can perform data augmentation by creating unordered and symmetrical pairs from sequences of scans of a scene. Given a series of three scans and their corresponding scene graphs $\{\mathcal{G}_1, \mathcal{G}_2, \mathcal{G}_3\}$, we can form training samples from $(\mathcal{G}_1, \mathcal{G}_2)$, $(\mathcal{G}_2, \mathcal{G}_1)$ (symmetrical), $(\mathcal{G}_1, \mathcal{G}_3)$ (unordered) and so on for a total of $n(n-1)$ samples.

While this significantly increases the number of samples, the data is typically further limited by a high degree of class imbalance. In our experiments, only 21% of nodes had an occurrence of state variability, 17% for position variability, and 13% for instance variability. We therefore employ a focal loss and perform importance sampling to reduce class imbalance. Samples with insufficient state information or instance variability are excluded from the position and state variability loss. This results in the following element-wise loss function:

$$l_i = -w_c \cdot (1 - p_i)^\gamma log(p_i) \qquad (2)$$

where p_i is equal to the probability the model assigns to the label class at element i, $\gamma = 0.5$ is a hyperparameter controlling the amplification and w_c is the class weight.

V. Evaluation

In this section, we provide results on the task of *semantic scene variability estimation* introduced in Sec. III. We demonstrate the effectiveness of our approach *DeltaVSG*, compare methods of embedding semantic and geometric information, and show the utility of VSGs for long-term robot operation.

A. Experimental Setup

We train and evaluate our method on the 3DSSG and 3RScan datasets from Wald *et al.* [7,51] which include 1482 scans of 478 unique indoor environments changing over time, and their associated scene graphs. Repeated measurements were taken on the scale from multiple hours to days, representing the time scale of variability addressed in this experiment. We use the ground truth annotations of the dataset to extract the graphs $\mathcal{G}$ and the variability labels Y. We further perform data augmentation as per Sec. IV-C, generating 3650 effective samples for training.

To accurately evaluate the quality of the variability prediction, we report individual results on *state*, *position*, and *instance* variability. For each case, we present the *accuracy* (Acc.) [%], i.e. the percentage of correct variation predictions, and the *F1-Score* [%] of the variation, i.e. the harmonic mean of precision and recall for objects that did change. In the tables, the highest number is shown in bold.

B. Variable Scene Graph Prediction Performance

To evaluate the capacity to estimate accurate VSGs, the scene variability prediction performance of different models is shown in Tab. I. We compare our DeltaVSG against two global context baselines, a Multi-Layer Perceptron (MLP) operating on our object class and attribute embeddings, and the LayoutTransformer [52] using self-attention to model scene context. We further compare against two graph-based architectures, the GNN of 3DSSG [7], where we replace the PointNet input features with the SG embeddings of [7]

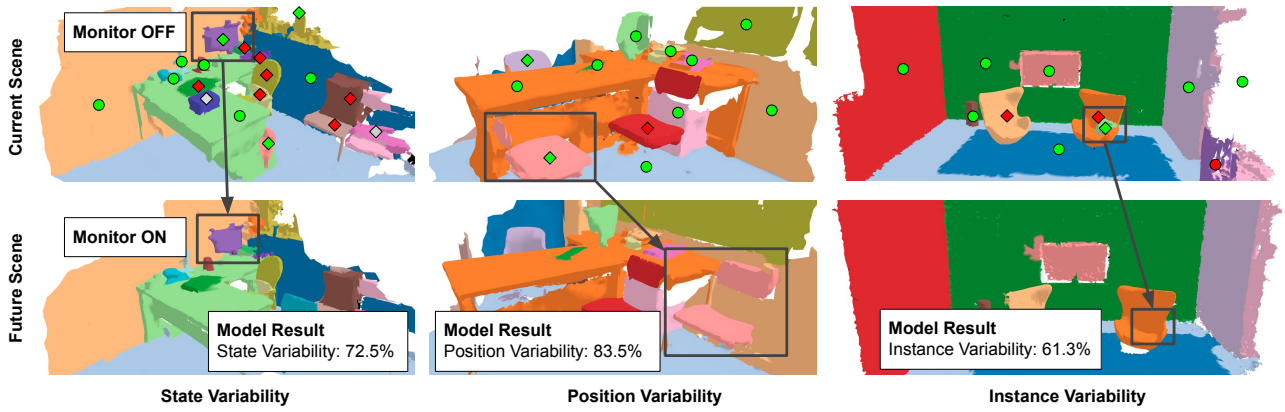

Fig. 3: Qualitative VSG prediction results of DeltaVSG. We show the current scene (top) overlaid with the vertices v_i of the VSG for different variabilities. Nodes are shaped if they are predicted to change ($\diamond$) or remain static ($\circ$). The color indicates **true** and **false** predictions, as compared to the realized future (bottom). Occurences of positive instance variability are **excluded** from the other variabilities.

to operate directly on a SG, and the recent architecture of Spatial Commonsense Graphs (SCG) [20] based on graph transformers [53].

We observe that equivalent fully connected architectures, such as MLP, show the lowest performance. The lack of convolutions may impact the ability to model local context, which significantly decreases recall leading to lower F1 scores. This is most apparent in position and instance variability. We further observe that the transformer-based architectures perform marginally worse across all categories than our graph-convolution-based approach. It is possible that learned attention relationships between object attributes and variability do not generalize as well to this problem. We find that our model matches or outperforms the 3DSSG model, which does not explicitly embed relative distances into object relationships. DeltaVSG, which explicitly embeds semantic and geometric relationships, thus tends to allow the model to generalize better to less common forms of variability, leading to better overall performance in terms of accuracy and recall.

Considering all variabilities, DeltaVSG achieves an accuracy of 77.1% with a recall of 72.3%. Across all approaches, models predict a large number of false positives, decreasing precision and thus F1-score, while accuracy and recall remain relatively high. We also see that the types of variability with fewer samples and a lower positive sample ratio, namely state and instance variability, show a significant drop in accuracy and F1. The overall lower precision reflects the fact that scenes can vary in many equally probable ways due to different possible human interactions. This reflects the challenging nature of the investigated task, as only one of many plausible future scenarios is realized in the evaluation data. Nonetheless, DeltaVSG can predict likely scene changes from local semantic and geometric context. Qualitative examples in Sec. V-C illustrate how this performance translates to a semantically sound understanding of human-driven changes in indoor environments. We also show in Sec. V-E that the achieved performance is sufficient to improve a long-term robot operation task.

TABLE I: VSG prediction performance of different models.

Model	State		Position		Instance	
	Acc.	*F1*	*Acc.*	*F1*	*Acc.*	*F1*
MLP	38.9	42.0	83.3	60.0	55.3	26.5
LayoutTransformer [52]	51.1	45.0	85.2	63.5	55.0	25.4
3DSSG [7]	61.5	34.2	**88.7**	**68.7**	63.1	**31.6**
SCG [20]	59.7	46.3	87.0	64.9	56.4	25.8
DeltaVSG (Ours)	**62.9**	**48.9**	88.1	66.6	**68.5**	30.1

C. Qualitative VSG Prediction Results

Qualitative results of the VSG prediction are shown in Fig. 3, which illustrate how DeltaVSG predictions often capture human intuition on scene variability. The model correctly identifies small handheld objects and frequently used objects such as chairs with a high probability of position or instance variability, as opposed to large pieces of furniture like desks or rugs. At a high level, DeltaVSG predictions are often informed by the larger layout of the scene. A desk configuration more consistent with working use leads to higher probabilities of a state change for a computer monitor. Chairs laid out for a temporary face-to-face meeting are predicted to move back in place. While some of these intuitively likely predictions yield false positives when compared to the specific future scene in the training sample, they still represent a consistent understanding of how objects can change in human-made environments.

D. Comparing Embedding Methods

We present several ablation experiments on the impact of embedding methods for DeltaVSG. First, we study the impact of local geometric context in Tab. II by varying the distance τ for establishing geometric relationship edges. Overall, we find a performance improvement when local geometric edges are added to the graph. However, this effect is lessened when many further away objects are considered, and can even reduce model performance. This indicates the importance of primarily local scene context, which appears to be sufficient information for semantic scene change prediction.

Second, we investigate the impact of the semantic taxonomy, i.e. the number distinguished semantic classes, in

TABLE II: DeltaVSG performance at different geometrical relationship thresholds τ as a percentile of the training distribution.

Percentile	τ [m]	State		Position		Instance	
		Acc.	*F1*	*Acc.*	*F1*	*Acc.*	*F1*
0^{th}	-	52.1	42.4	84.2	53.6	58.0	37.1
25^{th}	1.67	54.5	41.5	87.1	66.5	64.3	**37.5**
50^{th}	2.61	**66.6**	35.1	87.2	63.7	**68.7**	36.7
75^{th}	3.70	62.9	**48.9**	**88.1**	**66.6**	68.5	30.1
100^{th}	15.5	53.9	35.0	86.3	59.2	58.3	24.6

TABLE III: DeltaVSG performance on different class taxonomies.

Taxonomy	State		Position		Instance	
	Acc.	*F1*	*Acc.*	*F1*	*Acc.*	*F1*
RIO527 [7]	62.9	**48.9**	**88.1**	**66.6**	**68.5**	30.1
NYU40 [15]	42.3	31.0	84.6	57.0	57.5	27.3
RIO27 [51]	58.6	36.1	85.8	58.7	63.1	**31.0**
Eigen13 [54]	**64.1**	40.8	86.9	59.4	59.4	27.1

Tab. III. Note that the attributes A and thus the difficulty of the variability estimation task remain unchanged. We note a tradeoff between performance improvements caused by finer-grained semantic information, and degrading performance due to sparser embedding vectors and data samples. However, while our initial model is trained on RIO527 [7], generally high performance is maintained also when using taxonomies with a significantly reduced class set.

Lastly, we discuss the importance of compact embedding in Tab. IV. We note that our final method combining the large RIO object taxonomy and dimensionality reduction leads to the highest performance. This further hints at the trade-off between compact embeddings in learning tasks with sparse positive samples and detailed semantic information.

E. Application to Active Change Detection

Lastly, we demonstrate the utility of VSGs in the representative application of active change detection. In this task, a robot can move from object to object based on the map from the previous time, and at each location measure whether the object has changed. The goal is to identify n changes in the scene.

As a baseline, we employ an optimal *Coverage* path by casting all objects of the previous map into a Traveling Salesman Problem (TSP) [55]. In comparison, we show the utility of VSGs using a variability-aware method. *VSG-Planner* computes the VSG from the previous map using our DeltaVSG network, and then solves the TSP for the $n + 3$ objects with the highest variability probabilities. If less than n changes are observed within that path, the robot resorts to *Coverage* for the remaining objects.

The performance of both methods is shown in Fig. 4. Overall, we observe that our DeltaVSG model can be effectively used to improve task efficiency, especially when a larger number of changes must be detected. Again, we observe the challenge in long-term scene change prediction of different possible futures existing, reflected in the high variances. Nonetheless, VSG-Planner is able to identify a shorter path in 68.7% of cases, speeding up change detection by 66.0% on average.

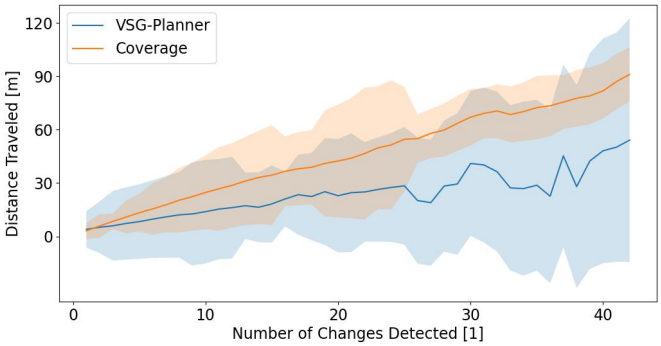

Fig. 4: Mean and standard deviation of distance traveled until n changed objects are detected in the active change detection task.

As such, we believe a robot performing inspections, collecting data, or performing other continuous tasks in human environments could leverage VSGs to assign uncertainty to the prior scene graph or environment map.

TABLE IV: DeltaVSG performance with different semantic object class embedding sizes d_v.

Embedding	d_v	State		Position		Instance	
		Acc.	*F1*	*Acc.*	*F1*	*Acc.*	*F1*
Binary	619	**81.4**	0.0	71.3	39.1	**87.8**	0.0
PCA	120	62.9	**48.9**	**88.1**	**66.6**	68.5	**30.1**
PCA	48	58.8	35.5	85.6	58.6	70.2	28.1

VI. Conclusions

In this work, we addressed the challenge of modeling semantic scene changes for robots continuously operating in shared environments. We thus formalized the problem of *semantic scene variability estimation*, identifying three major types of long-term semantic scene change, and formulate it as a supervised learning task. We presented 3D Variable Scene Graphs (VSG) as a natural and compact representation of scene variability. Lastly, we developed *DeltaVSG*, a novel approach operating on existing SG representations and combining explicit semantic and geometric embeddings with a GNN architecture to estimate the resulting VSG. We show in thorough experimental evaluation that our approach is able to capture intuitive scene variability predictions, achieving accuracy of 77.1% and recall of 72.3% on this challenging task. We demonstrate the utility of VSGs in the task of long-term change detection, speeding up task completion by an average of 66.0% compared to variability-unaware methods.

While we presented a first approach to estimate VSGs, there are numerous directions for future research in this task of semantic scene variability estimation. While we primarily focused on predicting change events, these can be extended to more detailed predictions, such as predicting the exact future state, where an object is likely to move, or also accounting for other topological changes such as where and which objects are likely to appear. Lastly, we hope to encourage more research in applying semantic change predictions and VSGs to different long-term robot autonomy tasks.

REFERENCES

[1] J. McCormac, A. Handa, A. Davison, and S. Leutenegger, "Semanticfusion: Dense 3d semantic mapping with convolutional neural networks," in *IEEE Int. Conf. on Robotics & Automation*, 2017, pp. 4628–4635.

[2] A. Rosinol, M. Abate, Y. Chang, and L. Carlone, "Kimera: An Open-Source Library for Real-Time Metric-Semantic Localization and Mapping," in *IEEE Int. Conf. on Robotics & Automation*, May 2020, pp. 1689–1696.

[3] Y. Jiang, X. Ma, F. Fang, and X. Kang, "Indoor instance-aware semantic mapping using instance segmentation," in *2021 33rd Chinese Control and Decision Conference (CCDC)*. IEEE, 2021, pp. 3549–3554.

[4] 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.

[5] L. Schmid, J. Delmerico, J. L. Schönberger, J. Nieto, M. Pollefeys, R. Siegwart, and C. Cadena, "Panoptic Multi-TSDFs: a Flexible Representation for Online Multi-resolution Volumetric Mapping and Long-term Dynamic Scene Consistency," in *2022 International Conference on Robotics and Automation (ICRA)*, May 2022, pp. 8018–8024.

[6] I. Armeni, Z.-Y. He, J. Gwak, A. R. Zamir, M. Fischer, J. Malik, and S. Savarese, "3d scene graph: A structure for unified semantics, 3d space, and camera," in *IEEE/CVF Int. Conf. on Computer Vision*, 2019, pp. 5664–5673.

[7] J. Wald, H. Dhamo, N. Navab, and F. Tombari, "Learning 3d semantic scene graphs from 3d indoor reconstructions," in *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, 2020, pp. 3961–3970.

[8] A. Rosinol, A. Gupta, M. Abate, J. Shi, and L. Carlone, "3d dynamic scene graphs: Actionable spatial perception with places, objects, and humans," *Proc. of Robotics: Science and Systems*, 2020.

[9] L. Han, T. Zheng, L. Xu, and L. Fang, "Occuseg: Occupancy-aware 3d instance segmentation," in *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, 2020, pp. 2940–2949.

[10] J. McCormac, R. Clark, M. Bloesch, A. Davison, and S. Leutenegger, "Fusion++: Volumetric object-level slam," in *Int. Conf. on 3D Vision*. IEEE, 2018, pp. 32–41.

[11] W. Yang, X. Wang, A. Farhadi, A. Gupta, and R. Mottaghi, "Visual semantic navigation using scene priors," *arXiv preprint arXiv:1810.06543*, 2018.

[12] A. Muzahid, W. Wan, F. Sohel, L. Wu, and L. Hou, "Curvenet: Curvature-based multitask learning deep networks for 3d object recognition," *IEEE/CAA Journal of Automatica Sinica*, vol. 8, no. 6, pp. 1177–1187, 2020.

[13] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma, *et al.*, "Visual genome: Connecting language and vision using crowdsourced dense image annotations," *International journal of computer vision*, vol. 123, no. 1, pp. 32–73, 2017.

[14] C. Lu, R. Krishna, M. Bernstein, and L. Fei-Fei, "Visual relationship detection with language priors," in *European Conf. on Computer Vision*. Springer, 2016, pp. 852–869.

[15] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus, "Indoor segmentation and support inference from rgbd images," in *European Conf. on Computer Vision*. Springer, 2012, pp. 746–760.

[16] M. Hassan, P. Ghosh, J. Tesch, D. Tzionas, and M. J. Black, "Populating 3d scenes by learning human-scene interaction," in *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, 2021, pp. 14 708–14 718.

[17] L. L. Wong, L. P. Kaelbling, and T. Lozano-Pérez, "Manipulation-based active search for occluded objects," in *2013 IEEE International Conference on Robotics and Automation*. IEEE, 2013, pp. 2814–2819.

[18] R. Druon, Y. Yoshiyasu, A. Kanezaki, and A. Watt, "Visual object search by learning spatial context," *IEEE Robotics and Automation Letters*, vol. 5, no. 2, pp. 1279–1286, 2020.

[19] J. Johnson, R. Krishna, M. Stark, L.-J. Li, D. Shamma, M. Bernstein, and L. Fei-Fei, "Image retrieval using scene graphs," in *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, 2015, pp. 3668–3678.

[20] F. Giuliari, G. Skenderi, M. Cristani, Y. Wang, and A. Del Bue, "Spatial commonsense graph for object localisation in partial scenes," in *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, 2022, pp. 19 518–19 527.

[21] J. Yang, J. Lu, S. Lee, D. Batra, and D. Parikh, "Graph r-cnn for scene graph generation," in *European Conf. on Computer Vision*, 2018, pp. 670–685.

[22] X. Lin, C. Ding, J. Zeng, and D. Tao, "Gps-net: Graph property sensing network for scene graph generation," in *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, 2020, pp. 3746–3753.

[23] S.-C. Wu, J. Wald, K. Tateno, N. Navab, and F. Tombari, "Scenegraphfusion: Incremental 3d scene graph prediction from rgb-d sequences," in *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, 2021, pp. 7515–7525.

[24] N. Hughes, Y. Chang, and L. Carlone, "Hydra: A real-time spatial perception engine for 3d scene graph construction and optimization," *Proc. of Robotics: Science and Systems*, 2022.

[25] T. Tahara, T. Seno, G. Narita, and T. Ishikawa, "Retargetable ar: Context-aware augmented reality in indoor scenes based on 3d scene graph," in *2020 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct)*. IEEE, 2020, pp. 249–255.

[26] C. Agia, K. M. Jatavallabhula, M. Khodeir, O. Miksik, V. Vineet, M. Mukadam, L. Paull, and F. Shkurti, "Taskography: Evaluating robot task planning over large 3d scene graphs," in *Int. Conf. on Robot Learning*. PMLR, 2022, pp. 46–58.

[27] Y. Zhu, J. Tremblay, S. Birchfield, and Y. Zhu, "Hierarchical planning for long-horizon manipulation with geometric and symbolic scene graphs," in *IEEE Int. Conf. on Robotics & Automation*. IEEE, 2021, pp. 6541–6548.

[28] Y. Qiu, A. Pal, and H. I. Christensen, "Learning hierarchical relationships for object-goal navigation," *arXiv preprint arXiv:2003.06749*, 2020.

[29] A. Kurenkov, R. Martín-Martín, J. Ichnowski, K. Goldberg, and S. Savarese, "Semantic and geometric modeling with neural message passing in 3d scene graphs for hierarchical mechanical search," in *IEEE Int. Conf. on Robotics & Automation*. IEEE, 2021, pp. 11 227–11 233.

[30] A. Luo, Z. Zhang, J. Wu, and J. B. Tenenbaum, "End-to-end optimization of scene layout," in *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, 2020, pp. 3754–3763.

[31] H. Dhamo, F. Manhardt, N. Navab, and F. Tombari, "Graph-to-3d: End-to-end generation and manipulation of 3d scenes using scene graphs," in *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, 2021, pp. 16 352–16 361.

[32] R. C. Daudt, B. Le Saux, and A. Boulch, "Fully convolutional siamese networks for change detection," in *IEEE International Conference on Image Processing (ICIP)*. IEEE, 2018, pp. 4063–4067.

[33] T. Suzuki, M. Minoguchi, R. Suzuki, A. Nakamura, K. Iwata, Y. Satoh, and H. Kataoka, "Semantic change detection," in *2018 15th International Conference on Control, Automation, Robotics and Vision (ICARCV)*. IEEE, 2018, pp. 1785–1790.

[34] J.-M. Park, J.-H. Jang, S.-M. Yoo, S.-K. Lee, U.-H. Kim, and J.-H. Kim, "Changesim: towards end-to-end online scene change detection in industrial indoor environments," in *IEEE/RSJ Int. Conf. on Intelligent Robots and Systems*. IEEE, 2021, pp. 8578–8585.

[35] W. Cheng, Y. Zhang, X. Lei, W. Yang, and G. Xia, "Semantic change pattern analysis," *arXiv preprint arXiv:2003.03492*, 2020.

[36] L. Ru, B. Du, and C. Wu, "Multi-temporal scene classification and scene change detection with correlation based fusion," *IEEE Transactions on Image Processing*, vol. 30, pp. 1382–1394, 2020.

[37] S. Kim, K.-n. Joo, and C.-H. Youn, "Graph neural network based scene change detection using scene graph embedding with hybrid classification loss," in *2021 International Conference on Information and Communication Technology Convergence (ICTC)*. IEEE, 2021, pp. 190–195.

[38] X. Li, H. Duan, Y. Tian, and F.-Y. Wang, "Exploring image generation for uav change detection," *IEEE/CAA Journal of Automatica Sinica*, vol. 9, no. 6, pp. 1061–1072, 2022.

[39] Y. Qiu, Y. Satoh, R. Suzuki, K. Iwata, and H. Kataoka, "3d-aware scene change captioning from multiview images," *IEEE Robotics and Automation Letters*, vol. 5, no. 3, pp. 4743–4750, 2020.

[40] T. Ku, S. Galanakis, B. Boom, R. C. Veltkamp, D. Bangera, S. Gangisetty, N. Stagakis, G. Arvanitis, and K. Moustakas, "Shrec 2021: 3d point cloud change detection for street scenes," *Computers & Graphics*, vol. 99, pp. 192–200, 2021.

[41] J. Li, P. Tang, Y. Wu, M. Pan, Z. Tang, and G. Hui, "Scene change detection: semantic and depth information," *Multimedia Tools and Applications*, vol. 81, no. 14, pp. 19 301–19 319, 2022.

[42] 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 *IEEE Int. Conf. on Robotics & Automation*. IEEE, 2017, pp. 5237–5244.
[43] R. Finman, T. Whelan, M. Kaess, and J. J. Leonard, "Toward lifelong object segmentation from change detection in dense rgb-d maps," in *European Conf. on Mobile Robots*, 2013, pp. 178–185.
[44] E. Langer, T. Patten, and M. Vincze, "Robust and efficient object change detection by combining global semantic information and local geometric verification," in *IEEE/RSJ Int. Conf. on Intelligent Robots and Systems*, 2020, pp. 8453–8460.
[45] Z. Liao, Q. Huang, Y. Liang, M. Fu, Y. Cai, and Q. Li, "Scene graph with 3d information for change captioning," in *Proceedings of the 29th ACM International Conference on Multimedia*, 2021, pp. 5074–5082.
[46] T. Krajník, J. Pulido Fentanes, M. Hanheide, and T. Duckett, "Persistent localization and life-long mapping in changing environments using the frequency map enhancement," in *IEEE/RSJ Int. Conf. on Intelligent Robots and Systems*, 2016, pp. 4558–4563.
[47] L. Wang, W. Chen, and J. Wang, "Long-term localization with time series map prediction for mobile robots in dynamic environments," in *IEEE/RSJ Int. Conf. on Intelligent Robots and Systems*. IEEE, 2020, pp. 1–7.
[48] S. Casas, C. Gulino, R. Liao, and R. Urtasun, "Spagnn: Spatially-aware graph neural networks for relational behavior forecasting from sensor data," in *IEEE Int. Conf. on Robotics & Automation*. IEEE, 2020, pp. 9491–9497.
[49] Z. Ravichandran, L. Peng, N. Hughes, J. D. Griffith, and L. Carlone, "Hierarchical representations and explicit memory: Learning effective navigation policies on 3d scene graphs using graph neural networks," in *IEEE Int. Conf. on Robotics & Automation*. IEEE, 2022, pp. 9272–9279.
[50] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," *arXiv preprint arXiv:1609.02907*, 2016.
[51] 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 Conf. on Computer Vision and Pattern Recognition*, 2019, pp. 7658–7667.
[52] K. Gupta, J. Lazarow, A. Achille, L. S. Davis, V. Mahadevan, and A. Shrivastava, "Layouttransformer: Layout generation and completion with self-attention," in *IEEE/CVF Int. Conf. on Computer Vision*, 2021, pp. 1004–1014.
[53] Y. Shi, Z. Huang, S. Feng, H. Zhong, W. Wang, and Y. Sun, "Masked label prediction: Unified message passing model for semi-supervised classification," *arXiv preprint arXiv:2009.03509*, 2020.
[54] C. Couprie, C. Farabet, L. Najman, and Y. LeCun, "Indoor semantic segmentation using depth information," *arXiv preprint arXiv:1301.3572*, 2013.
[55] M. Jünger, G. Reinelt, and G. Rinaldi, "The traveling salesman problem," *Handbooks in operations research and management science*, vol. 7, pp. 225–330, 1995.