EgoSG: Learning 3D Scene Graphs from Egocentric RGB-D Sequences

Chaoyi Zhang
University of Sydney

Xitong Yang
FAIR at Meta

Ji Hou
Meta GenAI

Kris Kitani
FAIR at Meta

Weidong Cai
University of Sydney

Fu-Jen Chu
FAIR at Meta

Abstract

Constructing a 3D scene graph of an environment is essential for agents and smart glasses assistants to develop an understanding of their surroundings and predict relationships between various entities within it. 3D Scene Graph Prediction (3DSGP) is commonly adopted to predict the spatial and semantic relationships between objects in a 3D environment reconstructed from posed (calibrated) RGB-D sequences, such as object containment or adjacency. However, reconstructing a scene can be time-consuming and computationally intensive, and requires specialized hardware like IMUs for accurate poses. The reliance on (1) robust algorithms and (2) accurate camera poses limits its applicability. Unlike existing 3DSGP methods, we propose to perform perception and reasoning on each frame without assuming available camera poses, which we call EgoSG, to estimate 3D scene graphs directly from egocentric frame sequences. In our method, per-frame instance features are acquired from a partial (per-frame) point cloud. By globally optimizing per-frame features, object instances are then associated across the egocentric frames, and graph representations are aggregated for 3D scene graph prediction. Compared to the state-of-the-art methods that heavily rely on 3D reconstruction, our approach is reconstruction-free and can be derived directly from unposed RGB-D sequences. We benchmark our EgoSG framework against existing reconstruction-based approaches on 3DSGP tasks. Our method outperforms the state-of-the-art methods by a large margin, achieving +44.63 R@1 in Object and +22.74 R@1 in Predicate from egocentric sequences without any reliance on reconstruction algorithms or camera poses.

1. Introduction

Egocentric agents capture 4D perception footage for scene understanding purposes while it actively traverses and interacts with the complex 3D world [13]. This process approximates human visual system (HVS) to some extent, in recognizing and reasoning the 3D visual clues surrounding us. Abstracting scene graphs from 3D environments can be an efficient and reasonable way to assist agents in structuring their spatial-awareness. Specifically, 3D objects and their relationships are usually captured and inferred as nodes and edges during the estimation of 3D scene graphs, with the objective of learning the contexts and relations between instances within a scene. Examples of applications that benefit from 3DSGP include interior adornment, robotic navigation, and AR/VR-powered intelligent applications.

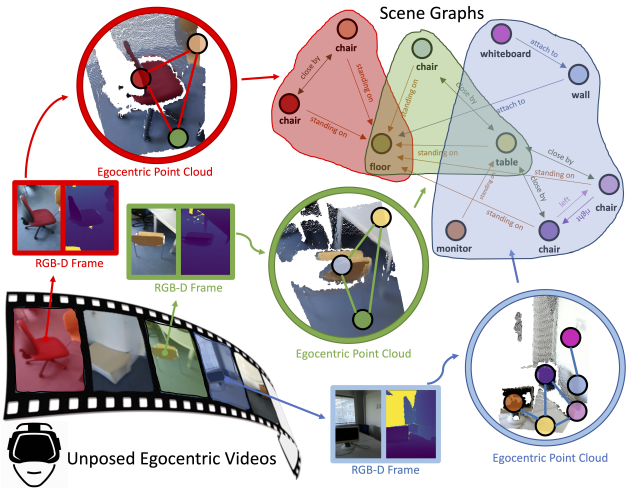

Figure 1. We present EgoSG to learn 3D scene graphs from unposed RGB-D videos.

Recently, a line of work [40, 44, 53] has been proposed to perform 3D scene graph prediction (3DSGP), which was an emerging 3D scene understanding task to jointly recognize localized objects and their relationships in 3D environments, as defined in [40]. However, these existing 3DSGP studies expect to take as inputs high-quality reconstructed

scenes, where their graph prediction network could be built upon. Unfortunately, this assumption has limited its applicability in mobile or egocentric applications, especially when accurate camera poses or SLAM algorithms become unavailable or unreliable, due to the hardware limitations and power consumption.

To this end, unlike all existing 3DSGP approaches, we propose EgoSG, a novel method that learns the scene graphs from unposed RGB-D sequences via a reconstruction-free manner. As the perception and reasoning processes of the inferred graphs are performed on each ego-view of the environments, we term these estimated graphs as egocentric scene graphs (EgoSGs), to differ them from the typical 3D scene graphs which are learnt directly from reconstructed global 3D scenes. To summarize, our contributions are in three-folds:

- We propose EgoSG that learns 3D scene graphs from egocentric RGB-D sequences, without any reliance on camera poses or SLAM techniques.
- Specifically, we propose to conduct local and global graph reasoning to make use of the spatial-aware context from both frame-level and video-level, respectively.
- Our approach outperforms the state-of-the-art approaches whose assumptions (*i.e.*, accurate point cloud reconstructions) might be hard to meet in many mobile devices such as VR/AR devices and mid-range phones.

2. Related Work

3D Scene Understanding. Indoor 3D scene understanding has been developed very fast since large-scale RGB-D datasets are introduced, such as ScanNet [5] and S3DIS [1]. Both benchmarks contribute to fast movement of the 3D scene understanding community. 3D scene understanding as a general concept contains a variety of 3D perception and reasoning tasks, which could be divided into three categories: low-level, middle-level and high-level tasks. Similar to 2D tasks where low-level vision deals with pixel-level features, 3D low-level tasks mainly operate on point cloud and focus on point-level tasks, such as point cloud registration [4, 18, 32, 50, 51, 54]. Middle-level tasks aim to predict aggregated entities in the scene, such as 3D semantic instance segmentation [8, 15, 16, 19, 28, 41, 56], and 3D object detection [31, 36, 47, 55]. As high-level tasks, they focus on understanding more abstract semantics, such as navigation, room layout estimation, and 3D scene graph prediction. In this paper, we focus on the high-level task end, more specifically 3D scene graph prediction problems, where mainly the relations of instance objects in the scene are inferred.

2D and 3D Scene Graph Analysis. Initially, scene graphs (SG) were utilized in computer vision to capture additional semantic knowledge about objects and their interrelationships for image retrieval [20], while it was inherited from visual relationship detection [27], to explore visual understanding beyond objects. Following the release of the visual genome dataset [23], which contains large-scale SG annotations on images, several image-based SG prediction and generation techniques were developed, including Xu et al.'s use of gated recurrent units (GRUs) to pass message iteratively between primal and dual graphs formed by the nodes and edges of the SG [45], and MotifNet's generation of SGs using global context parsed through bidirectional LSTM [52]. Most methods dealt with the SG detection problems are typically built upon advanced object detection networks, such as Graph R-CNN [48] proposed an attentional variant of GCN and combined it with Faster R-CNN [35] to process contextual information between objects and relationships. Later, [40] firstly introduced scene graph recognition task to 3D community, where 3D scene graphs are inferred from 3D reconstructed scenes, to describe the high-level relations between objects in the environment, and this task has been being developed fast recently [22, 43]. Some other work proposed to incrementally reconstruct the scenes form posed RGB-D sequences, where they built their graph prediction network on top of SLAM algorithms [44]. Here we add one new modality of learning SG, from unposed RGB-D sequences, which requires no reconstruction to be preformed prior to the scene graph recognition.

Egocentric perception and Reasoning. Egocentric perception and reasoning refer to the ability of an agent, such as a robot of a human wearing a camera, to perceive and reasoning about its environment form a first-person perspective. Different from traditional 3D Scene Understanding, where reconstructed point clouds are already available, Egocentric 3D Scene Understanding aims to understand our 3D environments without knowing 3D reconstructions in advance. One of the challenges in egocentric 3D scene understanding [13] is the lack of pre-existing 3D reconstruction data or camera poses. Instead, these methods [9, 12] reply on processing the visual data obtained from egocentric perspective to complete high-level understanding task, such as human-object interactions [3, 6], parsing social interactions [10, 49], and activity recognition [57]. To address these challenges, researchers are developing methods that can handle the variability in egocentric data and adapt to changes in the environment over time. In this paper, we push further this direction to 3D Scene Graph Prediction from egocentric frames, where no reconstructed point clouds and camera poses are given. And given such context of egocentric perception and reasoning, 3D scene graph prediction refers to the task of recognizing objects and their structural relationships in a scene from an egocentric perspective.

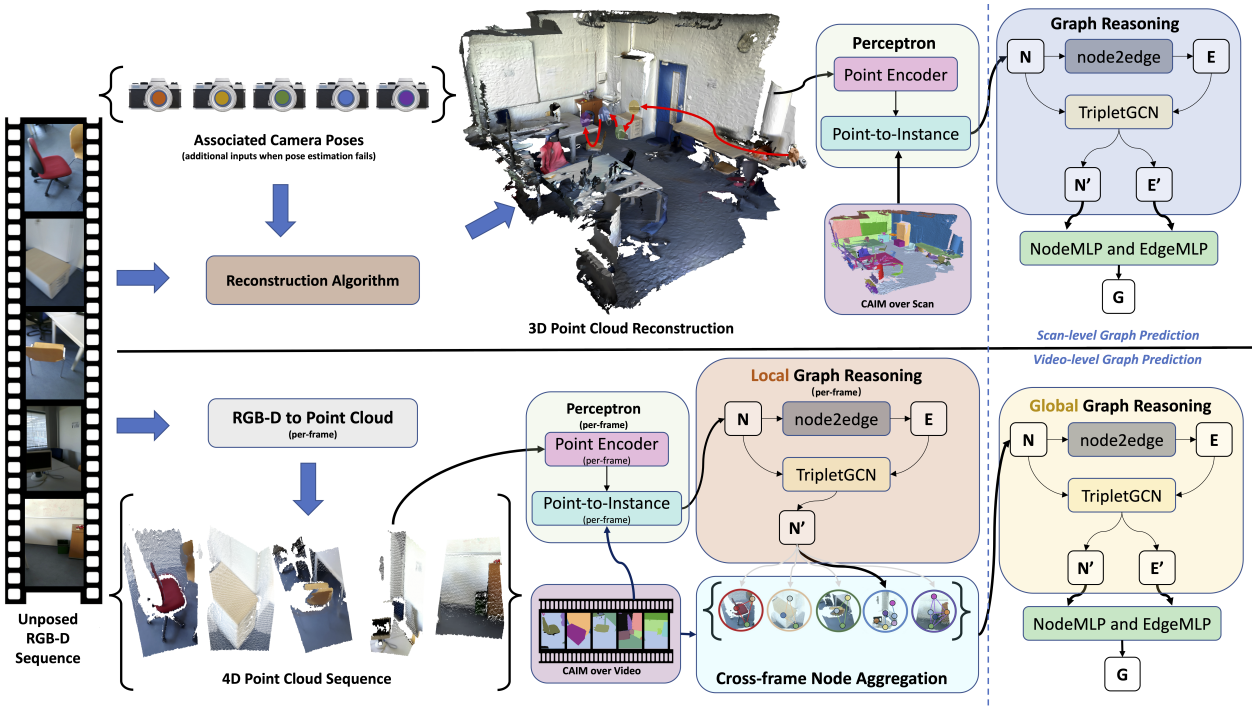

Figure 2. Revisiting 3D Scene Graph Prediction, given localized objects. We propose EgoSG, a reconstruction-free solution (bottom row) to infer scene graphs from unposed RGB-D sequences. Top row: reconstruction-based approach, which might require additional camera poses as inputs to derive 3D point cloud reconstruction (with red arrows denoted as the camera trajectories; zoom in to see cameras of associated poses). CAIM denotes class-agnostic instance mask. N, E, N', and E' denote node and edge features, as well as their refined variants, respectively, while G denotes the estimated 3D scene graph.

3. Method

We first revisit the 3D Scene Graph Prediction (3DSGP) task, including problem formulation, the existing reconstruction-based paradigm, and potential shortages (§3.1). Next, we present EgoSG, a reconstruction-free solution for 3DSGP, as well as its local and global graph reasoning designs (§3.2). Then, we elaborate details of a grouped `node2edge` aggregator, which was used in both our local and global graph reasoning modules (§3.3). Eventually, we present our training losses (§3.4).

3.1. Revisiting 3D Scene Graph Prediction

Inferring scenes graphs from reconstructed 3D environments. Given a 3D environment, 3DSGP aims at estimating a scene graph $\mathcal{G} = (\mathcal{N}, \mathcal{E})$, whose nodes $\mathcal{N}$ and edges $\mathcal{E}$ represent the objects and their structural relationships appeared in the environment, respectively. For reconstruction-based approaches (top row in Fig. 2), the 3D environments are typically represented as reconstructed point clouds $P \in R^{N\times 6}$, which contain N points with spatial coordinates (xyz) and color information (RGB) encoded. These points are fed into the point encoder $\mathcal{F}_{enc}(\cdot)$ to extract point-wise visual features $f_{point} \in R^{N\times 256}$, which are then aggregated to form instance-wise descriptors $f_{inst.} \in R^{m\times 256}$ for m localized objects. This point-to-instance aggregation is performed as in [40, 53], where a symmetric pooling function $g(\cdot)$ [29] is applied to f_{point} over the provided class-agnostic instance mask (as discussed below). Then, scan-level graph prediction is performed to infer final 3D scene graphs $\mathcal{G}$, where a graph reasoning ($\mathcal{GR}$) operation is usually employed to conduct message propagation over 3D contexts.

In the meantime, graph prediction network can also be built upon the reconstruction algorithm, such as [44], which proposed a SLAM-based SGP approach to iteratively reconstruct 3D scenes from posed RGB-D frames for scene graph inferences.

Shortage of reconstruction-based SGP. Several practical challenges can arise when applying SLAM techniques to real-world scenarios, including environmental complexity, noisy or incomplete sensor data, and the requirement of consistent mapping [2, 38]. These challenges can largely impede accurate and robust pose estimation, leading to accumulated errors and reduced downstream performance. And noticeably, as a result, the aforementioned prototype [44] relies heavily on the ground truth camera poses as auxiliary inputs, which might constrain its practical applicability to some extent. To alleviate these problems, a reconstruction-free framework will be investigated below to estimate scene graphs solely from unposed RGB-D se-

quences.

Class-Agnostic Instance Mask. The 3DSGP task is first introduced in 3DSSG [40] to benchmark the graph prediction capability of the network, without the requirements of object localization. Aligning it with the 2D vision community, this condition was inherited from RCNN [11] by VRD [27] to study how challenging it is to conduct relationship prediction without the limitation of 2D object detection, and later, itself and its variants have been broadly applied in a line of image-based SG learning tasks, including PredCls [11], PredDet and SGCls [27]. Specifically, this condition (i.e., given localized objects) is adapted to the 3DSG community by equipping networks with ground truth class-agnostic instance masks (CAIMs), which could be replaced by other point-to-instance indicators generated by any off-the-shell or trainable detectors in future work [40].

3.2. EgoSG: a reconstruction-free solution for SGP

The overall design of our EgoSG framework is shown as the bottom row of Fig. 2. For reconstructed-based approaches who predict graphs from 3D point cloud reconstructions, CAIM is given as a 3D point-to-instance mask to help localize instance from points. Since we take 4D point cloud sequence, the associated CAIM is now provided as two indicators, namely per-frame 3D point-to-instance mask, and cross-frame instance associations. The former would assist in localizing instances from each point cloud frame, while the latter would help to associate frame-wisely localized instance to video-wisely localized objects.

Ego-view perception. Given an egocentric footage traversing the 3D environment, we first adopt a default pinhole camera model frame-by-frame, to convert the unposed RGB-D frames into a 4D point cloud sequence $[P_1, P_2, ..., P_T]$ of length T. Each frame P_t contains an egocentric view of the partially observed 3D environment at time-step t. We then apply the point encoder $\mathcal{F}_{enc}$ on each point cloud frame P_t to extract the per-frame point-wise features $f^t_{point} \in R^{N_t \times 256}$, where N_t is the number of points in P_t.

Next, we extract the per-frame point-to-instance indicators $I^t_{p2i} \in R^{N_t \times m_t}$ from CAIM. It helps to localize m_t instances from N_t points within P_t. With such frame-based object localization given, we can use $g(\cdot)$ to aggregate f^t_{point} to form per-frame instance-wise features $f^t_{inst.} \in R^{m_t \times 256}$ at time-step t.

Ego-view reasoning. Unlike graph prediction networks who take 3D reconstructions as inputs, where global contexts can be directly obtained from reconstructed 3D scenes, EgoSG seems to miss the global 3D context information when applying reconstruction-free solutions over videos. To compensate our lack of global 3D context, we propose to learn per-frame local context by conducting local graph reasoning ($\mathcal{GR}_{\texttt{local}}$) operation.

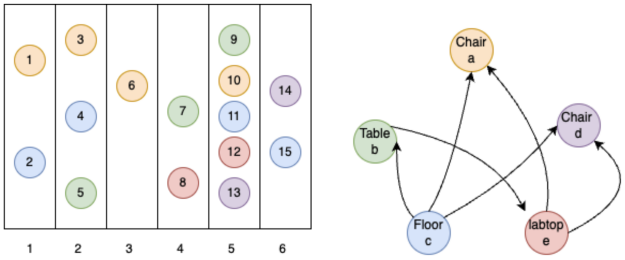

Figure 3. Cross-frame node association from M frame-wisely localized instances (left) to Q video-wisely localized objects (right), with $M = 15$ and $Q = 5$.

Specifically, our proposed $\mathcal{GR}_{\texttt{local}}$ module would (1) take the instance-wise features $f^t_{inst.}$ at time-step t, as the initial frame-level node features $\mathcal{N}^t_{frame}$, (2) derive per-frame edge features $\mathcal{E}^t_{frame}$ from nodes via `node2edge` scheme (with details in 3.3), (3) construct the per-frame $\mathcal{GR}_{\texttt{local}}$ structure $G^t_{frame} = (\mathcal{N}^t_{frame}, \mathcal{E}^t_{frame})$ and adopt a two-layer TripletGCN [21] to conduct message propagation between nodes and edges jointly. Eventually, as shown in Fig. 2, $\mathcal{GR}_{\texttt{local}}$ would only output refined $\mathcal{N}^{t\,\prime}_{frame}$ (*i.e.*, with $\mathcal{E}^{t\,\prime}_{frame}$ omitted), and passed them back to our main framework for the following cross-frame node association.

Noticeably, within the aforementioned $\mathcal{GR}_{\texttt{local}}$ process, T egocentric graphs G_{frame} are built as intermediate representations to conduct ego-view reasoning. It is theoretically feasible to add direct supervisions on these per-frame graph representations, and such supervisions lead the modal to produce local scene graphs $\mathcal{G}_t$ at time-step t. However, we decide not to introduce them, leaving our model to be trained with video-level supervisions alone, with the purpose of recognizing 3D relations that are not observed in ego-views.

Cross-frame node association. Previously we perform per-frame perception and reasoning over 4D point cloud sequence $\{P_t\}$, and, for each frame P_t, we obtain $\mathcal{N}^{t\,\prime}_{frame}$ refined node features for localized instances at time-step t. Now we concatenate them together to form $\mathcal{N}_{frames} \in R^{M \times 256}$ node features of all *frame-wisely* localized instances, where $M = \sum_{t=1}^{T} m_t$ is the total number of frame-wisely localized instances in entire point cloud sequence $\{P_t\}$.

Similar to point-to-instance indicators, we can also extract the frame-to-video association indicators $I_{f2v} \in R^{M \times Q}$ from CAIM, where Q denotes the number of *video-wisely* localized objects in entire point cloud sequence $\{P_t\}$. As illustrated in Fig. 3, based on I_{f2v}, we apply $g(\cdot)$ over $\mathcal{N}_{frames}$ to form the localized object features $\mathcal{N}_{video} \in R^{Q \times 256}$ for video-level graph predictions.

Video-level graph prediction. Taking as inputs the associated node features $\mathcal{N}_{video}$ at video-level, we can now perform video-level graph predictions. In order to recognize

the 3D relations that have not been observed in any frames, we proposed to conduct global graph reasoning ($\mathcal{GR}_{\texttt{global}}$) operation at video-level, to derive the 3D relation features from localized object features, and refine them via message propagation over global contexts.

Specifically, to achieve global graph reasoning, we propose to (1) employ the `node2edge` scheme (with details in 3.3) to obtain global relation features $\mathcal{E}_{video}$ from localized object features $\mathcal{N}_{video}$, (2) construct the video-level $\mathcal{GR}_{\texttt{global}}$ structure $G_{video} = (\mathcal{N}_{video}, \mathcal{E}_{video})$ and adopt another two-layer TripletGCN to conduct interactive message propagation between localized objects and their 3D relations. Eventually, as shown in Fig. 2, $\mathcal{GR}_{\texttt{global}}$ would output both refined graph features $\mathcal{N}'_{video}$ and $\mathcal{E}'_{video}$, which are further fed into two MLPs for object and relationship predictions towards ultimate 3D scene graph estimation.

Noticeably, for benchmark purpose, we do not include any SG-specific attention designs in (either frame-level or video-level) message propagation over G_{frame} and G_{video}, such as the Twinning Attentions [46, 53] or Feature-wise Attention [44], which have been reported to be helpful in SG learning tasks and could thus further improve the overall performance intuitively.

3.3. Grouped `node2edge` Aggregator

Nodes and edges are vital entities in scene graph representation learning, which encodes the object-centric visual appearances and the inter-object pairwise relations, respectively. 3DSSG [40] employs two separate encoders of same structure to independently captures object and relation features from 3D point cloud scans. To improve memory efficiency, both $\mathbf{SGG}_{\mathbf{point}}$ [53] and SGFusion [44] propose to *derive edge features from nodes*, where SGFusion computes edge features from heuristic node properties, such as std. dev. of spatial coordinates and bounding box volumes, while $\mathbf{SGG}_{\mathbf{point}}$ adopts the feature engineering (`diff` and `identity`) used in EdgeConv [42] to avoid hand-crafted aggregators. However, it's still troublesome to decide the appropriate ones between a variety of symmetric (e.g., `avg`) and asymmetric (e.g., `diff`) aggregators, as they might both capture desired inter-node properties to some extend.

Here, we propose a parameter-free feature aggregator `node2edge` to extend *this idea*. Inspired by grouped convolution [24], given a set of aggregation operators $\{\texttt{aggr}_i\}_{i=1}^{\texttt{G}}$ to be performed on edges (src, dst), we split their source (src) and destination (dst) node features $\mathcal{N}$ into `G` groups, apply associated $\texttt{aggr}_i(\cdot,\cdot)$ for each paired group of src_i and dst_i, and group them back as directed edge features $\mathcal{E} : src \rightarrow dst$ with $d_{edge} = d_{node}$. Specifically, it can be described as

$$\mathcal{E} = \underbrace{aggr_1(\mathcal{N}_1^{src}, \mathcal{N}_1^{dst})}_{\text{group 1}} +\!\!+ \cdots +\!\!+ \underbrace{aggr_{\texttt{G}}(\mathcal{N}_{\texttt{G}}^{src}, \mathcal{N}_{\texttt{G}}^{dst})}_{\text{group G}}, \quad (1)$$

where $+\!\!+$ denotes the concatenation operation.

Our design is parameter-free yet effective in automatically engineering edge features of fairly good quality, as it manages to pass the buck of selecting appropriate edge descriptors to the optimization process of its prior object-centric encoders via back propagation. In our aforementioned EgoSG framework, we invoke this module in construction of both frame-level graphs $G_{\texttt{frame}}$ and video-level graphs $G_{\texttt{video}}$.

3.4. Training Losses

The 3DSSG task is overall supervised by two SG-specific recognition losses [40, 44, 53], which include a multi-class cross-entropy loss L_{obj} for object classification, and a multi-label cross-entropy loss L_{rel} for relation classification. To tackle the class imbalance problem, [40], we extend both two losses with focal loss [25], which is found to be less effective in our preliminary experiments, so we instead simply compute SG-specific recognition losses with class weights. More specifically, we compute normalized inverse frequency for multi-class L_{obj} and per-class positive weights for multi-label L_{rel}, respectively. Unlike previous work [44] that mixed up the multi-label and multi-class settings of L_{rel}, we insist on its multi-label setup with the purpose of maintaining a consistent benchmark for comparison. To sum up, the total training loss is computed as $L_{total} = L_{obj} + L_{rel}$.

4. Experiments

We first compare our EgoSG with existing reconstructed-based approaches on 3D scene graph prediction over localized objects in § 4.1, with ablation studies demonstrated in § 4.2. A qualitative analysis is presented in § 4.3.

Dataset Details. For method comparisons, we choose 3RScan dataset [39] with scene graph annotations released in 3DSSG [40]. In 3RScan, objects were annotated into 20 classes following the NYUv2 format [37]. For relationships, we follows SGFusion [44] to filter out the rare relationships, forming a 7-class predicate set, *i.e.* attached to, build in, connected to, hanging on, part of, standing on and supported by. The official data split is applied [40, 44].

Implementation Details We set $d_{node} = d_{edge} = 256$ and adopt instance normalization as normalization layers in our model. Our EgoSG is trained over 50 epochs, with a learning rate of 0.0003. During training, we adopt uniform temporal sampling to form a 20-frame clip of RGB-D videos, and we randomly sample 8192 points per point cloud frame. The training is completed over 8 Nvidia Tesla V100 16GB GPUs. Due to the large amount of GPU memory required for processing 4D point cloud sequences, we process one single sequence per GPU at any given time, and set gradient accumulation step to 8.

Approach	Modality	$\mathcal{R}$-free	$\mathcal{P}_{GT}$	$\mathcal{D}_{GT}$	$\mathcal{N}_{GT}$	Object		Predicate		Relationship	
						R@1	R@3	R@1	R@2	R@1	R@3
3DSSG [40]	3D Scan	✗	✓	✓	✗	59.37	84.87	81.02	97.58	45.86	62.48
SGFusion [44]											
w/ estimated pose	Sequence	✗	✗	✓	✓	24.37	41.03	53.44	81.31	0.04	14.75
w/ GT pose			✓	✓	✓	77.81	-	89.19	-	60.27	-
EgoSG-point (ours)											
w/ unposed RGBD frames	Sequence	✓	✗	✓	✗	69.00	89.86	76.18	87.21	91.58	97.27
w/ unposed RGB frames			✗	✗	✗	61.42	-	72.34	-	**89.32**	-

Table 1. 3D Scene Graph Prediction with localized objects (given CAIM). $\mathcal{R}$-free denotes system reliance on 3D scene reconstruction. $\mathcal{P}_{GT}$, $\mathcal{D}_{GT}$, and $\mathcal{N}_{GT}$ indicate whether the system takes as inputs GT camera pose (which is obtained from IMU sensor for reconstruction purpose), depth, or normal information, respectively. Note: the depth information in last row is estimated via MiDaS (2021). All approaches are benchmarked with PointNet as $\mathcal{F}_{\mathcal{B}}(\cdot)$.

4.1. Scene Graph Predictions on Localized Objects

Evaluation settings. We adopt the same evaluation metrics as used in 3DSSG [40], which are the top-n accuracies for object, predicate and relationship prediction. The evaluation of relationship triplets $< \mathtt{S}, \mathtt{R}, \mathtt{O} >$ measures the overall recognition of subjects $\mathtt{S}$, objects $\mathtt{O}$, and their in-between predicates $\mathtt{R}$, by multiplying their scores jointly.

Overall results. Our method achieved the best result in relationship triplet recognition, which adopt the joint recognition metrics over both objects and predicates. More importantly, ours actually requires the least assumptions, compared to reconstruction-based approaches, which require GT camera poses and normal.

Discussion with 3DSSG. To our surprise, as shown in Tbl. 1, our reconstruction-free EgoSG outperforms 3DSSG who takes entire 3D reconstructed scenes as inputs, in R@1 of object and relationship recognition. This validates our hypothesis that compared to 3D point-based reconstructions, unposed RGB-D sequence do contain sufficient 3D spatial cues, to support the estimation of 3D scene graphs via a reconstruction-free manner.

Discussion with SGFusion. As it is built upon InSeg SLAM, SGFusion [44] proposes to infer *segment-level* SGs from sequence inputs (rather than to generate *object-level* SGs). To make fair object-level comparisons with 3DSSG and ours, we thus aggregate its outputs with GT CAIMs to derive the *corresp.* object-level SGs.

We notice that ours may fall short in comparison with SGFusion when GT poses are provided. However, the graph prediction capability of SGFusion rely heavily on the GT camera pose to obtain scene reconstruction of acceptable quality, and it could restrict the hardware requirements in achieve egocentric scene understanding with mobile devices. This argument of SGFusion is verified when replacing GT poses with estimated ones from their built-in SLAM system, resulting in a *significant* performance drop, which can be easily outperformed by our EgoSG without any reliance of camera poses or SLAM techniques. Besides, SGFusion takes normal as additional inputs, while ours only requires xyz and RGB.

Method	Encoder $\mathcal{F}_{\mathcal{B}(\cdot)}$	SG Recognition	
		Obj.	Pred.
EgoSG-image	ResNet18 11.2M	26.79	57.00
	ResNet50 23.5M	40.14	71.31
	ViT-B-16 86.8M	36.94	77.09
	SwinV2-T 28.7M	41.48	70.97
EgoSG-point	PointNet 0.5M	69.00	76.18
	PointNet++ 6.4M	73.02	76.65
	PointNet++* 6.4M	74.65	77.69

Table 2. EgoSG with different perception settings, reported with their numbers of trainable parameters. R@1 is reported. * denotes the PointNet++ weights pretrained on CSC [17].

4.2. Ablative Studies

Different perception schemes. Considering the advanced 2D vision architecture with available pretrained weights, we could also equip EgoSG with image encoders and have it evaluated on 2D inputs instead. Unlike the reconstruction-based approaches [40, 44], our encoders could be flexibly switched when depth information becomes unavailable. Among a variety of image encoders with promising pretrained weights on ImageNet, we choose the popular ResNet18 and ResNet50 [14], and the cutting-edge transformer-based ViT-B-16 [7] and SwinV2-T [26]. We finetune them with RGB frames, removing depth information. For 3D perceptions, we add two more entries for analysis, namely training PointNet++ [30] from scratch, and finetuning PointNet++ which was pretrained in CSC [17].

As shown in Tbl 2, all point-based entries outperform their competitors by large margins, yet maintains much less numbers of trainable parameters. It validates that depth information encodes vital 3D priors for 3D scene graph recognition, and point encoders are capable of extracting them efficiently. Moreover, Tbl. 2 also show our overall performance could be further improved with better point encoder designs and pretrained weights. We choose the

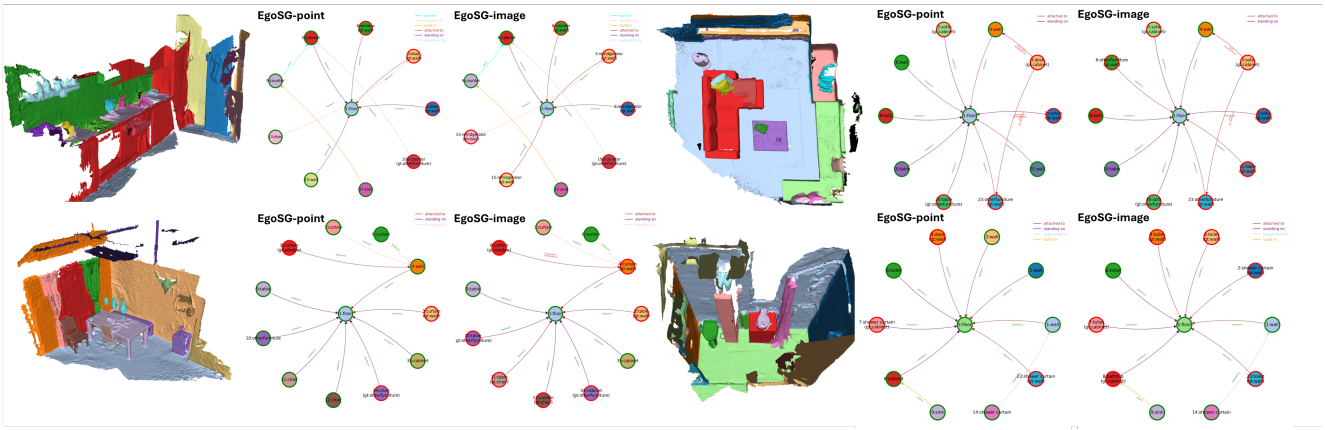

Figure 4. Qualitative analysis of 3D Scene Graph Prediction. EgoSG-point: ours with unposed RGB-D sequences; EgoSG-image: ours with RGB sequences. Zoom in for details. Note that the reconstructed scenes are only shown as a reference for scene graphs, while none of them are used in our approach. Objects and relations are outlined and annotated, respectively, as green if they are correctly recognized, otherwise in red with ground truth annotated. More detailed visualizations are shown in supplementary.

best SwinV2-T as our image encoder for benchmarking in Tbl. 1, while training PointNet from scratch is selected to make fair comparison with others in Tbl. 2.

Local and Global Graph Reasoning. We proposed to use $\mathcal{GR}_{local}$ to perform frame-based graph reasoning over instances observed at same point cloud frames, and proposed to use $\mathcal{GR}_{global}$ to jointly refine features of 3D relations and localized objects. As shown in Tbl. 3, although both graph reasoning designs could lead to overall improvement, we notice an immediate performance gain when $\mathcal{GR}_{local}$ is added to conduct message propagation over each ego-view frame. Combining our previous results in comparing with reconstruction-based SOTA approaches, it suggests that reasoning over ego-view contexts could be an feasible way to compensate the lack of 3D contexts when 3D reconstructions are not readily available.

Does `node2edge` build effective edge features? We choose different combinations of `max`, `avg` and `diff` operators, for ablative studies of `node2edge`. We compare the grouped aggregations with their non-grouped variants, where an extra learnable MLP ($\mathtt{G}\times d, d$) is added for channel mappings to produce edge features with $d_{edge} = d_{node} = d$. As shown in Tbl. 4, different $aggr(\cdot,\cdot)$ may contribute to object and predicate recognition differently, while mixing them is an intuitive way to take joint benefits. To make node and edge features consistent in channel numbers, the grouped aggregations are found to be more efficient than their non-grouped versions. Practically, we select `max`, `avg` and `diff` operators to form our aggregator group ($\mathtt{G} = 3$), and split node features into 3 groups to apply associated $aggr(\cdot,\cdot)$ individually, which help to largely improve the predicate recognition and slightly improve the node recognition.

Model ID	Graph Reasoning		SG Recognition	
	$\mathcal{GR}_{\mathtt{local}}$	$\mathcal{GR}_{\mathtt{global}}$	Object	Predicate
1			33.92	28.40
2	✓		42.29	38.53
3		✓	33.71	31.63
4	✓	✓	42.66	38.40

Table 3. Ablation studies on local and global graph reasoning effects ($\mathcal{GR}_{\mathtt{local}}$ and $\mathcal{GR}_{\mathtt{global}}$). We choose more challenging NYU40 classes for object recognition. mR@1 is reported.

G	$\{\mathtt{aggr}(\cdot,\cdot)\}$			Grouped	SG Recognition	
	`diff`	`max`	`avg`		Obj.	Pred.
1	✓			N/A	37.61	29.88
		✓			38.59	22.07
			✓		39.27	29.95
2	✓	✓		✓	39.50	31.49
				✗	34.89	33.05
	✓		✓	✓	38.75	38.10
				✗	36.39	33.79
		✓	✓	✓	40.00	29.18
				✗	36.97	28.53
3	✓	✓	✓	✓	**39.60**	**38.13**
				✗	33.93	32.68

Table 4. Ablation studies on grouped `node2edge` aggregator in deriving edge features from nodes. mR@1 metric is reported on both Object (Obj.) and Predicate (Pred.).

4.3. Qualitative Analysis

Here we show visualizations of scene graphs estimated by our EgoSG framework. Specifically, we tested our EgoSG with both point and image encoders, where we choose PointNet as point encoder and SwinV2-T as the image encoder, respectively. Correspondingly, the EgoSG-point is

deployed on RGB-D sequence data (i.e., 4D point cloud sequences), and EgoSG-image is deployed on RGB sequences (i.e., video data).

As shown in Fig.4, EgoSG-point achieves better results in 3DSGP task than EgoSG-image, especially in the recognition of 3D localized objects. Although we observed that EgoSG-image might tend to introduce more errors in the predicate recognition, compared to EgoSG-point, we are delighted to note that EgoSG-image manages to capture the overall graph layout as close as how EgoSG-point performs, which could be a promising signals in further investigation of 2D vision framework with purely 2D inputs, and seek to catch up with the 3D models which takes benefits from the depth information. This qualitative findings also validate our observation in Tbl. 1, where both EgoSG-point and EgoSG-image demonstrate comparably great capability in recognizing the relationship triplets.

4.4. Towards 3DSGP in the Wild

We present two future directions for the pursuit of 3DSGP in the wild, which performs 3DSGP with estimated depths or associations.

3DSGP with Estimated Depth. Due to our reconstruction-free design, our EgoSG could be flexibly adapted to even more challenging use-cases (i.e., when depth information becomes unavailable). Surprisingly, even with depth removed, our EgoSG designs would still outperform others (with depth information) in object and relationship triplet recognition. Noticeably, due to others' reconstruction-based design, their systems would completely fail when depth becomes unavailable, as shown in Tbl. 1. In detail, we further benchmark the 3DSGP performance with the depth images estimated by the off-the-shell estimator MiDaS [34]. As shown in Tbl. 5, although it falls short as expected on comparisons with EgoSG-point using GT depth, it still outperforms EgoSG-image with *no* depth by margin, which validates the contribution and necessity of depth information towards 3DSGP.

3DSGP with Estimated Association. Another direction in pursuing 3DSGP in the wild is to gradually relax the requirements of the given class-agnostic instance masks (CAIMs) and conduct the ultimate 3D scene graph generations (3DSGGen) without given localized objects. As explained in the main paper, our CAIM contains two indicators to pass the spatial-temporal object localization for 3DSGP task, namely the per-frame object detection and cross-frame object association.

Practically, we first relax the assumption of the given ground truth cross-frame association (GT-CFA) and compare two heuristic solutions. Specifically, we propose to apply the GT-CFA in the training of our EgoSG scene graph recognition network and, during inference, we estimate CFA on-the-fly via two intuitive object linking techniques, which are (1) hierarchical clustering and (2) computing objects' similarity scores. For hierarchical clustering, we link objects by thresholding via their embedding differences, while, for the latter option, we associate objects by calculating pairwise similarity scores. For evaluation, we modify VRD [27] to EgoVRD to make it compatible with 4D point cloud sequences, where an object is correctly localized if it achieves > 0.5 overlapping between its predicted video trajectory and its ground truth. The results are shown in Tbl. 6.

Approach	$\mathcal{D}$	Obj. R@1	Pred. R@1	Rel. R@1
EgoSG-point	GT	69.00	76.18	91.58
EgoSG-point	**Est.**	**61.42**	**72.34**	**89.32**
EgoSG-image	✗	41.48	70.97	84.23

Table 5. Comparisons of 3DSGP with estimated depths. $\mathcal{D}$ denotes whether the ground truth or estimated depths (via [34]) are adopted. We adopt the popular PointNet and SwinV2-T as the point encoder and image encoder for EgoSG-point and EgoSG-image, respectively.

Assumption	Det.	Asso.	EgoVRD (R@50)
given 3D localized objects	✓	✓	19.4
given 2D detected objects	✓	✗	4.5 (1) 6.3 (2)
3DSGGen in wild	✗	✗	N/A

Table 6. From top to bottom, with assumptions of frame-based detection (Det.) and cross-frame association (Asso.) gradually relaxed, the 3DSGP task could be converted to an ultimate 3DSGGen task in the wild step-by-step. (1) and (2) denote the results achieved by hierarchical clustering and pairwise similarity prediction to estimate the association between frames during inference.

5. Conclusion

In this paper, we discover the new modality of 3D scene graph prediction (3DSGP), i.e., from unposed RGB-D sequences, and propose the first reconstruction-free baseline for the 3DSGP task. Specifically, we propose to conduct local and global reasoning over graph representations on frame-level and video-level, respectively, to make use of the spatial contexts of the unposed videos. By leveraging the graph reasoning operations to capture the frame-based and video-based contexts, our method manage to outperform current state-of-the-arts reconstruction-based designs, which rely heavily on accurate point cloud reconstructions.

References

[1] Iro Armeni, Ozan Sener, Amir R. Zamir, Helen Jiang, Ioannis Brilakis, Martin Fischer, and Silvio Savarese. 3D semantic parsing of large-scale indoor spaces. In *ICCV*, 2016. 2

[2] Josep Aulinas, Yvan Petillot, Joaquim Salvi, and Xavier Lladó. The slam problem: a survey. *Artificial Intelligence Research and Development*, pages 363–371, 2008. 3

[3] Minjie Cai, Kris M Kitani, and Yoichi Sato. Understanding hand-object manipulation with grasp types and object attributes. In *Robotics: Science and Systems*. Ann Arbor, Michigan;, 2016. 2

[4] Christopher Choy, Jaesik Park, and Vladlen Koltun. Fully convolutional geometric features. In *ICCV*, 2019. 2

[5] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3D reconstructions of indoor scenes. In *CVPR*, 2017. 2

[6] Dima Damen, Teesid Leelasawassuk, and Walterio Mayol-Cuevas. You-do, i-learn: Egocentric unsupervised discovery of objects and their modes of interaction towards video-based guidance. *Computer Vision and Image Understanding*, 149:98–112, 2016. 2

[7] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 6

[8] Francis Engelmann, Martin Bokeloh, Alireza Fathi, Bastian Leibe, and Matthias Nießner. 3D-MPA: Multi-Proposal Aggregation for 3D Semantic Instance Segmentation. In *CVPR*, 2020. 2

[9] Alireza Fathi, Xiaofeng Ren, and James M Rehg. Learning to recognize objects in egocentric activities. In *CVPR 2011*, pages 3281–3288. IEEE, 2011. 2

[10] Alircza Fathi, Jessica K Hodgins, and James M Rehg. Social interactions: A first-person perspective. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1226–1233. IEEE, 2012. 2

[11] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jagannath Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *CVPR*, 2014. 4

[12] Raghav Goyal, Effrosyni Mavroudi, Xitong Yang, Sainbayar Sukhbaatar, Leonid Sigal, Matt Feiszli, Lorenzo Torresani, and Du Tran. Minotaur: Multi-task video grounding from multimodal queries. *arXiv preprint arXiv:2302.08063*, 2023. 2

[13] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18995–19012, 2022. 1, 2

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 6

[15] Ji Hou, Angela Dai, and Matthias Nießner. 3D-SIS: 3D Semantic Instance Segmentation of RGB-D Scans. In *CVPR*, 2019. 2

[16] Ji Hou, Angela Dai, and Matthias Nießner. RevealNet: Seeing Behind Objects in RGB-D Scans. In *CVPR*, 2020. 2

[17] Ji Hou, Benjamin Graham, Matthias Nießner, and Saining Xie. Exploring data-efficient 3d scene understanding with contrastive scene contexts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15587–15597, 2021. 6

[18] Shengyu Huang, Zan Gojcic, Mikhail Usvyatsov, Andreas Wieser, and Konrad Schindler. Predator: Registration of 3d point clouds with low overlap. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, pages 4267–4276, 2021. 2

[19] Li Jiang, Hengshuang Zhao, Shaoshuai Shi, Shu Liu, Chi-Wing Fu, and Jiaya Jia. PointGroup: Dual-Set Point Grouping for 3D Instance Segmentation. In *CVPR*, 2020. 2

[20] Justin Johnson, Ranjay Krishna, Michael Stark, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. Image retrieval using scene graphs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3668–3678, 2015. 2

[21] Justin Johnson, Agrim Gupta, and Li Fei-Fei. Image generation from scene graphs. 2018. 4

[22] Sebastian Koch, Pedro Hermosilla, Narunas Vaskevicius, Mirco Colosi, and Timo Ropinski. Sgrec3d: Self-supervised 3d scene graph learning via object-level scene reconstruction. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3404–3414, 2024. 2

[23] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. Visual Genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision (IJCV)*, 123(1):32–73, 2017. 2

[24] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2012. 5

[25] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *ICCV*, 2017. 5

[26] Ze Liu, Han Hu, Yutong Lin, Zhuliang Yao, Zhenda Xie, Yixuan Wei, Jia Ning, Yue Cao, Zheng Zhang, Li Dong, et al. Swin Transformer V2: Scaling Up Capacity and Resolution. *arXiv preprint arXiv:2111.09883*, 2021. 6

[27] Cewu Lu, Ranjay Krishna, Michael Bernstein, and Li Fei-Fei. Visual relationship detection with language priors. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*, pages 852–869. Springer, 2016. 2, 4, 8

[28] Yinyu Nie, Ji Hou, Xiaoguang Han, and Matthias Nießner. Rfd-net: Point scene understanding by semantic instance reconstruction. In *CVPR*, 2021. 2

[29] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3D classification and segmentation. *CVPR*, 2017. 3

[30] Charles R Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *NeurIPS*, 2017. 6

[31] Charles R Qi, Or Litany, Kaiming He, and Leonidas J Guibas. Deep hough voting for 3d object detection in point clouds. In *proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9277–9286, 2019. 2

[32] Zheng Qin, Hao Yu, Changjian Wang, Yulan Guo, Yuxing Peng, and Kai Xu. Geometric transformer for fast and robust point cloud registration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11143–11152, 2022. 2

[33] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 12179–12188, 2021. 6

[34] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(3), 2022. 8

[35] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In *NeurIPS*, 2015. 2

[36] Shaoshuai Shi, Xiaogang Wang, and Hongsheng Li. Pointrcnn: 3d object proposal generation and detection from point cloud. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 770–779, 2019. 2

[37] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from RGB-D images. *ECCV*, 2012. 5

[38] Takafumi Taketomi, Hideaki Uchiyama, and Sei Ikeda. Visual slam algorithms: A survey from 2010 to 2016. *IPSJ Transactions on Computer Vision and Applications*, 9(1):1–11, 2017. 3

[39] Johanna Wald, Armen Avetisyan, Nassir Navab, Federico Tombari, and Matthias Nießner. Rio: 3d object instance relocalization in changing indoor environments. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7658–7667, 2019. 5

[40] Johanna Wald, Helisa Dhamo, Nassir Navab, and Federico Tombari. Learning 3d semantic scene graphs from 3d indoor reconstructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3961–3970, 2020. 1, 2, 3, 4, 5, 6

[41] Xinlong Wang, Shu Liu, Xiaoyong Shen, Chunhua Shen, and Jiaya Jia. Associatively segmenting instances and semantics in point clouds. In *CVPR*, 2019. 2

[42] Yue Wang, Yongbin Sun, Ziwei Liu, Sanjay E. Sarma, Michael M. Bronstein, and Justin M. Solomon. Dynamic graph cnn for learning on point clouds. *ACM Transactions on Graphics (TOG)*, 2019. 5

[43] Ziqin Wang, Bowen Cheng, Lichen Zhao, Dong Xu, Yang Tang, and Lu Sheng. Vl-sat: Visual-linguistic semantics assisted training for 3d semantic scene graph prediction in point cloud. *arXiv preprint arXiv:2303.14408*, 2023. 2

[44] Shun-Cheng Wu, Johanna Wald, Keisuke Tateno, Nassir Navab, and Federico Tombari. SceneGraphFusion: Incremental 3D Scene Graph Prediction from RGB-D Sequences. In *Proceedings IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 1, 2, 3, 5, 6

[45] Danfei Xu, Yuke Zhu, Christopher B Choy, and Li Fei-Fei. Scene graph generation by iterative message passing. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5410–5419, 2017. 2

[46] Xu Yan, Zhihao Yuan, Yuhao Du, Yinghong Liao, Yao Guo, Zhen Li, and Shuguang Cui. Clevr3d: Compositional language and elementary visual reasoning for question answering in 3d real-world scenes. *arXiv preprint arXiv:2112.11691*, 2021. 5

[47] Bin Yang, Wenjie Luo, and Raquel Urtasun. Pixor: Real-time 3d object detection from point clouds. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 7652–7660, 2018. 2

[48] Jianwei Yang, Jiasen Lu, Stefan Lee, Dhruv Batra, and Devi Parikh. Graph R-CNN for scene graph generation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 670–685, 2018. 2

[49] Ryo Yonetani, Kris M Kitani, and Yoichi Sato. Recognizing micro-actions and reactions from paired egocentric videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2629–2638, 2016. 2

[50] Hao Yu, Fu Li, Mahdi Saleh, Benjamin Busam, and Slobodan Ilic. Cofinet: Reliable coarse-to-fine correspondences for robust pointcloud registration. *Advances in Neural Information Processing Systems*, 34:23872–23884, 2021. 2

[51] Hao Yu, Ji Hou, Zheng Qin, Mahdi Saleh, Ivan Shugurov, Kai Wang, Benjamin Busam, and Slobodan Ilic. Riga: Rotation-invariant and globally-aware descriptors for point cloud registration. *arXiv preprint arXiv:2209.13252*, 2022. 2

[52] Rowan Zellers, Mark Yatskar, Sam Thomson, and Yejin Choi. Neural Motifs: Scene graph parsing with global context. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 2

[53] Chaoyi Zhang, Jianhui Yu, Yang Song, and Weidong Cai. Exploiting edge-oriented reasoning for 3d point-based scene graph analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9705–9715, 2021. 1, 3, 5

[54] Yu Zhang, Junle Yu, Xiaolin Huang, Wenhui Zhou, and Ji Hou. Pcr-cg: Point cloud registration via deep explicit color and geometry. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part X*, pages 443–459. Springer, 2022. 2

[55] Zaiwei Zhang, Bo Sun, Haitao Yang, and Qixing Huang. H3dnet: 3d object detection using hybrid geometric primitives. In *European Conference on Computer Vision*, pages 311–329. Springer, 2020. 2

[56] Lin Zhao and Wenbing Tao. Jsnet: Joint instance and semantic segmentation of 3d point clouds. In *Proceedings of*

the AAAI Conference on Artificial Intelligence, pages 12951–12958, 2020. 2

[57] Yipin Zhou and Tamara L Berg. Temporal perception and prediction in ego-centric video. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4498–4506, 2015. 2