PUF: Plug-and-Play Uncertainty-Aware Fusion for Online 3D Scene Graph Generation

Yi Yang[1], Myrna Castillo[2,3], Bodo Rosenhahn[1], and Michael Ying Yang[3]*

[1] Leibniz Universität Hannover, Germany
[2] Istituto Italiano di Tecnologia, Italy
[3] University of Bath, UK

Abstract. Online 3D scene graph generation builds a persistent, structured representation of a scene by incrementally fusing 2D observations into a global 3D graph. Existing online methods treat this fusion as a fully deterministic pipeline, where we identify three sources of uncertainty that are overlooked: observation, 2D model, and 3D representation. We propose PUF: a Plug-and-play, Uncertainty-aware, and training-free Fusion framework. Scene graph node association is reformulated as a probabilistic likelihood over semantic and spatial factors, replacing binary accept/reject gates. Dirichlet evidence accumulation distributes class and relationship evidence across plausible candidates proportional to association likelihood. An optional class-conditional prior completes edges for sparsely or never co-observed object pairs. We instantiate PUF with both a 3D Gaussian and a 3D voxel backend and observe consistent improvements, demonstrating its ability to generalize across different representations. Experiments on the 3DSSG and ReplicaSSG benchmarks show that our method substantially outperforms existing approaches while maintaining real-time latency. These results establish uncertainty-aware fusion as a principled and effective paradigm for online 3D scene understanding. The source code is publicly available at `https://github.com/yyyyangyi/PUF`.

Keywords: Scene graph generation · Uncertainty · Plug-and-Play

1 Introduction

A 2D Scene Graph (SG) describes an image as a structured representation in which nodes correspond to detected objects and directed edges encode their pairwise relationships. A 3D scene graph lifts this abstraction into metric space. The task of online 2D-to-3D Scene Graph Generation (SGG) is to construct a global 3D SG incrementally from a stream of RGB-D frames. 3D SGs provide the high-level scene abstraction needed for downstream tasks such as embodied navigation [26], robotic manipulation [42], and spatial question answering [36]. Incremental construction supports real-time applications for which a complete

* Corresponding author.

scene reconstruction is unavailable or impractical [24], and lifting 2D predictions online into 3D leverages powerful 2D SGG models trained on richly annotated datasets [16] without the prohibitive cost of 3D SG annotation [31, 32, 39].

In the online setting, a global 3D SG is built by fusing a continuous stream of 2D observations, *e.g.*, a video sequence, into a persistent graph structure. With the camera's field of view covering only a fraction of the scene at any instant, every such observation is partial and inherently uncertain. Accurately fusing these partial observations is non-trivial, as it requires reasoning about *how much* to trust each observation. However, existing online methods discard this information by committing to hard decisions at every stage. We identify three distinct sources of uncertainty, as illustrated in Fig. 1. *(a)* Observations are uncertain as objects are truncated. Relationships are particularly subject to such noise since they can only be predicted when both endpoint objects are simultaneously well observed. *(b)* 2D SGG models are uncertain. They produce soft class and relationship distributions that cannot be collapsed to hard predictions without loss. *(c)* Object representation in 3D is approximate. Back-projecting a 2D bounding box through a single noisy depth frame without committing to a full 3D reconstruction yields uncertain 3D position and spatial extent.

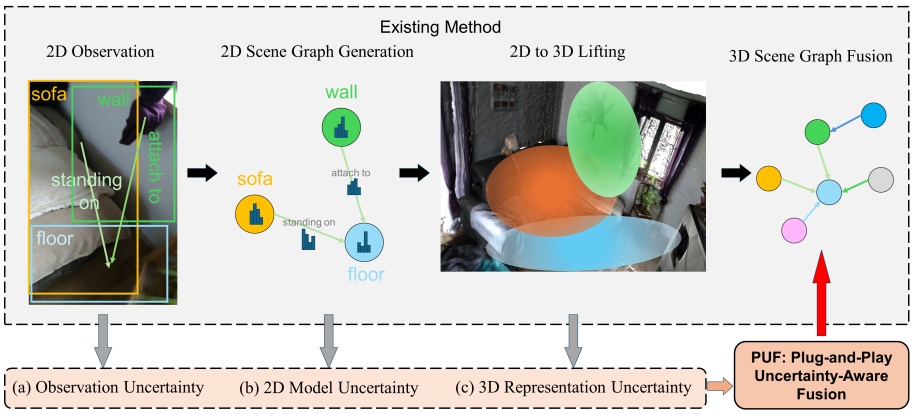

Fig. 1: Our method is aware of 3 types of uncertainty which are overlooked in existing online 3D SGG methods. (a) Uncertainty from partial observation in 2D; (b) Uncertainty encoded in softmax output from 2D SGG model; (c) Uncertainty in 3D lifting, e.g. 3D Gaussian with estimated depth in FROSS [11].

Offline 3D SGG methods rely on 3D point cloud input and global message-passing over the reconstructed graph, making them accurate but incompatible with real-time operation [6, 10, 15, 35]. Online methods [7, 9, 11, 13, 34, 40] process frames incrementally and achieve low latency, but treat the 2D-to-3D merging stage as a fully deterministic pipeline. This rigidity overlooks the sources of uncertainty above and therefore introduces three systematic limitations. *(a)* A detection is either merged or rejected based on its argmax label, which prevents

2D model uncertainty from propagating into 3D. *(b)* Thresholding on raw 3D distance or spatial overlap treats back-projected 3D positions as exact, ignoring the inherent ambiguity in depth-based localization. *(c)* Hard label transfer during merging does not redistribute evidence across candidate nodes, nor complete relational evidence for sparsely co-observed edges.

We propose **PUF**: a *plug-and-play*, *uncertainty-aware* and *training-free* fusion framework that compensates for these failures simultaneously. Our framework wraps any 2D SGG model that outputs soft class and relation probability distributions, and propagates these distributions directly into our proposed merging process. 2D model uncertainty is hence preserved throughout 3D reconstruction. Scene graph node association is formulated as a probabilistic problem that jointly accounts for semantic uncertainty and 3D representation uncertainty. Each observation is matched against existing nodes through a likelihood, combining class-distribution similarity with a spatial overlap term, so that uncertain geometry attenuates rather than vetoes an association. Upon association, both node semantics and edge labels are updated through Dirichlet accumulation that redistributes evidence across all plausible candidates proportional to association likelihood. An optional relation prior compensates for poorly observed or structurally unobservable edges. PUF is representation-agnostic, and we instantiate it with both a 3D Gaussian and a 3D voxel backend to demonstrate its ability to generalize across different representations. We evaluate PUF on the 3DSSG [32, 35] and ReplicaSSG [11, 27] benchmarks, where our method significantly and consistently outperforms the strongest baseline. Overall, our **main contributions** are as follows:

- A plug-and-play, training-free fusion framework for online 3D scene graph generation that propagates label and representation uncertainty throughout the merging pipeline without additional training.
- A probabilistic node association that jointly models 2D scene graph uncertainty and 3D representation uncertainty, replacing hard association gates with a principled likelihood-based formulation.
- A Dirichlet evidence accumulation scheme for both node semantics and edge labels, augmented with a class-conditional prior that completes edges for sparsely and never co-observed object pairs.
- State-of-the-art results on the 3DSSG and ReplicaSSG benchmarks, improving relationship Recall@1 by 18.1 points over the strongest online baseline on 3DSSG, while maintaining real-time latency at 15 ms per frame.

2 Related Work

Offline 3D Scene Graph Generation. Early 3D scene graph methods treat the task as a global inference problem over a complete 3D point cloud. Wald *et al.* [32] introduced the 3DSSG dataset and a graph convolutional network [14] baseline that encodes PointNet [20] features extracted from fully reconstructed meshes. Subsequent works further refine node or edge representations, yet operate identically on a pre-built point cloud graph. Wang *et al.* [33] improve long-tail

relation prediction via multi-modal knowledge distillation at training time, and Heo *et al*. [10] design a discriminative object feature encoder with contrastive pretraining. These methods produce dense and accurate segmentation at the point cloud level, but are incompatible with real-time operation.

Online 2D-to-3D Scene Understanding. A parallel line of work projects 2D predictions into 3D incrementally. General online lifting approaches demonstrate that rich 3D representations can be built without offline reconstruction. For example, ConceptGraphs [9] fuses 2D foundation-model detections into a 3D object graph via multi-view association. More recent open-vocabulary methods [29, 41] extend this to semantic segmentation masks, merging 2D outputs into persistent 3D structures using semantic similarity and spatial overlap.

In online 3D scene graph generation specifically, early methods rely on sparse point cloud reconstruction with SLAM technology [3,30]. SGFN [35] and its predecessor MonoSSG [34] infer local frame graphs from incremental segmentation and merge them deterministically into a global structure. Feng *et al*. [7] introduce hyperrectangle embeddings to debias relation prediction. Kim *et al*. [13] build a sparse semantic graph from RGB-D streams without running reconstruction, but rely on hard association rules. More recently, FROSS [11] pairs a real-time 2D SGG model with Gaussian back-projection for 3D node representation. This methodology achieves faster-than-real-time operation by removing the dependency on SLAM reconstruction, but the 2D-to-3D merging is fully deterministic. SCRSSG [40] pursue a post-hoc confidence rescoring strategy that re-weights predictions using node-to-node and node-to-edge co-occurrence statistics from a global graph. This method focuses on alleviating prediction imbalance, and does not address the three sources of uncertainty identified in this work. All the above methods adopt deterministic online fusion or global message passing, which are not uncertainty-aware.

Uncertainty-Aware 3D Perception and Scene Graphs. Bayesian filtering has a long history in sensor fusion. The Joint Probabilistic Data Association Filter [1, 21, 22] computes per-observation marginal association probabilities across all candidate tracks simultaneously, replacing hard data association with a principled probabilistic assignment. More recent research extends this method for joint landmark and pose estimation in SLAM [2, 28], and track-let prediction in multi-target tracking [5, 23]. These techniques address spatial representation uncertainty in tracking but do not model semantic class distributions or graph-structured relationships.

In the 2D scene graph community, several works have recognized that relationship prediction is inherently uncertain. Yang *et al*. [38] propose a plug-and-play Probabilistic Uncertainty Modeling module that represents each subject-object union region as a Gaussian distribution rather than a deterministic feature vector, enabling diverse and semantically ambiguous predictions. Similarly, Li *et al*. [17] adopt Bayesian classifier reparameterization to handle annotation noise in relationship labels. These approaches improve 2D prediction quality by preserving distribution-level information during model training, yet they address uncertainty only within the 2D inference stage.

Our work bridges these two lines. Inspired by the literature on probabilistic data association, we replace hard association gates with a probabilistic node association formulation that jointly models semantic and spatial uncertainty. We then preserve the full softmax output of the 2D SGG model throughout the fusion process via Dirichlet evidence accumulation, propagating prediction uncertainty directly into the 3D scene graph rather than discarding it at the 2D-to-3D boundary. We adopt voxel- and Gaussian-based 3D representations from state-of-the-art online methods [11, 29] and demonstrate that our PUF is effective on both, and is thus generalizable across 3D representations.

3 Method

3.1 Problem Definition

We formulate online 3D scene graph generation as the incremental construction of a directed graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where each node $v_i \in \mathcal{V}$ represents a 3D object with semantic label $c_i \in \{1, \ldots, C\}$ and 3D position $\boldsymbol{\mu}_i$, and each directed edge $(i, j) \in \mathcal{E}$ carries a relationship label $r_{ij} \in \{1, \ldots, R\}$. Given a stream of RGB-D frames $\{(\mathbf{I}_t, \mathbf{D}_t, \mathbf{T}_t)\}_{t=1}^{T}$, where $\mathbf{T}_t$ is the known camera pose, the goal is to build $\mathcal{G}$ incrementally without retaining the full video history.

3.2 Framework Overview

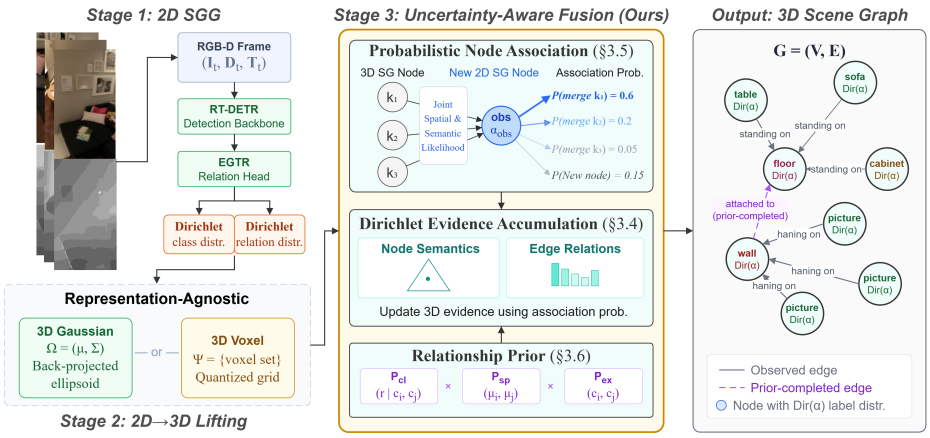

Fig. 2: Overview of our proposed PUF. **Stage 1:** An RGB-D frame is processed by a 2D SGG model. We use Dirichlet-parameterization for object and relation label distributions. **Stage 2:** Each detection is lifted to 3D via a Gaussian or voxel backend (representation-agnostic). **Stage 3:** Our uncertainty-aware fusion stage performs probabilistic node association and Dirichlet evidence accumulation, with an optional relationship prior for sparsely observed edges. **Output** 3D scene graph has nodes and edges carry full Dirichlet distributions over their respective label spaces.

Figure 2 illustrates our uncertainty-aware fusion framework. At each frame, a 2D SGG model produces soft class and relation distributions, which are lifted to 3D observation nodes via either a Gaussian or voxel backend. We use the Dirichlet distribution to accumulate object and relationship evidence on the global 3D SG (Sec. 3.4). Probabilistic node association (Sec. 3.5) computes a joint association likelihood between a new observation and all existing global nodes, which replaces hard binary gates, and redistributes 2D evidence to 3D accumulators. An optional relationship prior (Sec. 3.6) completes sparsely observed edges.

Our framework is *plug-and-play*: it wraps any 2D SGG model that outputs soft class and relation distributions without modifying model architecture; and is *training-free*: it requires no additional neural-network training.

3.3 Preliminaries

2D scene graph model. We use RT-DETR detection backbone [4, 18] with EGTR [12] for 2D SGG, following FROSS [11]. RT-DETR produces object bounding boxes and feature representations at real-time latency. EGTR predicts pairwise relationship logits via an edge-conditioned transformer.

PUF departs from existing hard-fusion methods by consuming the *full* softmax outputs: the per-object class probability $\hat{p}^c \in \Delta^{C-1}$ (the (C-1)-simplex) and the per-pair relation probability $\hat{p}^r \in \Delta^{R-1}$, rather than only their argmax. In this way, the prediction uncertainty is propagated into the fusion layer.

3D representation — Gaussian. Each detected bounding box is back projected to a 3D Gaussian $\Omega_i = (\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$ using per-pixel depth and the camera intrinsics/extrinsics. The mean $\boldsymbol{\mu}_i$ is the depth-weighted centroid of the back-projected pixel cloud inside the box; the covariance $\boldsymbol{\Sigma}_i$ is the empirical covariance of the cloud. Implementation is identical to FROSS [11] for fair comparison; the details are provided in Supplementary Sec. B.

3D representation — Voxel. Each bounding box is converted to a discrete voxel set, inspired by OnlineAnySeg [29]. Depth-filtered pixels are first back-projected to the world coordinates and then quantized to a uniform grid of resolution δ. Depth filtering retains only pixels whose depth deviates by at most ε_d from the box-centre depth, isolating foreground from background clutter. The full lifting and merging procedure is detailed in Supplementary Sec. B.

3.4 Dirichlet Representation for Nodes and Edges

We use the Dirichlet distribution to represent the semantic labels of SG nodes and edges. It naturally replaces one-hot based label with a distribution over all categories, therefore carries 2D SGG model uncertainty. Its posterior update is straightforward and facilitates the evidence accumulation in the merging step.

Node class model. The class probability $\hat{p}_k^c$ of global 3D node $\mathcal{V}_k$ is modeled by a Dirichlet:

$$\hat{p}_k^c \sim \mathrm{Dir}(\boldsymbol{\alpha}_k). \tag{1}$$

where $\boldsymbol{\alpha}_k \in \mathbb{R}^C$ is the evidence accumulator for $\mathcal{V}_k$ over the semantic class label space, and interpreted as the concentration parameters of the Dirichlet distribution. Note that $\boldsymbol{\alpha}_k$ is real-valued and not normalized to facilitate association update (Eq. (7)). The normalized Dirichlet posterior mean $\bar{\boldsymbol{\alpha}}_k = \boldsymbol{\alpha}_k / \|\boldsymbol{\alpha}_k\|_1$ gives a proper distribution over class labels, and the hard class prediction for the final output is $c_k = \arg\max \bar{\boldsymbol{\alpha}}_k$.

Edge relation model. Similarly to nodes, for each directed edge $\mathcal{E}_{ij}$, we model the relationship class probability $\hat{p}^r_{ij}$ as:

$$\hat{p}^r_{ij} \sim \mathrm{Dir}(\boldsymbol{\phi}_{ij}). \tag{2}$$

where $\boldsymbol{\phi}_{ij} \in \mathbb{R}^R$ is the accumulator for edge $\mathcal{E}_{ij}$. The final predicted relationship is the posterior mean argmax: $r_{ij} = \arg\max \bar{\boldsymbol{\phi}}_{ij}$.

3.5 Probabilistic Node Association

Association likelihood. Between each pair of newly observed node *obs* and an existing node k in global 3D SG, we compute a factored likelihood:

$$L[obs,\, k] = L_{\mathrm{sp}}(obs,\, k) \;\cdot\; L_{\mathrm{se}}(obs,\, k). \tag{3}$$

The spatial factor, unlike hard thresholds in existing methods, enters the likelihood as a continuous term that modulates association probabilities, and preserves as well as propagates 3D representation uncertainty into fusion. Its specific form depends on the representation:

$$L_{\mathrm{sp}}(obs,\, k) = \begin{cases} BC(\Omega_{obs}, \Omega_k) & \text{(Gaussian)}, \\ |\Psi_{obs} \cap \Psi_k| \;/\; |\Psi_{obs}| & \text{(Voxel)}. \end{cases} \tag{4}$$

For the Gaussian backend, BC denotes the Bhattacharyya coefficient and measures spatial overlap between two Gaussians, and $BC = 1 - H^2$ where H is the Hellinger distance [19]. For the voxel representation, we use the containment score as defined in Eq. (4) to measure spatial overlap between an observation with voxel set Ψ_{obs} and a global node with accumulated set Ψ_k.

The semantic factor compares the Dirichlet posterior means of the observation and the global node via Jensen–Shannon divergence (JSD):

$$L_{\mathrm{se}}(obs,\, k) = \exp\!\left(- \frac{JSD(\hat{p}^c_{obs},\, \bar{\boldsymbol{\alpha}}_k)}{\sigma_{se}} \right). \tag{5}$$

The bandwidth σ_{se} controls semantic selectivity. A larger value increases the semantic factor and encourages more divergent $(\hat{p}^c_{obs},\, \bar{\boldsymbol{\alpha}}_k)$ pairs to merge.

Probabilistic marginals of node association. Inspired by the Joint Probabilistic Data Association Filter [1, 22], we model the probability that the observation associates with global node k as $\beta[obs,\, k]$. $\beta[obs,\, k]$ is obtained by

marginalizing the likelihood vector $\{L[obs,\, k]\}_{k\in\mathcal{K}}$ together with a prior probability that *obs* is a new object:

$$\beta[obs,\, k] = \frac{L[obs,\, k]}{Z}, \quad \beta_{birth} = \frac{\lambda_{birth}}{Z}, \quad Z = \lambda_{birth} + \textstyle\sum_{k\in\mathcal{K}} L[obs,\, k]. \tag{6}$$

where λ_{birth} is a birth term encoding the prior probability of observing a new object. By construction $\beta_{birth} + \sum_k \beta[obs,\, k] = 1$. If $\beta_{birth} > 0.5$, the observation spawns a new global node; otherwise it associates with existing nodes. To ensure real-time computation we adopt per-observation independent normalization Eq. (6) rather than the expensive joint hypothesis enumeration as in the original Joint Probabilistic Data Association Filter [1]. This is also due to the fact that DETR's non-maximum suppression mechanism [18] already prevents two detections of the same object within a single frame. We provide a detailed analysis in Supplementary Sec. A.

Association update. If the new observation merges to the global 3D SG ($\beta_{birth} \leq 0.5$), we apply semantic and spatial updates as follows:

$$\text{Semantic:} \quad \boldsymbol{\alpha}_k \;\leftarrow\; \boldsymbol{\alpha}_k + \beta[obs,\, k] \cdot \hat{p}^{c}_{obs}, \quad \forall\, k : \beta[obs,\, k] \geq \beta_{min}, \tag{7}$$

$$\text{Spatial:} \quad k^* = \arg\max_k \; \beta[obs,\, k]. \tag{8}$$

where β_{min} is an evidence cut-off threshold. Class evidence across all candidates with non-negligible $\beta[obs,\, k]$ is distributed to the global node k, which lets node k's Dirichlet accumulator absorb uncertainty across near-tie associations. 3D representation (Gaussian mean, covariance/voxel union) is updated only for k^* to avoid contaminating or drifting the positions of non-target nodes: a predicted label can be 0.6 chair - 0.4 table, but the object is not part chair, part table, so semantics are softly merged while spatial occupation is not.

A soft relationship transfer is applied to the SG edges. Let $\hat{p}^{r}_{obs,j}$ denote the 2D relation prediction between the observation node and global node j in the current frame. The relationship evidence is redistributed to the same candidates and with the same weights as the semantic update:

$$\boldsymbol{\phi}_{kj} \;\leftarrow\; \boldsymbol{\phi}_{kj} + \beta[obs,\, k] \cdot \hat{p}^{r}_{obs,j}, \quad \boldsymbol{\phi}_{ik} \;\leftarrow\; \boldsymbol{\phi}_{ik} + \beta[obs,\, k] \cdot \hat{p}^{r}_{i,obs}, \tag{9}$$

where the first term covers the observation acting as subject and the second as object. This update preserves consistency between the class and relation accumulators, giving evidence accumulation to poorly observed nodes and edges. After *obs* is merged to existing nodes, it is invalidated.

Otherwise, if $\beta_{birth} > 0.5$, then node *obs* is a newly observed object and initializes a new 3D node from the soft 2D outputs:

$$\boldsymbol{\alpha}_{obs} \leftarrow \hat{p}^{c}_{obs}. \tag{10}$$

$$\boldsymbol{\phi}_{i,obs} \leftarrow \hat{p}^{r}_{i,obs}\,, \quad \boldsymbol{\phi}_{obs,j} \leftarrow \hat{p}^{r}_{obs,j}\,. \tag{11}$$

Then *obs* is appended to the list of global 3D nodes.

3.6 Relationship Prior for Edge Evidence Enhancement

An edge in the SG is only well-observed if both of its associated nodes are well-observed, making edges more susceptible to observation noise. To address this, we leverage Dirichlet conjugacy and propose an informative prior. For a pair with object classes c_i, c_j and 3D centroids $\boldsymbol{\mu}_i, \boldsymbol{\mu}_j$, the prior factorizes as:

$$\boldsymbol{\phi}_{ij}^{prior} = P_{cl}(r \mid c_i, c_j) \cdot P_{sp}(\boldsymbol{\mu}_i, \boldsymbol{\mu}_j) \cdot P_{ex}(c_i, c_j). \tag{12}$$

Class-conditional prior $P_{cl}(r \mid c_i, c_j)$ is precomputed from training-set annotation counts with Laplace smoothing:

$$P_{cl}(r \mid c_i, c_j) = \frac{n_{ij}^r + \varepsilon}{\sum_{r'} n_{ij}^{r'} + R\varepsilon}, \tag{13}$$

where n_{ij}^r counts relation r between object-class pair (c_i, c_j) in training data, and Laplace smoothing $\varepsilon = 0.1$. This requires a single offline pass over annotations without gradient-based training.

Spatial prior $P_{sp}(\boldsymbol{\mu}_i, \boldsymbol{\mu}_j)$ is computed online from 3D positions, decaying the prior as spatial distance increases: $P_{sp}(\boldsymbol{\mu}_i, \boldsymbol{\mu}_j) = e^{-\frac{|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j|}{2}}$. The combination of the class and spatial priors is normalized before multiplying with $P_{ex}(c_i, c_j)$.

Existence gate $P_{ex}(c_i, c_j)$ is the fraction of training-set pairs of class (c_i, c_j) that carry any annotated relation. It scales the prior down for class pairs that rarely interact, suppressing false-positive completions for rarely co-occurring class pairs.

For observed edges, the final distribution is the Bayesian posterior:

$$\hat{p}_{ij}^r \sim \mathrm{Dir}(\boldsymbol{\phi}_{ij}^{post.}), \qquad \boldsymbol{\phi}_{ij}^{post.} = \boldsymbol{\phi}_{ij}^{prior} + \boldsymbol{\phi}_{ij}, \tag{14}$$

where $\boldsymbol{\phi}_{ij}$ is the accumulated evidence by Eq. (9). For unobserved edges, the prior alone is used if $\max(\boldsymbol{\phi}_{ij}^{prior}) > 0.5$. In the case no prior is used, the relationship label is predicted from Eq.(2).

4 Experiments

4.1 Experimental Setup

Datasets. The *3DSSG* dataset [32] extends 3RScan [31], comprising 1,482 RGB-D indoor scans with dense instance segmentation and directed inter-object relationship annotations. We follow the standard protocol of [11,32,34,35], mapping 160 object categories to the 20 NYUv2 classes [25] and retaining the 7 most frequent predicate types. After mapping, the dataset contains 21,974 annotated object instances and 16,324 inter-object relationships. We use the official training, validation, and test splits throughout.

ReplicaSSG [11] extends the Replica dataset [27] with inter-object relationship annotations across 18 photorealistic indoor scenes. Its label space is adopted from Visual Genome [16] and contains 33 object and 9 relationship categories, facilitating zero-shot transfer from pretrained 2D SG models. Of the 18 scenes, 7 serve as a validation set and 11 as the test set; there is no training split.

Evaluation Metrics. We adopt the recall-based protocol of [11, 34]. *Object recall* measures the fraction of ground-truth instances matched to a correctly classified node. *Predicate recall* measures correct relationship classification among pairs whose endpoints are both detected. *Relationship recall* requires both endpoints to be correctly detected *and* the predicate correctly classified. We additionally report *mean recall* (mRecall), which averages per-class recall to mitigate the severe predicate class imbalance present in both datasets [11, 34, 35].

Implementation Details. We use EGTR [12] with RT-DETRv2-M [18] as the *2D scene graph model*, following FROSS [11] for fair comparison. For 3DSSG experiments, the model is trained on 2D scene graphs extracted from the 3DSSG training split. For ReplicaSSG experiments, we use a model trained on Visual Genome [16] with no additional fine-tuning on Replica data. At inference, object detections below a confidence threshold of 0.7 are discarded and the top-10 pairwise relation scores per frame are retained, matching the FROSS configuration. All models are evaluated on an Intel XEON Gold 6534 processor with a single CPU core, and an NVIDIA L40S GPU.
3D representations. For the Gaussian backend, the back-projection and covariance estimation follow FROSS [11] without modification. For the voxel backend, we quantize depth-filtered back-projected points to a grid of $\delta = 2\,\text{cm}$ and retain only pixels whose depth deviates by at most $\varepsilon_d = 0.3\,\text{m}$ from the bounding box center depth for foreground isolation.
Hyperparameters. For Probabilistic Node Association and on both 3DSSG and ReplicaSSG, the semantic bandwidth is set to $\sigma_{se} = 0.3$; the birth density is $\lambda_{birth} = 0.4$; the minimum association weight for soft updates is $\beta_{min} = 0.05$. For the Relationship Prior, the class-conditional prior P_{cl} is precomputed offline from the 3DSSG training split, and no prior is used for ReplicaSSG since it does not have a training set. All hyperparameters are selected by grid search on the respective validation splits with relationship recall as the objective. Detailed analysis is provided in Section 4.3 for the 3D Gaussian representation and in Supplementary Sec. D.2 for the voxel.

4.2 Quantitative Results and Comparison

Tables 1 and 2 present performance and latency comparisons on the 3DSSG and ReplicaSSG benchmarks, respectively. On 3DSSG, our PUF-Gaussian variant achieves the highest relationship recall of 46.0%, improving over FROSS by 18.1 points while operating at 15 ms per frame. Compared to the SGG recall gain, the additional fusion latency is negligible, and our method remains faster than all RGB-D+SLAM baselines by an order of magnitude. This further substantiates the advantages of lifting scene graphs from 2D images over direct point cloud reasoning. On ReplicaSSG, where no relationship prior is used due to the absence of a training split, PUF-Gaussian variant consistently outperforms FROSS across all metrics. This demonstrates that our proposed fusion framework itself provides reliable gains, independently of the prior. This is further supported by

our ablation studies Tab. 3 and 4 in Sec. 4.3. Our Voxel variant further demonstrates that the improvements are representation-agnostic: consistent gains over FROSS are observed with both backends across both benchmarks, confirming that the framework generalizes beyond any single 3D representation choice.

Table 1: Performance comparison of 3D SGG methods using different input modality on the 3DSSG test set. The end-to-end latency for baseline methods are obtained without the environmental mapping step. The best and second-best results are highlighted in **bold**, and underline, respectively. Latencies are reported in milliseconds.

Method	Recall (%)↑			mRecall (%)↑		Latency↓	Input Modality
	Rel.	Obj.	Pred.	Obj.	Pred.		
3DSSG [32]	12.9	37.4	22.0	26.2	14.4	-	Point Cloud
VL-SAT [33]	23.5	53.7	28.9	42.1	25.3	-	Point Cloud
OCRL [10]	25.2	58.5	30.1	49.6	27.1	-	Point Cloud
SGFN [35]	22.0	51.6	27.5	37.7	24.0	245	RGB-D+SLAM
MonoSSG [34]	23.3	53.8	28.4	43.8	26.6	283	RGB-D+SLAM
SCRSSG [40]	25.7	61.1	27.6	60.5	27.8	350	RGB-D+SLAM
VGfM [8]	19.6	50.0	20.4	34.8	11.0	379	RGB
IMP [37]	19.7	49.5	20.9	34.7	13.8	-	RGB-D
Kim *et al.* [13]	9.1	59.0	7.1	51.0	8.0	454	RGB-D
FROSS [11]	27.9	62.5	33.2	63.8	18.1	**13**	RGB-D
PUF-Voxel	40.3	65.5	46.1	64.1	21.8	31	RGB-D
PUF-Gaussian	**46.0**	**69.7**	**51.4**	**65.8**	**28.2**	15	RGB-D

Table 2: Performance comparison of 3D SGG methods on the ReplicaSSG test set and the end-to-end latency. The best and second-best results are highlighted in **bold**, and underline, respectively. Latencies are reported in milliseconds.

Method	Recall (%)↑			mRecall (%)↑		Latency↓
	Rel.	Obj.	Pred.	Obj.	Pred.	
FROSS [11]	22.5	26.2	28.0	29.1	20.6	**14**
PUF-Voxel	22.4	27.9	26.9	30.2	17.9	30
PUF-Gaussian	**25.3**	**31.0**	**35.6**	**33.7**	**26.2**	16

4.3 Ablation Study

Module Effectiveness. Tables 3 and 4 report ablation results on 3DSSG and ReplicaSSG, respectively. We ablate three components: Dirichlet node representation, Dirichlet edge representation, and the relationship prior, across both the

Gaussian and voxel backends. Probabilistic association is only enabled if the Dirichlet node representation is used.

Table 3: Ablation study of model components with recall and mean recall metrics on 3DSSG test set. Components Node and Edge means using Dirichlet representation for SG nodes and edges, respectively. Latencies are reported in milliseconds.

Method	Component			Gaussian						Voxel					
				Recall@1↑			mRecall@1↑		Latency↓	Recall@1↑			mRecall@1↑		Latency↓
	Node	Edge	Prior	Rel.	Obj.	Pred.	Obj.	Pred.		Rel.	Obj.	Pred.	Obj.	Pred.	
FROSS				27.9	62.4	33.2	63.8	18.1	**13.2**	23.7	61.6	29.5	62.6	16.5	**28.9**
PUF	✓			31.3	69.0	36.6	65.8	20.9	14.6	28.6	64.9	33.4	64.1	17.8	31.0
	✓	✓		33.9	69.0	39.2	65.8	22.4	14.7	31.3	64.9	35.2	64.1	18.3	31.3
		✓		27.9	62.9	33.7	64.3	21.9	13.5	24.2	61.8	30.6	62.8	18.0	29.2
		✓	✓	41.4	62.9	47.8	64.3	27.4	13.5	37.2	61.8	40.0	62.8	19.6	29.2
	✓	✓	✓	**46.0**	**69.7**	**51.5**	**65.8**	**28.2**	14.8	**40.3**	**65.5**	**46.1**	**64.1**	**21.8**	31.3

On 3DSSG, adding the Dirichlet node representation alone improves relationship recall by 3.4 points over FROSS under the Gaussian backend, confirming that propagating 2D class uncertainty into the fusion layer benefits node association and semantic accumulation. Incorporating Dirichlet edge representation and enabling probabilistic association further improves relationship recall to 33.9%, as soft relational evidence is redistributed across plausible candidates rather than collapsed to hard votes. On top of the probabilistic fusion, the relationship prior additionally contributes a large gain: enabling it alongside the edge Dirichlet representation lifts relationship recall to 41.4%, as it completes evidence for sparsely and never co-observed pairs. The full model combining all three components achieves 46.0% relationship recall. Consistent trends are observed with the voxel backend, confirming that each component's contribution is representation-agnostic.

Table 4: Ablation study of model components with recall and mean recall metrics on ReplicaSSG test set. Components Node and Edge means using Dirichlet representation for SG nodes and edges, respectively. Latencies are reported in milliseconds.

Method	Component		Gaussian						Voxel					
			Recall@1↑			mRecall@1↑		Latency↓	Recall@1↑			mRecall@1↑		Latency↓
	Node	Edge	Rel.	Obj.	Pred.	Obj.	Pred.		Rel.	Obj.	Pred.	Obj.	Pred.	
FROSS			22.5	26.2	28.0	29.1	20.6	**13.8**	18.6	21.3	21.1	23.4	13.5	**28.8**
PUF	✓		23.1	**31.0**	30.5	**33.7**	22.5	15.4	21.7	**27.9**	24.5	**30.2**	16.8	29.5
		✓	23.0	26.1	29.7	28.8	23.7	15.2	20.8	25.4	24.3	26.8	16.9	29.3
	✓	✓	**25.3**	**31.0**	**35.6**	**33.7**	**26.2**	15.6	**22.4**	**27.9**	**26.9**	**30.2**	**17.9**	29.7

On ReplicaSSG, where no relationship prior is used, the node and edge Dirichlet components each independently improve over FROSS, and combining them

with the probabilistic association yields the best performance of 25.3% relationship recall under the Gaussian backend. This consistent improvement in the absence of any prior corroborates that the core uncertainty-aware fusion framework is effective on its own, and that the gains observed on 3DSSG are not solely attributable to the prior.

Hyperparameter choices. Table 5 reports the sensitivity of the two main hyperparameters on the 3DSSG (w/o prior) and ReplicaSSG validation sets under the Gaussian backend. The semantic bandwidth σ_{se} controls the selectivity of the semantic likelihood factor (Eq. (5)): values below 0.3 over-penalize distributional differences and fragment semantically similar nodes, while values above 0.3 over-merge distinct objects, degrading both object and relationship recall. The birth density λ_{birth} governs the prior probability of spawning a new global node (Eq. (6)). A large λ_{birth} under-merges the graph, leaving object recall roughly intact, but depressing predicate and relationship recall as relational evidence is scattered across duplicate nodes. A small λ_{birth} over-merges distinct objects, harming all metrics. Notably, the same setting (σ_{se}=0.3, λ_{birth}=0.4) is optimal on both datasets, indicating that the framework is not sensitive to precise tuning. Voxel-backend hyperparameter analysis is provided in Supplementary Sec. D.2 and exhibits consistent trends.

Table 5: Performance comparison on (w/o prior) 3DSSG/ReplicaSSG validation sets with different semantic bandwidth σ_{se} and birth density λ_{birth}.

Dataset	Recall(%)↑	σ_{se} 0.2	0.25	0.3	0.35	0.4	λ_{birth} 0.3	0.35	0.4	0.45	0.5
3DSSG	Rel.	32.4	33.7	**34.2**	31.0	29.3	30.9	31.7	**34.2**	**34.2**	**34.2**
	Obj.	65.1	66.5	**66.9**	65.8	64.0	65.3	66.0	66.9	67.7	**68.1**
	Pred.	36.6	37.4	**38.2**	35.7	34.6	34.8	35.6	**38.2**	37.8	36.5
ReplicaSSG	Rel.	16.1	16.3	**17.6**	14.7	13.6	12.8	13.9	**17.6**	14.6	14.6
	Obj.	23.3	24.5	**24.5**	23.6	22.2	21.3	21.7	**24.5**	23.0	23.0
	Pred.	20.8	22.0	**22.4**	19.9	18.3	19.0	19.4	**22.4**	20.3	19.1

Relationship Prior on 3DSSG. The 3DSSG dataset presents a notable co-visibility challenge: despite an average of ∼245 frames per scan, 25.3% of ground-truth object pairs accumulate fewer than 10 joint frame observations, and 2.7% never co-appear at all. Figure 3 stratifies relationship recall by the number of 2D co-observations per ground-truth triplet. Two findings stand out. First, even without the relationship prior, our uncertainty-aware fusion framework ("w/o Prior") consistently outperforms FROSS, confirming that probabilistic node association and soft evidence redistribution alone improve recall for both well- and

poorly-observed edges. This is further confirmed by a quantitative breakdown in Supplementary Sec. C. Second, the relationship prior yields its largest gains in the sparsest bins, and its marginal contribution diminishes monotonically as co-observation count increases. This suggests that the prior behaves correctly: the prior fills in evidence that the observation stream cannot provide, while deferring to accumulated observations when they are plentiful.

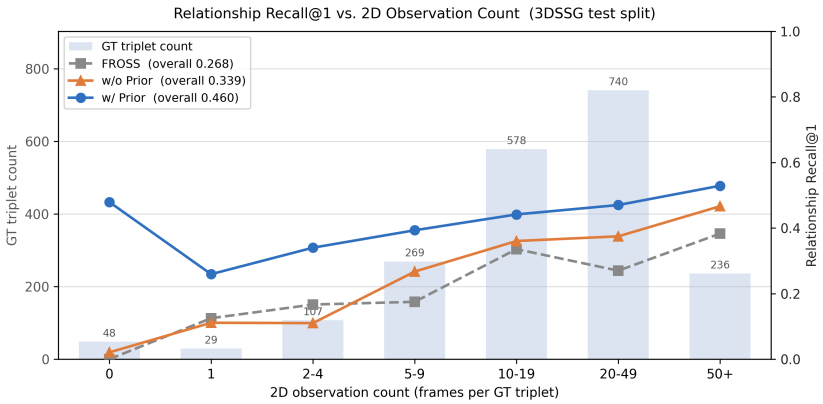

Fig. 3: Relationship recall@1 under different 2D observation counts on 3DSSG test set. W/o prior denotes our full fusion framework without using relationship prior.

4.4 Qualitative Results

Figure 4 compares a representative scene from 3DSSG under FROSS and our method using the Gaussian backend. FROSS misclassifies several rare nodes in the `other` class (marked in red) and misses the rare predicate `hanging on` between `curtain` and `wall`, defaulting instead to the dominant `standing on`. Our method correctly recovers the object and the rare relationship, illustrating the benefit of soft evidence accumulation over hard label transfer. Additional qualitative outputs on ReplicaSSG and with the voxel backend are provided in Supplementary Sec. E, which show consistent results.

Because the Dirichlet representation maintains a full distribution over labels, we can additionally visualize per-node and per-edge predictive uncertainty via the normalized entropy $H/H_{\max}$, where H_{max} is the entropy over an even Dirichlet distribution. As shown in the right-hand graph of Fig. 4, correctly classified nodes and edges exhibit low normalized entropy, while the few remaining errors, such as the `standing on` edge and the misclassified `counter`, carry notably higher uncertainty. This calibration property is a direct consequence of propagating the three sources of uncertainty identified in Sec. 1 through the fusion pipeline, and could be leveraged by downstream consumers to filter or re-query unreliable predictions.

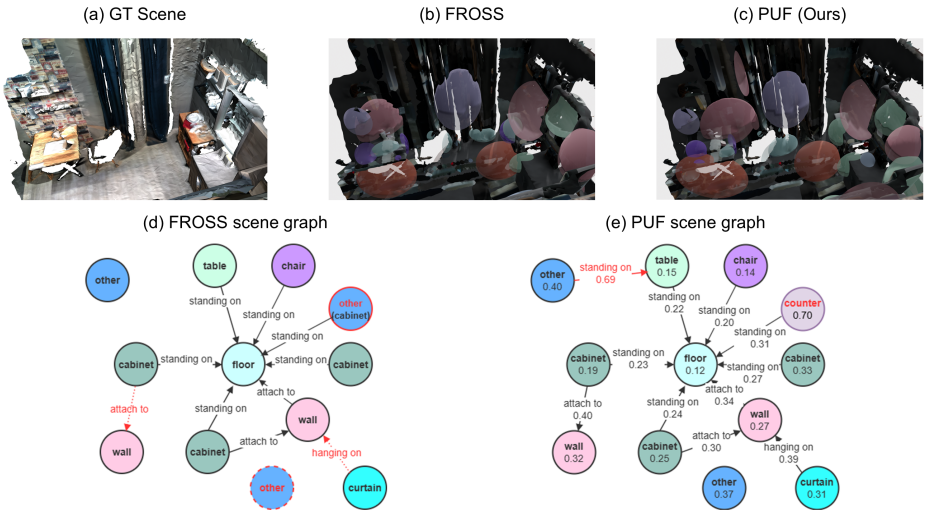

Fig. 4: Qualitative between our PUF and FROSS [11] on 3DSSG dataset with Gaussian 3D representation. Node colors correspond to the respective scene graphs. False and missing predictions are marked in red, with missing predictions in dashed line.

5 Conclusion

We presented a plug-and-play, training-free fusion framework, **PUF**, for online 3D scene graph generation. Our framework replaces the deterministic merging pipelines of prior works with uncertainty-aware probabilistic association and Dirichlet evidence accumulation. By preserving 2D prediction distributions, modeling 3D representation uncertainty in node association, and optionally completing sparsely observed edges with a class-conditional prior, our framework yields substantial gains in relationship recall on both the 3DSSG and ReplicaSSG benchmarks while adding negligible latency. Its representation-agnostic design is validated with both Gaussian and voxel backends, and makes it readily applicable as a drop-in wrapper around any 2D scene graph model that outputs soft class and relation distributions.

Acknowledgements

This work has been supported by the Centre for Spatial Intelligence (RCSI) at University of Bath, the European Union under grant agreement no. 101136006-XTREME, the European Innovation Council under grant agreement no. 101257536-CEREBRIS, the MWK of Lower Saxony within Hybrint (VWZN4219) and LCIS (VWZN4704), the DFG under Germany's Excellence Strategy within the Cluster of Excellence PhoenixD (EXC2122) and Quantum Frontiers (EXC2123).

References

1. Bar-Shalom, Y., Daum, F., Huang, J.: The probabilistic data association filter. IEEE Control Systems Magazine **29**(6), 82–100 (2009)
2. Bowman, S.L., Atanasov, N., Daniilidis, K., Pappas, G.J.: Probabilistic data association for semantic slam. In: 2017 IEEE International Conference on Robotics and Automation (ICRA). pp. 1722–1729. IEEE (2017)
3. Campos, C., Elvira, R., Rodríguez, J.J.G., Montiel, J.M., Tardós, J.D.: Orb-slam3: An accurate open-source library for visual, visual–inertial, and multimap slam. IEEE Transactions on Robotics **37**(6), 1874–1890 (2021)
4. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: European Conference on Computer Vision. pp. 213–229. Springer (2020)
5. Chiu, H.k., Li, J., Ambruş, R., Bohg, J.: Probabilistic 3d multi-modal, multi-object tracking for autonomous driving. In: 2021 IEEE International Conference on Robotics and Automation (ICRA). pp. 14227–14233. IEEE (2021)
6. Feng, M., Hou, H., Zhang, L., Wu, Z., Guo, Y., Mian, A.: 3d spatial multimodal knowledge accumulation for scene graph prediction in point cloud. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 9182–9191 (2023)
7. Feng, M., Yan, C., Wu, Z., Dong, W., Wang, Y., Mian, A.: Hyperrectangle embedding for debiased 3d scene graph prediction from rgb sequences. IEEE Transactions on Pattern Analysis and Machine Intelligence (2025)
8. Gay, P., Stuart, J., Del Bue, A.: Visual graphs from motion (vgfm): Scene understanding with object geometry reasoning. In: Asian Conference on Computer Vision. pp. 330–346. Springer (2018)
9. Gu, Q., Kuwajerwala, A., Morin, S., Jatavallabhula, K.M., Sen, B., Agarwal, A., Rivera, C., Paul, W., Ellis, K., Chellappa, R., et al.: Conceptgraphs: Open-vocabulary 3d scene graphs for perception and planning. In: 2024 IEEE International Conference on Robotics and Automation (ICRA). pp. 5021–5028. IEEE (2024)
10. Heo, K., Kim, G., Kim, S., Cho, M.: Object-centric representation learning for enhanced 3d semantic scene graph prediction. In: The Thirty-ninth Annual Conference on Neural Information Processing Systems (2025)
11. Hou, H.Y., Lee, C.Y., Sonogashira, M., Kawanishi, Y.: Fross: Faster-than-real-time online 3d semantic scene graph generation from rgb-d images. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 28818–28827 (2025)
12. Im, J., Nam, J., Park, N., Lee, H., Park, S.: Egtr: Extracting graph from transformer for scene graph generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 24229–24238 (2024)
13. Kim, U.H., Park, J.M., Song, T.J., Kim, J.H.: 3-d scene graph: A sparse and semantic representation of physical environments for intelligent agents. IEEE Transactions on Cybernetics **50**(12), 4921–4933 (2019)
14. Kipf, T.N., Welling, M.: Semi-supervised classification with graph convolutional networks. In: International Conference on Learning Representations (2017)
15. Koch, S., Hermosilla, P., Vaskevicius, N., Colosi, M., Ropinski, T.: 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 (WACV). pp. 3404–3414 (2024)

16. Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L.J., Shamma, D.A., et al.: Visual genome: Connecting language and vision using crowdsourced dense image annotations. International Journal of Computer Vision **123**(1), 32–73 (2017)
17. Li, X., Wu, T., Zheng, G., Yu, Y., Li, X.: Uncertainty-aware scene graph generation. Pattern Recognition Letters **167**, 30–37 (2023)
18. Lv, W., Zhao, Y., Chang, Q., Huang, K., Wang, G., Liu, Y.: Rt-detrv2: Improved baseline with bag-of-freebies for real-time detection transformer. arXiv preprint arXiv:2407.17140 (2024)
19. Murrugarra-Llerena, J., Kirsten, L.N., Zeni, L.F., Jung, C.R.: Probabilistic intersection-over-union for training and evaluation of oriented object detectors. IEEE Transactions on Image Processing **33**, 671–681 (2024)
20. Qi, C.R., Su, H., Mo, K., Guibas, L.J.: Pointnet: Deep learning on point sets for 3d classification and segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 652–660 (2017)
21. Rasmussen, C., Hager, G.D.: Probabilistic data association methods for tracking complex visual objects. IEEE Transactions on Pattern Analysis and Machine Intelligence **23**(6), 560–576 (2001)
22. Rezatofighi, S.H., Milan, A., Zhang, Z., Shi, Q., Dick, A., Reid, I.: Joint probabilistic data association revisited. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). pp. 3047–3055 (2015)
23. Saleh, F., Aliakbarian, S., Rezatofighi, H., Salzmann, M., Gould, S.: Probabilistic tracklet scoring and inpainting for multiple object tracking. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 14329–14339 (2021)
24. Shah, D., Eysenbach, B., Kahn, G., Rhinehart, N., Levine, S.: Ving: Learning open-world navigation with visual goals. In: 2021 IEEE International Conference on Robotics and Automation (ICRA). pp. 13215–13222. IEEE (2021)
25. Silberman, N., Hoiem, D., Kohli, P., Fergus, R.: Indoor segmentation and support inference from rgbd images. In: European Conference on Computer Vision. pp. 746–760. Springer (2012)
26. Singh, K.P., Salvador, J., Weihs, L., Kembhavi, A.: Scene graph contrastive learning for embodied navigation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 10884–10894 (2023)
27. Straub, J., Whelan, T., Ma, L., Chen, Y., Wijmans, E., Green, S., Engel, J.J., Mur-Artal, R., Ren, C., Verma, S., et al.: The replica dataset: A digital replica of indoor spaces. arXiv preprint arXiv:1906.05797 (2019)
28. Strecke, M., Stuckler, J.: Em-fusion: Dynamic object-level slam with probabilistic data association. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 5865–5874 (2019)
29. Tang, Y., Zhang, J., Lan, Y., Guo, Y., Dong, D., Zhu, C., Xu, K.: Onlineanyseg: Online zero-shot 3d segmentation by visual foundation model guided 2d mask merging. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3676–3685 (2025)
30. Tateno, K., Tombari, F., Navab, N.: Real-time and scalable incremental segmentation on dense slam. In: 2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 4465–4472. IEEE (2015)
31. Wald, J., Avetisyan, A., Navab, N., Tombari, F., Nießner, M.: Rio: 3d object instance re-localization in changing indoor environments. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 7658–7667 (2019)

32. Wald, J., Dhamo, H., Navab, N., Tombari, F.: Learning 3d semantic scene graphs from 3d indoor reconstructions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3961–3970 (2020)
33. Wang, Z., Cheng, B., Zhao, L., Xu, D., Tang, Y., Sheng, L.: Vl-sat: Visual-linguistic semantics assisted training for 3d semantic scene graph prediction in point cloud. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 21560–21569 (2023)
34. Wu, S.C., Tateno, K., Navab, N., Tombari, F.: Incremental 3d semantic scene graph prediction from rgb sequences. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5064–5074 (2023)
35. Wu, S.C., Wald, J., Tateno, K., Navab, N., Tombari, F.: Scenegraphfusion: Incremental 3d scene graph prediction from rgb-d sequences. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 7515–7525 (2021)
36. Wu, Z., Li, H., Chen, G., Yu, Z., Gu, X., Wang, Y.: 3d question answering with scene graph reasoning. In: Proceedings of the 32nd ACM International Conference on Multimedia. pp. 1370–1378 (2024)
37. Xu, D., Zhu, Y., Choy, C.B., Fei-Fei, L.: Scene graph generation by iterative message passing. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5410–5419 (2017)
38. Yang, G., Zhang, J., Zhang, Y., Wu, B., Yang, Y.: Probabilistic modeling of semantic ambiguity for scene graph generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 12527–12536 (2021)
39. Yang, J., Cen, J., Peng, W., Liu, S., Hong, F., Li, X., Zhou, K., Chen, Q., Liu, Z.: 4d panoptic scene graph generation. Advances in Neural Information Processing Systems **36**, 69692–69705 (2023)
40. Yeo, Q.X., Li, Y., Lee, G.H.: Statistical confidence rescoring for robust 3d scene graph generation from multi-view images. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 24999–25008 (2025)
41. Yin, H., Xu, X., Wu, Z., Zhou, J., Lu, J.: Sg-nav: Online 3d scene graph prompting for llm-based zero-shot object navigation. Advances in Neural Information Processing Systems **37**, 5285–5307 (2024)
42. Zhu, Y., Tremblay, J., Birchfield, S., Zhu, Y.: Hierarchical planning for long-horizon manipulation with geometric and symbolic scene graphs. In: 2021 IEEE International Conference on Robotics and Automation (ICRA). pp. 6541–6548. Ieee (2021)