NoPA: Non-Parametric Online 3D Scene Graph Generation

Qi Xun Yeo[1], Seungjun Lee[1], Yan Li[1], and Gim Hee Lee[1]

Department of Computer Science, National University of Singapore
{qixunyeo, seungjun.lee}@u.nus.edu, {yan.li,gimhee.lee}@nus.edu.sg

Abstract. Classic 3D scene graph generation approaches fail to work in real-time due to the heavy computational cost of environment mapping and the need to generate intermediate point-cloud representations. To alleviate this issue, a recent work eschews point clouds in favor of a lightweight Gaussian distribution for each object. This approximation drastically speeds up inference and enables real-time 3D scene graph generation. However, the representation has two key weaknesses. **1)** Each object is approximated by a single 3D Gaussian, which causes a severe loss of 3D geometric detail. **2)** The discrepancy between this approximation and the true object geometry exacerbates the inaccurate merging of object candidates during online inference. To address these issues, we propose **NoPA**, which represents each object as a separate non-parametric distribution. This formulation retains 3D geometric information while preserving real-time inference of the parametric Gaussian formulation. To build upon our novel object representation, we propose a tailored merging strategy to recover coherent object instances. Specifically, we leverage maximum mean discrepancy on kernel density estimates to enable robust merging of object candidates during online exploration while minimizing added computational complexity. The key is to maintain a fixed particle set per object. Furthermore, to rectify the relation loss caused by misclassified objects, NoPA propagates relationships between objects with high affinity. Experiments show that NoPA substantially outperforms current methods without sacrificing real-time inference speed.

Keywords: 3D Scene Graph · Non-Parametric Distribution · Kernel Density Function

1 Introduction

We study *online 3D scene graph generation* from streaming RGB-D images. A 3D semantic scene graph (3D SSG) encodes objects and their relationships in a structured representation and serves as a critical abstraction for embodied AI and robotics. It enables downstream tasks such as navigation [4, 12, 26, 30, 31, 33, 38, 39], scene generation [5, 22, 36, 40, 41], and manipulation [2, 5, 9, 20, 38]. These capabilities are essential in medical [10, 24, 25], construction [3, 18, 21, 23], and autonomous driving domains [7, 8, 19, 44].

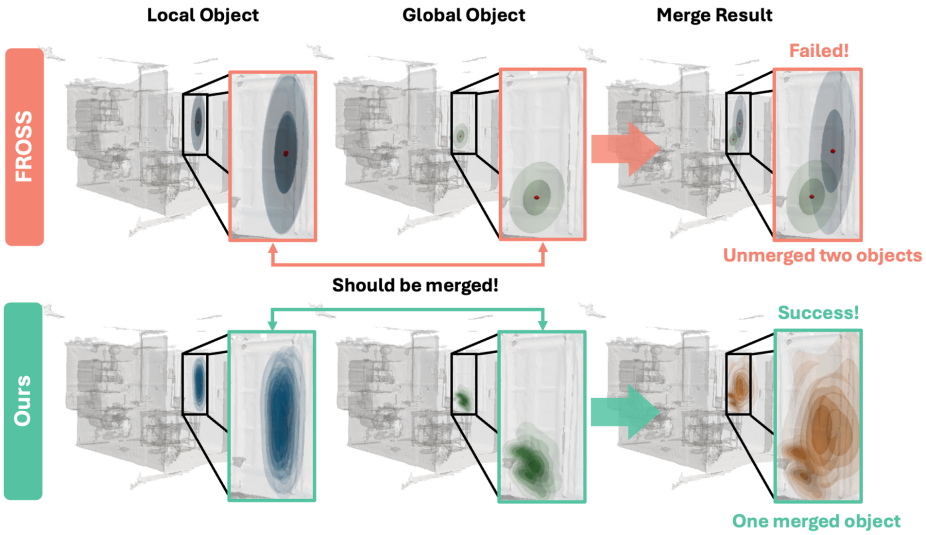

Fig. 1: Gaussian (FROSS [11]) vs Non-Parametric (Ours). We show the problem of under-merging when merging objects represented as 3D Gaussians on a single window object instance. **Top Row (FROSS):** Under Gaussian parameterization, the local object (1st column) fails to merge with the global object (2nd column) which results in under-merging (3rd column). After the merging process, the window instance is incorrectly represented by two smaller Gaussians instead of a single larger one. This fragmentation produces splintered 3D scene graphs that do not reflect the true scene structure. Strict post-processing filters would then remove fragmented objects together with their relations. As a result, the quality of the final 3D SSG degrades. **Bottom Row (Ours):** Our non-parametric formulation allows the local object to merge with the global object and preserves richer geometric support. This reduces under-merging and produces more accurate and consistent 3D SSGs.

Despite its importance, most prior works address 3D SSG generation in an offline setting without strict real-time constraints [1, 6, 28, 29, 32, 35, 37, 42]. To the best of our knowledge, only three works tackle the online setting [11, 34, 35] with real-time performance. SceneGraphFusion [35] and MonoSSG [34] rely on simultaneous localization and mapping (SLAM) pipelines to reconstruct geometry before scene graph prediction, which introduces significant computational overhead and limits scalability. FROSS [11] avoids explicit mapping by lifting 2D scene graphs into 3D and achieves high frame rates by approximating each object as a Gaussian distribution.

Although FROSS achieved real-time performance without SLAM, their Gaussian parameterization imposes a restrictive geometric assumption where each object is modeled as an ellipsoid defined by its covariance. This approximation discards fine geometric structure and makes merging fragile. As shown in Fig. 1, thin or planar structures such as pictures and windows often produce near-singular covariance matrices leading to under-merging. Different viewpoints of the same

object often yields Gaussian ellipsoids with inconsistent covariances and spatial offsets causing incorrect merges. Consequently, merging decisions are prone to be unstable since incorrect merges and undermerges accumulate over time and progressively degrade the global 3D SSG. These limitations arose fundamentally from the parametric assumption instead of implementation details.

To overcome both the computational burden of SLAM-based pipelines and the geometric limitations of Gaussian modeling, we introduce **NoPA** (**Non-PArametric** Online 3D Scene Graph Generation). Our NoPA replaces Gaussian object modeling with a fixed-size non-parametric particle set that preserves geometric support. This formulation removes the restrictive ellipsoidal assumption while maintaining constant memory and runtime complexity. Merging two object candidates proceeds by estimating a kernel density over their unified particle support and resampling a fixed-size set from it. This integrates multi-view geometric evidence while preserving a constant-size representation. Consequently, NoPA matches the real-time efficiency of SLAM-free parametric methods while retaining substantially richer geometric structure.

Adopting a non-parametric representation shifts the merging problem from covariance comparison to distribution comparison across views. We empirically find that covariance similarity is insufficient and unreliable under viewpoint variation, often resulting in under-merging (*cf*. Fig. 1). We address this by introducing a principled distribution-level merging criterion based on Maximum Mean Discrepancy (MMD). MMD measures similarity directly between particle sets in feature space and provides a stable signal even when geometric support differs across views or when 2D predictions are noisy. This significantly improves merging robustness in ambiguous cases and prevents cascading structural errors in the final 3D SSG. To further strengthen the global consistency of the 3D SSG, our NoPA incorporates a relationship propagation mechanism as a post-processing step. We cluster object candidates using the previously computed MMD scores and propagate relationships across clusters to recover missing edges. This mitigates structural damage caused by imperfect merging and enhances overall graph completeness without sacrificing runtime efficiency.

In summary, our main contributions are as follows:

- We introduce **NoPA**, a non-parametric formulation for online 3D scene graph generation that eliminates restrictive Gaussian assumptions while preserving fixed memory usage and real-time computational complexity.
- We design a principled **distribution-level merging framework** based on Maximum Mean Discrepancy that replaces fragile covariance similarity and improves robustness under viewpoint variation and noisy predictions.
- We develop a **relationship propagation mechanism** guided by distribution similarity to recover missing relations and reinforce graph consistency.
- We achieve state-of-the-art (SOTA) performance on multiple online 3D SSG benchmarks, which demonstrates superior accuracy while maintaining competitive real-time efficiency.

2 Related Work

Offline 3D SSG. Offline 3D SSG generation approaches aim to estimate a 3D scene graph using ground truth 3D geometry [1, 28, 29, 32, 42] or multi-view RGB-D images [6, 9, 27, 37, 43] in a non-incremental manner. Wald et al. [28] first proposed the problem of 3D SSG generation and attempted to solve it by modeling pairwise relationships to predict the graph. Most modern approaches rely on multi-view RGB-D images. Wang et al. [32] leverage pretrained model priors by distilling knowledge from a multimodal oracle model into a 3D model. Yeo et al. [37] propose a statistical confidence rescoring mechanism to refine low-confidence predictions and use SegmentAnything (SAM) [15] instance masks to enhance node features. Koch et al. [16] distill knowledge from visual language models (VLMs) into a 3D graph neural network (GNN). Gu et al. [9] run a class-agnostic segmentation model to obtain candidate objects, associate them across views using geometric and semantic similarity, instantiate nodes in a 3D SSG refined by VLMs, and prompt a large language model (LLM) with object pairs to infer spatial relations. Koch et al. [17] build a relationship-aware 3D representation that supports node and predicate queries. Zhang et al. [43] introduce functional relationships and contribute a functional 3D SSG dataset. These offline systems typically aggregate information over a fixed set of frames and perform expensive global association and refinement. This becomes challenging under strict online latency and bounded-memory constraints. In contrast, our NoPA targets online 3D SSG generation with constant memory per object. We replace Gaussian object modeling with fixed-size particle sets, use a distribution-level merging criterion to improve cross-view association, and propagate relations within affinity clusters to recover missed edges during incremental fusion.

Online 3D SSG. The first work to tackle 3D SSG generation in an online setting is by Kim et al. [14]. The work focuses on predicting local 3D SSGs that combine into a global 3D SSG. However, it fails to reach real-time speeds required for practical deployment. Wu et al. [35] introduce a graph convolutional network (GCN) based aggregation function, abbreviated as FAN, to improve the predicted 3D SSG while running RGB-D SLAM to obtain dense intermediate 3D representations. MonoSSG [34] proposes an entity association approach that lifts 2D entities into 3D, enhances ORBSLAM, and introduces a geometric gate that fuses geometric information with multi-view image features. More recently, FROSS [11] approximates objects as 3D Gaussians to avoid heavy point cloud processing or environment mapping, which accelerates inference. However, by removing precise localization, it fails to use geometric information available in 3D for merging and instead relies on an approximation of the 2D object shape from RGB-D observations. In contrast, our NoPA balances the trade-off between preserving geometric detail and improving inference speed through a fixed number of particles per object. This design maintains real-time performance while retaining richer 3D geometry than FROSS, which improves merging and overall model performance.

3 Problem Definition

Given N multi-view RGB images of a 3D scene, denoted as $\{I_i\}_{i=1}^{N}$, our goal is to estimate a 3D scene graph:

$$G^{3D} = (O, R), \tag{1}$$

where $O = \{o_j\}_{j=1}^{M}$ is the set of M object nodes and $R = \{r_{k\to j}\}_{j,k=1}^{M}$ is the set of directed relationship edges over object pairs. Node j has an object (category) label o_j, and the directed edge from node k to node j has a predicate label $r_{k\to j}$. Equivalently, the scene graph can be represented as a set of triplets $\{(o_k, r_{k\to j}, o_j)\}$.

We study the *online* setting. The method does not assume access to the full image set $\{I_i\}_{i=1}^{N}$ at test time. Instead, it receives a sequential stream of partial observations and incrementally updates G^{3D} as new images arrive during scene exploration.

4 Preliminaries

We build our framework based on FROSS [11], where it updates the scene graph incrementally from a stream of RGB frames and avoids explicit environment mapping. Given RGB observations $\{I_i\}_{i=1}^{N}$, FROSS first predicts a per-frame 2D scene graph $G_i^{2D} = g_\phi(I_i)$, where g_ϕ is a pretrained 2D SSG detector. It then lifts this 2D graph into the world frame using the per-frame depth map d_i and camera pose $P_i \in SE(3)$, and fuses the lifted result into the global 3D scene graph:

$$G_i^{3D} = \mathcal{B}\big(G_i^{2D} \mid d_i, P_i\big) \odot G_{i-1}^{3D}, \tag{2}$$

where G_{i-1}^{3D} and G_i^{3D} denote the global 3D scene graph before and after processing frame i. The operator $\mathcal{B}(\cdot \mid d_i, P_i)$ maps each 2D node in G_i^{2D} to a 3D object hypothesis by back-projecting image evidence with d_i and transforming it to the world frame with P_i. The fusion operator $\odot$ performs data association between the lifted hypotheses and existing nodes in G_{i-1}^{3D}, followed by merging and state updates.

FROSS represents each 3D object node o_j with a single Gaussian in $\mathbb{R}^3$:

$$p(\mathbf{x} \mid o_j) = \mathcal{N}(\mathbf{x}; \mu_j, \Sigma_j)\,, \tag{3}$$

where $\mathbf{x} \in \mathbb{R}^3$ is a 3D point, $\mu_j \in \mathbb{R}^3$ is the object centroid, and $\Sigma_j \in \mathbb{R}^{3\times 3}$ is the covariance. This Gaussian is initialized by lifting a 2D Gaussian estimated from the predicted 2D bounding box. During fusion, it merges a lifted object hypothesis i with an existing global object j when their semantic labels match and their Hellinger distance $d_H(i, j)$ falls below a threshold δ_H. Approximating each object with a Gaussian $\mathcal{N}(\mu, \Sigma)$, the Hellinger distance admits a closed form:

$$d_H(i, j) = \sqrt{1 - \exp\big(-d_B(i, j)\big)}. \tag{4}$$

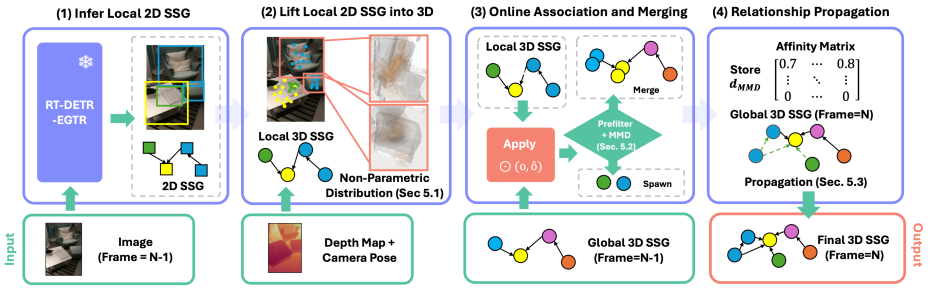

Fig. 2: Overview of our online 3D scene graph generation pipeline. (1) A pretrained RT-DETR-EGTR [13,45] model predicts a local 2D scene graph from each RGB frame. (2) For every object node, we sample pixels inside its 2D bounding box, back-project them with depth, and obtain a 3D particle set in the world frame. (3) We associate each local particle set with existing global objects using a two-stage test: a constant-time Hellinger pre-filter followed by MMD on ambiguous cases. (4) We propagate relations within high-affinity clusters to reduce relation dropouts from the 2D predictor.

where $d_B(i,j)$ is the Bhattacharyya distance between two Gaussians $\mathcal{N}(\mu_i, \Sigma_i)$ and $\mathcal{N}(\mu_j, \Sigma_j)$:

$$d_B(i,j) = \frac{1}{8}\Delta\mu_{ij}^\top \Sigma^{-1} \Delta\mu_{ij} + \frac{1}{2}\ln\left(\frac{\det\Sigma}{\sqrt{\det\Sigma_i \,\det\Sigma_j}}\right),$$
$$\Delta\mu_{ij} = \mu_i - \mu_j, \qquad \Sigma = \frac{\Sigma_i + \Sigma_j}{2}. \tag{5}$$

Limitations of FROSS. FROSS represents each object with a single Gaussian, which coarsely approximates the object geometry with an ellipsoidal shape, removing the fine structural details of the instances. This approximation loss is accumulated across the streaming images, resulting in disjoint object instances from fragile merging and incorrect associations.

5 Our Method

Overview. Fig. 2 summarizes our online 3D scene graph generation framework. We improve FROSS by replacing its single-Gaussian object model with a non-parametric particle representation, and redesigning online fusion around distributional similarity. At timestep i, a pretrained RT-DETR-EGTR model infers a *local 2D scene graph* G_i^{2D} from the RGB frame I_i. G_i^{2D} contains object nodes (2D boxes with class labels) and relation edges (pairwise predicates) in the image plane. We lift each detected 2D object node into a 3D particle set using the depth map d_i and camera pose P_i (Sec. 5.1), and fuse these local 3D candidates into the global 3D scene graph from the previous timestep (Sec. 5.2). Fusion uses a fast two-stage association rule: a constant-time Hellinger pre-filter for clear matches/mismatches, and a maximum mean discrepancy (MMD) test

for borderline pairs. We then stabilize the relation set by propagating relations within affinity clusters, which helps to recover the relations missed by the 2D predictor in a single view (Sec. 5.3). This design preserves object support in $\mathbb{R}^3$, improves cross-view association, and maintains constant memory per object.

5.1 Non-Parametric Object Representation

From 2D boxes to 3D particles. For each detected 2D object node in G_i^{2D} with bounding box b, we sample n pixels $\{\mathbf{u}_k\}_{k=1}^n$ uniformly within b. Using the depth map d_i, we back-project each pixel to a camera-frame 3D point $\mathbf{X}_k^c = \pi^{-1}(\mathbf{u}_k, d_i(\mathbf{u}_k))$ and transform it to the world frame with the camera pose $P_i \in SE(3)$:

$$\mathbf{x}_k = P_i \, \mathbf{X}_k^c, \qquad \mathbf{x}_k \in \mathbb{R}^3. \tag{6}$$

The lifted object is represented by the particle set $\mathcal{X}(o) = \{\mathbf{x}_k\}_{k=1}^n$.

Kernel density view. We treat the particle set as samples from an unknown object occupancy distribution and form a kernel density estimate (KDE):

$$\hat{f}(\mathbf{x} \mid o) = \frac{1}{n} \sum_{k=1}^{n} \kappa(\mathbf{x}, \mathbf{x}_k), \tag{7}$$

where $\kappa(\cdot, \cdot)$ is an RBF kernel:

$$\kappa(\mathbf{x}, \mathbf{y}) = \exp\Big(- \tfrac{\|\mathbf{x}-\mathbf{y}\|_2^2}{2\sigma^2} \Big). \tag{8}$$

Remark. A single 3D Gaussian enforces an ellipsoidal prior, which blurs multi-part structures and amplifies approximation error across views. As shown in Fig. 3, a particle set preserves object support in $\mathbb{R}^3$ and admits multi-modality, which reduces over-merging and under-merging under partial observations.

5.2 Online Association and Merging

We maintain a set of global objects, each with particles $\mathcal{X}(o)$. Given a local object candidate $\hat{o}$ with particles $\mathcal{X}(\hat{o})$ and an existing global object o with particles $\mathcal{X}(o)$, we decide whether $\hat{o}$ corresponds to o and should be merged, or whether it should spawn a new global object. Our association uses a two-stage criterion: 1) a ***constant-time Hellinger pre-filter*** that resolves clear cases, followed by 2) an ***MMD test*** on ambiguous pairs.

Stage 1: Constant-time pre-filter. To obtain a cheap moment-level proxy, we fit a *unimodal Gaussian* to each particle set by matching its first two moments: (μ, Σ) for o and $(\hat{\mu}, \hat{\Sigma})$ for $\hat{o}$. We then compute the Hellinger distance d_H between the two fitted Gaussians. Instead of committing to a hard threshold at δ_H, we introduce a *margin band* of width 2ϵ:

$$\text{merge if } d_H < \delta_H - \epsilon, \qquad \text{spawn if } d_H > \delta_H + \epsilon. \tag{9}$$

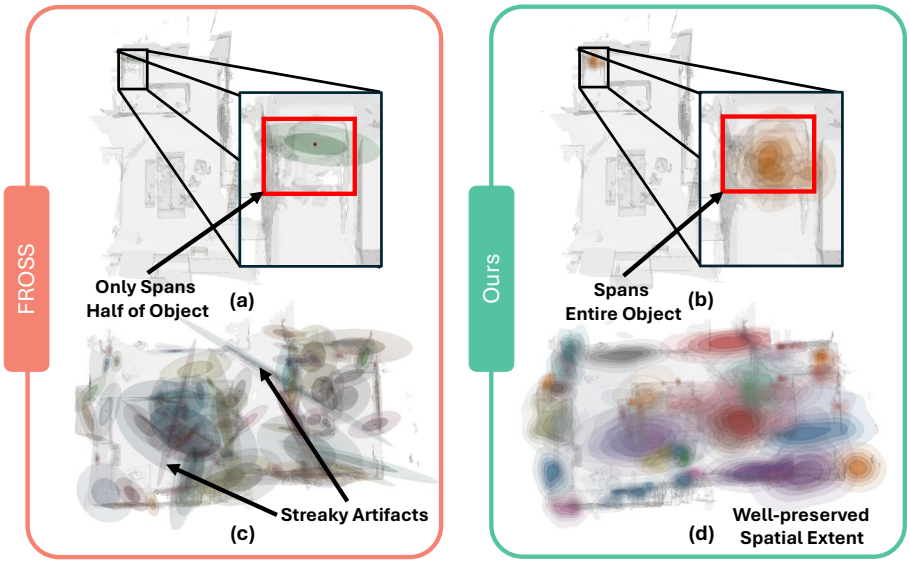

Fig. 3: The visualization of objects in Scene `41385849` from the 3DSSG dataset. In the top row, we visualize an instance of the sink class localized by a red bounding box. In the bottom row, we visualize the global 3D object instances. **Left:** Visualized Gaussian blobs from FROSS [11]. The Gaussian blob only encompasses half of the sink in (a). Spurious blobs spanning across the scene in (c) visualizes the impact of incorrect merging. **Right:** Visualized kernel densities for our particle set. The entire sink is associated with our particle set in (b). Our representation attains good coverage on objects across the scene without large artifacts in (d).

The band prevents unstable decisions when d_H fluctuates under depth noise, truncation, or limited view overlap. Pairs within $[\,\delta_H - \epsilon,\ \delta_H + \epsilon\,]$ remain undecided and move to Stage 2.

Stage 2: MMD for ambiguous pairs. For candidates inside the margin band $[\,\delta_H - \epsilon,\ \delta_H + \epsilon\,]$, we compute the maximum mean discrepancy (MMD) between the two KDEs:

$$d^2_{\mathrm{MMD}}(o, \hat{o}) = \mathbb{E}_{\mathbf{x},\mathbf{x}' \sim \mathcal{X}(o)}[\kappa(\mathbf{x}, \mathbf{x}')] + \mathbb{E}_{\mathbf{y},\mathbf{y}' \sim \mathcal{X}(\hat{o})}[\kappa(\mathbf{y}, \mathbf{y}')] - 2\,\mathbb{E}_{\mathbf{x} \sim \mathcal{X}(o),\,\mathbf{y} \sim \mathcal{X}(\hat{o})}[\kappa(\mathbf{x}, \mathbf{y})]\,. \tag{10}$$

We set σ^2 with the median heuristic over random pairs from $\mathcal{X}(o) \cup \mathcal{X}(\hat{o})$. We merge if $d_{\mathrm{MMD}}(o, \hat{o}) \leq \delta_{\mathrm{MMD}}$, and spawn otherwise, where δ_{MMD} is a fixed threshold calibrated on a held-out sequence to match the desired precision–recall trade-off.

Why MMD? The Stage 1 Gaussian fit is intentionally coarse, where different particle sets can share similar (μ, Σ) under partial views, thin structures, or multi-part objects. MMD directly compares the *full distributions* induced by

the KDEs in a reproducing kernel Hilbert space, which makes it sensitive to support mismatch beyond first- and second-order moments. It is also model-free and works naturally with our particle representation, with the requirement of only kernel evaluations without meshing or explicit points correspondence.

Decision rule. Following the online update in Eq. 2, the fusion operator $\odot$ implements the two-stage association between a local candidate $\hat{o}$ and a global object o:

$$\odot(o,\hat{o}) = \begin{cases} \text{merge} & d_H < \delta_H - \epsilon, \\ \text{spawn} & d_H > \delta_H + \epsilon, \\ \text{merge} & \text{otherwise and } d_{\text{MMD}}(o,\hat{o}) \le \delta_{\text{MMD}}, \\ \text{spawn} & \text{otherwise.} \end{cases} \tag{11}$$

It first accepts or rejects unambiguous pairs using d_H, and invokes d_{MMD} only within the margin band to resolve borderline cases where moment matching proves unreliable.

Merge update with constant memory. After each merge, we take the union of the particles, fit a KDE over the union support, and resample a fixed-size set of n particles:

$$\mathcal{X}(o) \leftarrow \text{Resample}_n\big(\mathcal{X}(o) \cup \mathcal{X}(\hat{o})\big). \tag{12}$$

The resampling operation Resample_n acts on the union of the global particle set and the observed particle set in 3D space. Specifically, resampling corresponds to the proposal step of Monte-Carlo Markov Chain (MCMC) sampling, where we uniformly sample particle indices and perturb them with Gaussian noise:

$$I_j \sim \text{Uniform}\{1,2,\ldots,n\}, \qquad \varepsilon_j \sim \mathcal{N}(0, h^2\Sigma), \tag{13}$$

$$\mathcal{X}_j(o) = \mathcal{X}_{I_j}(o) + \varepsilon_j. \tag{14}$$

where $h^2\Sigma$ is the bandwidth matrix computed from scaling factor h and sample covariance Σ. In vectorized form:

$$\mathcal{X}(\mathbf{o}) = \mathcal{X}_{:,\mathbf{I}}(\mathbf{o}) + L\,\mathbf{Z}, \qquad LL^\top = h^2\Sigma \quad \text{(Cholesky)}, \tag{15}$$

where $\mathbf{Z} \in \mathbb{R}^{d\times n}$ contains *i.i.d.* Gaussian samples. This step preserves geometric information from both candidates and prevents particle growth over time.

Remark. The Hellinger pre-filter keeps most association decisions inexpensive. MMD focuses computation on hard cases where first- and second-order moments agree but distribution support differs. As shown in Fig. 4, resampling maintains constant memory and avoids collapsing the representation toward a single mode, which stabilizes long-horizon fusion.

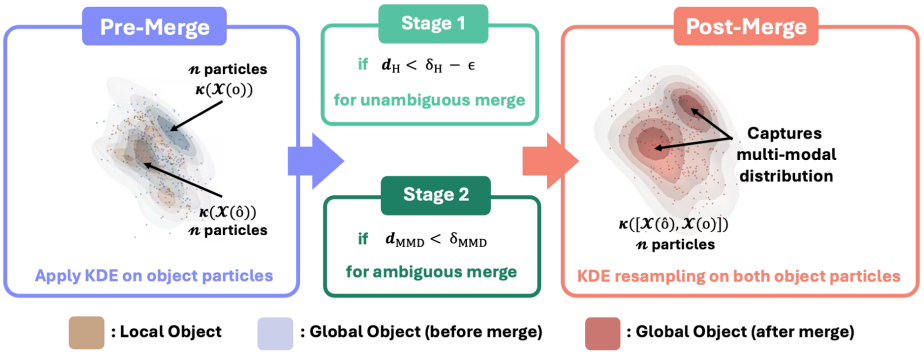

Fig. 4: Visualization of the merging process for an object in Scene `7272e16c` from the 3DSSG dataset. If the local particle set and the global particle set yields a small Hellinger distance ($d_H < \delta_H - \epsilon$) after fitting a unimodal Gaussian (*Stage 1*), their covariances clearly matches, and the merge decision is straightforward. If the Hellinger distance falls within the margin band (*Stage 2*), covariance alignment alone is insufficient to determine merging. Consequently, we apply the more robust MMD criterion and merge the particle sets if $d_{\text{MMD}} < \delta_{\text{MMD}}$. **Left:** Initial kernel density estimates of the local and global particle sets. **Right:** Kernel density estimate of the merged particle set after fusion. The number of particles remain constant after merging due to KDE resampling while the updated particle set captures the multi-modal distribution.

5.3 Relationship Propagation with Affinity Clusters

Online 3D SSG generation is sensitive to relation dropouts, where low-confidence predicate edges are suppressed, and a missing edge in one frame may never be recovered later. To increase robustness, we exploit geometric redundancy across objects with similar 3D support. Specifically, we reuse the MMD scores computed during ambiguous association to build an affinity matrix over object candidates. The affinity function is calculated as:

$$\mathbf{A} = [\max(0, 1 - \frac{d_{\text{MMD}}(\mathcal{X}_i(o), \mathcal{X}_j(o))}{2\delta_{\text{MMD}}})]_{i,j=1}^{n,n}, \tag{16}$$

which satisfies the criteria of boundedness and monotonicity thereby ensuring a well-formed formulation. If two segments possess low affinity score below affinity threshold τ, they are prevented from being grouped into the same cluster. This filtering mechanism avoids forming overly large clusters which improves amortized computational efficiency.

For a newly merged or spawned node, we first copy its observed 2D relations to the corresponding 3D node. We then propagate candidate relations within its affinity cluster and finalize the relation type by majority voting on the available evidence. This aggregation recovers relations missed in a single view, reduces sensitivity to relation confidence thresholds, and limits relation drift.

Remark. Affinity-based propagation recovers relations that are missed in a single view by borrowing consistent evidence from geometrically similar neighbors.

Majority voting reduces the impact of occasional false positives and prevents relation drift.

6 Experiments

6.1 Experimental Setup

Datasets. We evaluate on the 3DSSG [28] and ReplicaSSG datasets following the same evaluation set up as FROSS [11]. 3DSSG contains 1482 scenes which are annotated with 21974 objects and 16324 predicate relationship between two objects. Each scene encompasses multiple short videos with different motion trajectories capturing the entire scene. The dataset is challenging due to the poor quality of images captured with significant motion blur. ReplicaSSG contains 18 scenes which are annotated with 1526 objects and 582 predicate relationship between two objects. The dataset contains high quality images and meshes inherited from the Replica dataset. The large number of objects and predicates per scene makes evaluation non-trivial.

Baseline Methods. For 3DSSG, we compare our method with the other online 3D SSG generation methods that utilize RGB-D images with the ground truth pose information including Kim's framework [14], JointSSG [34][1], FROSS [11]. We exclude MonoSSG [34] since it only uses RGB images. We reproduce the approaches using their respective GitHub repositories[2].

For ReplicaSSG, we only compare with FROSS since it is the only prior approach that evaluates on the dataset.

Implementation Details. We choose to keep the top 20 relationships from the pretrained RT-DETR-EGTR instead of the top 10 to reduce the loss of the relationship edge from the final 3D SSG when running FROSS. This ensures a fair comparison between the competing approaches.

Our implementation is built using the PyTorch framework. Following FROSS [11], we pretrain the initial 2D SSG generation model - RT-DETR-EGTR for a maximum of 50 epochs. All experiments are conducted on a single RTX 3090 GPU for a fair comparison. We utilized $n = 256$ particles to represent each object. δ_{MMD} is set to 0.7 for 3DSSG and 0.6 for ReplicaSSG. ϵ is kept at 0.05. τ is set at 0.25.

Similar to FROSS, we omit the *None* class for both object and predicate predictions that were previously implemented for SceneGraphFusion [35]. The advantage of removing the *None* class is the prevention of possible overfitting to the *None* class which is the most prevalent ground truth annotation.

We follow the same strict criteria as FROSS to match the predicted object candidates to the ground truth object instances: (1) The majority of points

[1] JointSSG is the framework using RGB-D input from the same paper as MonoSSG.
[2] For JointSSG, we follow https://github.com/ShunChengWu/3DSSG. For Kim's framework and FROSS, we follow https://github.com/Howardkhh/FROSS.

Table 1: Comparison with SOTA online 3D SSG generation approaches on the 3DSSG dataset with 20 object classes and 7 predicate classes. The top group of results are the reported results from the respective papers. The middle group of results marked with the † are the reproduced results from the respective GitHub repositories. The **Best** and Second Best results are highlighted, respectively.

Method	Recall% (↑)			mRecall% (↑)		Latency (ms ↓)	VRAM (MB↓)
	Rel.	Obj.	Pred.	Obj.	Pred.		
JointSSG [34]	25.5	58.1	27.3	43.0	33.3	191	-
Kim [14]	9.1	59.0	7.1	51.0	8.0	310	-
FROSS [11]	27.9	62.4	33.0	63.8	18.0	7	-
JointSSG† [34]	23.4	55.4	27.0	45.4	**35.3**	284	3252
Kim†[3] [11]	0.9	52.1	1.1	44.2	0.4	488	**1204**
FROSS† [11]	25.7	60.6	30.7	62.4	17.7	**22**	**1204**
Ours($n = 128$)	49.9	68.5	58.5	65.7	30.2	26	1206
Ours($n = 256$)	**53.2**	**69.0**	**61.4**	**66.4**	29.4	27	1206

(more than 50%) sampled from our continuous particle set should have its nearest ground truth point mapped to the corresponding matched ground truth object. (2) The fraction of overlap counts belonging to the second-largest matched ground truth object compared to the overlap counts belonging to the largest matched ground truth object must not exceed 75%. These criteria enforces one-to-one correspondence between predicted and ground truth objects which increases the robustness of the evaluation.

Evaluation Metrics. In terms of evaluating the 3D scene graph centric performance, we follow SceneGraphFusion [35] and MonoSSG [34] to report the overall **top-1** recall (Recall) for the object class estimation (Obj.), the predicate estimation (Pred.), and the relationship triplet estimation (Rel.). We also report the mean recall (mRecall) for the object class estimation (Obj.), the predicate estimation (Pred.) only. In terms of evaluating the runtime efficiency for the online setting, we report the latency as in [11] and additionally compare the memory requirements of each method.

6.2 Quantitative Results

We present the main performance comparison for the top-1 recall and mean recall against other baseline methods in Tab. 1. Our method surpasses all baselines in top-1 recall and object mRecall while maintaining competitive latency and comparable VRAM usage for real-time inference. More impressively, NoPA with 128 particles already surpasses all baselines. These results validate our hypothesis that the expressiveness of our continuous non-parametric formulation and our improved merging process contribute to the superior performance of NoPA.

[3] Kim's method crashed in the 132nd scene out of 157 scenes from RAM OOM. The point clouds stored are further resized down by 5x to fit the memory possibly leading to the large drop in performance.

Table 2: Comparison with FROSS on the ReplicaSSG dataset with 34 object classes and 9 predicate classes. † refers to the reproduced results. The **Best** and Second Best results are highlighted, respectively.

Method	Recall% (↑)			mRecall% (↑)		Latency (ms ↓)	VRAM (MB↓)
	Rel.	Obj.	Pred.	Obj.	Pred.		
FROSS [11]	22.3	26.1	27.8	28.8	20.4	7	-
FROSS† [11]	22.3	25.3	27.5	27.6	12.6	**17**	**1206**
Ours($n = 128$)	32.4	28.0	34.6	29.6	16.5	22	1230
Ours($n = 256$)	**36.9**	**28.6**	**39.5**	**29.8**	**18.6**	23	1230

Table 3: Ablation study on the test split of the 3DSSG dataset. NP. refers to the usage of our non-parametric particle set distribution instead of Gaussian distribution to represent objects in the scene. Merge. refers to the usage of our MMD-based merging approach. Prop. refers to the usage of our relationship propagation mechanism. The **Best** results are shown in bold.

Method	NP.	Merge.	Prop.	Recall% (↑)			mRecall% (↑)	
				Rel.	Obj.	Pred.	Obj.	Pred.
FROSS	×	×	×	25.7	60.6	30.7	62.4	17.7
+ NP.	✓	×	×	17.6	66.1	20.8	64.8	11.1
+ Merge.	✓	✓	×	26.3	**69.0**	31.0	**66.4**	17.1
Ours	✓	✓	✓	**53.2**	**69.0**	**61.4**	**66.4**	**29.4**

For ReplicaSSG in Tab. 2, NoPA surpasses FROSS across all metrics with a particularly significant margin for relationship recall (65.5% increase).

6.3 Qualitative Results

We visualize the qualitative results compared to FROSS in Fig. 5. Compared to FROSS, our approach can differentiate objects with thin structures such as the window class with a higher success rate. For ambiguous cases such as the misclassification of the counter instance, the failure is understandable since both methods rely on coarse approximations while differentiating between a counter and a cabinet is non-trivial. Notably, NoPa correctly associates partial wall observations with the appropriate wall instance despite their textureless appearance. In contrast, this ambiguity poses a challenge for FROSS, which often produces incorrect merges and therefore exhibits poorer performance. Additional qualitative results are provided in the supplementary material.

6.4 Ablation Studies

To validate the contribution of each component of our proposed approach, we ablate each component in Tab. 3. Replacing the Gaussian distribution with the particle set distribution increases recall for objects at the expense of predicate

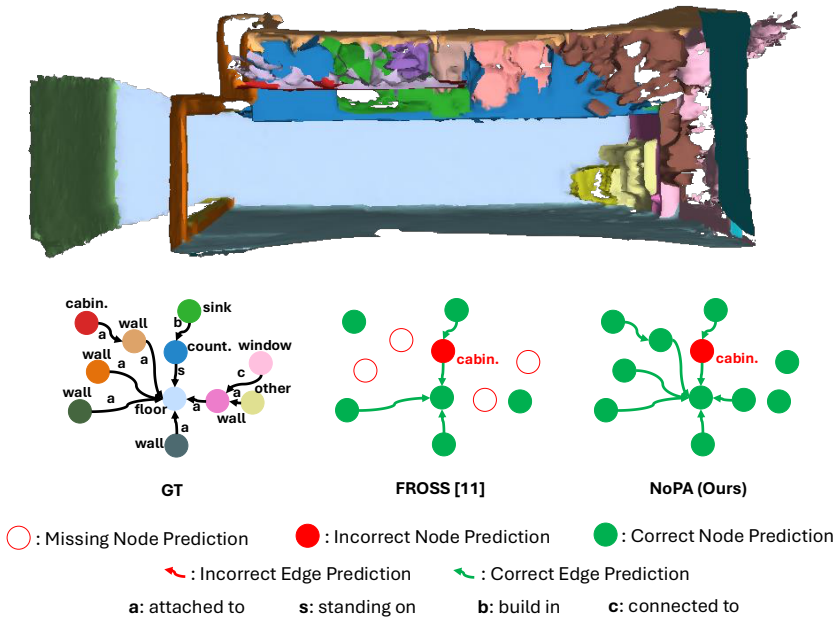

Fig. 5: We compare the qualitative results between FROSS and our proposed approach for scene 321c867e from the 3DSSG dataset. The scene shows a kitchen from bird's eye view (BEV). FROSS fails to predict a majority of the wall background class. As a consequence, a majority of the predicate relationships are lost. Our method correctly classifies most objects, except for the counter instance, while correctly predicting the majority of predicate relationships.

and relationship recall. The merging approach in FROSS for parametric representations clearly degrades performance when applied to our non-parametric representation due to fundamental incompatibilities between the two formulations. Integrating the distribution replacement with our tailored merging approach for non-parametric representations yields further improvements across all metrics. The relationship propagation mechanism further boosts both relationship and predicate recall with no degradation in object recall since the position and spatial extent of all objects are unaffected by its use.

We analyze the effects of our merging approach in Tab. 4. We vary the values of δ_{MMD} between the values of 0.6 and 0.9. We observe that a lower threshold improves object recall at the expense of predicate recall due to stricter merging criteria. Conversely, a higher threshold improves predicate and relationship recall with a trade-off in object recall. This corroborates the findings in [11]. The matching criteria discard predicted objects after merging that either lack sufficient overlap with the ground truth or excessively overlap with other matched objects due to their increased spatial extent. Relaxing the merging criterion produces more merged objects with a larger spatial extent, thereby reducing object recall.

Table 4: Comparison between different values of MMD threshold on the validation split of the 3DSSG dataset δ_{MMD}. The **Best** results are shown in bold.

δ_{MMD}	0.6	0.65	0.7	0.75	0.8	0.85	0.9
Rel.	36.3	47.9	**53.7**	51.3	46.4	46.5	46.3
Obj.	**68.5**	67.0	66.0	63.3	59.2	59.3	59.2
Pred.	40.6	53.3	**61.0**	58.8	53.7	53.7	53.5

Calculating MMD for particle set distribution similarity is superior to the approximation of the particle set distribution into a 3D Gaussian needed for the calculation of Hellinger distance especially for ambiguous cases. A theoretical explanation is the possible presence of misclassified objects from the predictions of the 2D SSG. The covariance calculated in Hellinger distance alone cannot capture the full distributional difference in object class that the use of MMD captures. This overreliance on covariance alone possibly leads to incorrect merges between objects from different classes which may explain the superiority of our approach. For more ablations and analysis, refer to the supplementary material.

Limitations. Similar to prior works that lift 2D SSG to 3D, our approach heavily depends on the accuracy of the 2D SSG predictions from the pretrained models, especially for object prediction. The quality of 2D detections and relations cap the upper bound of performance for our approach.

7 Conclusion

We present **NoPA**, a non-parametric framework for online 3D scene graph generation from multi-view RGB-D observations. Our method replaces Gaussian object models with fixed-size particle sets that preserve geometric support while maintaining constant memory and runtime complexity. This design removes the restrictive ellipsoidal assumption and yields more stable multi-view object associations. We introduce a distribution-level merging criterion based on MMD that compares particle sets directly in feature space and improves robustness under viewpoint variation and noisy predictions. A lightweight Hellinger distance prefilter preserves efficiency by avoiding unnecessary MMD evaluations. We further propose a relationship propagation mechanism that recovers missing relations and improves global graph consistency. Experiments show SOTA performance on multiple online 3D scene graph benchmarks while maintaining competitive real-time efficiency. These results show that non-parametric object representations can viably replace parametric modeling for online 3D SSG generation.

Acknowledgments

This research / project is supported by the National Research Foundation (NRF) Singapore, under its NRF-Investigatorship Programme (Award ID. NRF-NRFI09-0008), and the Tier 2 grant MOET2EP20124-0015 from the Singapore Ministry of Education.

References

1. Armeni, I., He, Z.Y., Gwak, J., Zamir, A.R., Fischer, M., Malik, J., Savarese, S.: 3d scene graph: A structure for unified semantics, 3d space, and camera. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 5664–5673 (2019)
2. Buechner, M., Roefer, A., Engelbracht, T., Welschehold, T., Bauer, Z., Blum, H., Pollefeys, M., Valada, A.: Articulated 3d scene graphs for open-world mobile manipulation. arXiv preprint arXiv:2602.16356 (2026)
3. Çelen, A., Han, G., Schindler, K., Gool, L.V., Armeni, I., Obukhov, A., Wang, X.: I-design: Personalized LLM interior designer. In: Bue, A.D., Canton, C., Pont-Tuset, J., Tommasi, T. (eds.) Computer Vision - ECCV 2024 Workshops - Milan, Italy, September 29-October 4, 2024, Proceedings, Part II. Lecture Notes in Computer Science, vol. 15624, pp. 217–234. Springer (2024). https://doi.org/10.1007/978-3-031-92387-6_17
4. Chang, Y., Ballotta, L., Carlone, L.: D-lite: Navigation-oriented compression of 3d scene graphs for multi-robot collaboration. IEEE Robotics Autom. Lett. **8**(11), 7527–7534 (2023). https://doi.org/10.1109/LRA.2023.3320011
5. Dhamo, H., Manhardt, F., Navab, N., Tombari, F.: Graph-to-3d: End-to-end generation and manipulation of 3d scenes using scene graphs. In: IEEE International Conference on Computer Vision (ICCV) (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. pp. 9182–9191 (2023)
7. Fischer, T., Porzi, L., Bulò, S.R., Pollefeys, M., Kontschieder, P.: Multi-level neural scene graphs for dynamic urban environments. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024. pp. 21125–21135. IEEE (2024). https://doi.org/10.1109/CVPR52733.2024.01996
8. Greve, E., Büchner, M., Vödisch, N., Burgard, W., Valada, A.: Collaborative dynamic 3d scene graphs for automated driving pp. 11118–11124 (2024). https://doi.org/10.1109/ICRA57147.2024.10610112
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. Guo, D., Lin, M., Pei, J., Tang, H., Jin, Y., Heng, P.A.: Tri-modal confluence with temporal dynamics for scene graph generation in operating rooms. In: MICCAI. Springer (2024)
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) (October 2025)
12. Huang, X., Zhao, S., Wang, Y., Lu, X., Zhang, W., Qu, R., Li, W., Wang, Y., Wen, C.: Msgnav: Unleashing the power of multi-modal 3d scene graph for zero-shot embodied navigation (2026), https://arxiv.org/abs/2511.10376
13. 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 (June 2024)
14. Kim, U.H., Park, J.M., Song, T.J., Kim, J.H.: 3d-scene-graph: A sparse and semantic representation of physical environments for intelligent agents. IEEE Cybernetics (2019)
15. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4015–4026 (2023)
16. Koch, S., Vaskevicius, N., Colosi, M., Hermosilla, P., Ropinski, T.: Open3dsg: Open-vocabulary 3d scene graphs from point clouds with queryable objects and open-set relationships. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2024)
17. Koch, S., Wald, J., Colosi, M., Vaskevicius, N., Hermosilla, P., Tombari, F., Ropinski, T.: Relationfield: Relate anything in radiance fields. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2025)
18. Lee, S., Lee, G.H.: Diet-gs: Diffusion prior and event stream-assisted motion deblurring 3d gaussian splatting. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 21739–21749 (2025)
19. Lee, S., Lee, G.H.: Segment any events with language. arXiv preprint arXiv:2601.23159 (2026)
20. Lee, S., Zhao, Y., Lee, G.H.: Segment any 3d object with language. arXiv preprint arXiv:2404.02157 (2024)
21. Lin, C., Mu, Y.: Instructscene: Instruction-driven 3d indoor scene synthesis with semantic graph prior. In: International Conference on Learning Representations (ICLR) (2024)
22. Liu, Y., Li, X., Zhang, Y., Qi, L., Li, X., Wang, W., Li, C., Li, X., Yang, M.H.: Controllable 3d outdoor scene generation via scene graphs. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 28052–28062 (October 2025)
23. Nyffeler, J., Tombari, F., Barath, D.: Hierarchical 3d scene graphs construction outdoors. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 26817–26826 (October 2025)
24. Özsoy, E., Czempiel, T., Holm, F., Pellegrini, C., Navab, N.: LABRAD-OR: lightweight memory scene graphs for accurate bimodal reasoning in dynamic operating rooms. In: Greenspan, H., Madabhushi, A., Mousavi, P., Salcudean, S.E., Duncan, J., Syeda-Mahmood, T.F., Taylor, R.H. (eds.) Medical Image Computing and Computer Assisted Intervention - MICCAI 2023 - 26th International Conference, Vancouver, BC, Canada, October 8-12, 2023, Proceedings, Part IX. Lecture Notes in Computer Science, vol. 14228, pp. 302–311. Springer (2023). https://doi.org/10.1007/978-3-031-43996-4_29
25. Özsoy, E., Örnek, E., Eck, U., Czempiel, T., Tombari, F., Navab, N.: 4d-or: Semantic scene graphs for or domain modeling. In: Wang, L., Dou, Q., Fletcher, P., Speidel, S., Li, S. (eds.) Medical Image Computing and Computer Assisted Intervention – MICCAI 2022 - 25th International Conference, Proceedings. pp. 475–485. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), Springer Science and Business Media Deutschland GmbH (2022). https://doi.org/10.1007/978-3-031-16449-1_45
26. Seiwald, P., Wu, S.C., Sygulla, F., Berninger, T.F.C., Staufenberg, N.S., Sattler, M.F., Neuburger, N., Rixen, D., Tombari, F.: Lola v1.1 – an upgrade in hardware

and software design for dynamic multi-contact locomotion. In: 2020 IEEE-RAS 20th International Conference on Humanoid Robots (Humanoids). IEEE (2021). https://doi.org/10.1109/humanoids47582.2021.9555790
27. Sonogashira, M., Iiyama, M., Kawanishi, Y.: Towards open-set scene graph generation with unknown objects. IEEE Access **10**, 11574–11583 (2022). https://doi.org/10.1109/ACCESS.2022.3145465
28. 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. pp. 3961–3970 (2020)
29. Wald, J., Navab, N., Tombari, F.: Learning 3d semantic scene graphs with instance embeddings. International Journal of Computer Vision **130**(3), 630–651 (2022)
30. Wang, Z., Lee, S., Dai, G., Lee, G.H.: D3d-vlp: Dynamic 3d vision-language-planning model for embodied grounding and navigation. arXiv preprint arXiv:2512.12622 (2025)
31. Wang, Z., Lee, S., Lee, G.H.: Dynam3d: Dynamic layered 3d tokens empower vlm for vision-and-language navigation. arXiv preprint arXiv:2505.11383 (2025)
32. 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. pp. 21560–21569 (2023)
33. Werby, A., Huang, C., Büchner, M., Valada, A., Burgard, W.: Hierarchical open-vocabulary 3d scene graphs for language-grounded robot navigation. Robotics: Science and Systems (2024)
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. 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. pp. 7515–7525 (2021)
36. Yang, Z., Lu, K., Zhang, C., Qi, J., Jiang, H., Ma, R., Yin, S., Xu, Y., Xing, M., Xiao, Z., et al.: Mmgdreamer: Mixed-modality graph for geometry-controllable 3d indoor scene generation. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 9391–9399 (2025)
37. 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 (October 2025)
38. Yin, H., Wei, H., Xu, X., Guo, W., Zhou, J., Lu, J.: Gc-vln: Instruction as graph constraints for training-free vision-and-language navigation. arXiv preprint arXiv:2509.10454 (2025)
39. Yin, H., Xu, X., Wu, Z., Zhou, J., Lu, J.: SG-nav: Online 3d scene graph prompting for LLM-based zero-shot object navigation. In: The Thirty-eighth Annual Conference on Neural Information Processing Systems (2024), https://openreview.net/forum?id=HmCmxbCpp2
40. Zhai, G., Örnek, E.P., Chen, D.Z., Liao, R., Di, Y., Navab, N., Tombari, F., Busam, B.: Echoscene: Indoor scene generation vianbsp;information echo overnbsp;scene graph diffusion. In: Computer Vision – ECCV 2024: 18th European Conference, Milan, Italy, September 29–October 4, 2024, Proceedings, Part XXI. p. 167–184. Springer-Verlag, Berlin, Heidelberg (2024). https://doi.org/10.1007/978-3-031-72664-4_10

41. Zhai, G., Örnek, E.P., Wu, S.C., Di, Y., Tombari, F., Navab, N., Busam, B.: Commonscenes: Generating commonsense 3d indoor scenes with scene graphs. In: Thirty-seventh Conference on Neural Information Processing Systems (2023), https://openreview.net/forum?id=1SF2tiopYJ
42. Zhang, C., Yu, J., Song, Y., Cai, W.: Exploiting edge-oriented reasoning for 3d point-based scene graph analysis. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9705–9715 (2021)
43. Zhang, C., Delitzas, A., Wang, F., Zhang, R., Ji, X., Pollefeys, M., Engelmann, F.: Open-vocabulary functional 3d scene graphs for real-world indoor spaces. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2025)
44. Zhang, Y., Qian, D., Li, D., Pan, Y., Chen, Y., Liang, Z., Zhang, Z., Liu, Y., Mei, J., Fu, M., Ye, Y., Liang, Z., Shan, Y., Du, D.: Graphad: Interaction scene graph for end-to-end autonomous driving. In: Kwok, J. (ed.) Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence, IJCAI-25. pp. 2422–2430. International Joint Conferences on Artificial Intelligence Organization (8 2025). https://doi.org/10.24963/ijcai.2025/270, main Track
45. Zhao, Y., Lv, W., Xu, S., Wei, J., Wang, G., Dang, Q., Liu, Y., Chen, J.: Detrs beat yolos on real-time object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 16965–16974 (June 2024)