

# Statistical Confidence Rescoring for Robust 3D Scene Graph Generation from Multi-View Images

Qi Xun Yeo Yanyan Li Gim Hee Lee

Department of Computer Science, National University of Singapore

{qixunyeo, yan.li, gimhee.lee}@nus.edu.sg

## Abstract

Modern 3D semantic scene graph estimation methods utilize ground truth 3D annotations to accurately predict target objects, predicates, and relationships. In the absence of given 3D ground truth representations, we explore leveraging only multi-view RGB images to tackle this task. To attain robust features for accurate scene graph estimation, we must overcome the noisy reconstructed pseudo point-based geometry from predicted depth maps and reduce the amount of background noise present in multi-view image features. The key is to enrich node and edge features with accurate semantic and spatial information and through neighboring relations. We obtain semantic masks to guide feature aggregation to filter background features and design a novel method to incorporate neighboring node information to aid robustness of our scene graph estimates. Furthermore, we leverage on explicit statistical priors calculated from the training summary statistics to refine node and edge predictions based on their one-hop neighborhood. Our experiments show that our method outperforms current methods purely using multi-view images as the initial input. Our project page is available at <https://qixun1.github.io/projects/SCRSSG>.

## 1. Introduction

The semantic scene graph (SSG) is a crucial intermediate representation that enhances higher-level scene understanding. It plays a key role in tasks such as image captioning [17, 34, 37], image retrieval [13, 22, 23, 38], image editing [3, 5, 43], and medical applications [7, 12, 15, 20, 39] by capturing both semantic and, more importantly, relational information between objects and their surroundings. Initially developed for 2D images, SSG has since expanded to the 3D domain [6, 9, 19]. 3D scene graphs provide a high-level representation of an entire 3D scene using inputs such as multi-view RGB images or LiDAR point clouds. Unlike their 2D counterparts, they incorporate spatial relation-

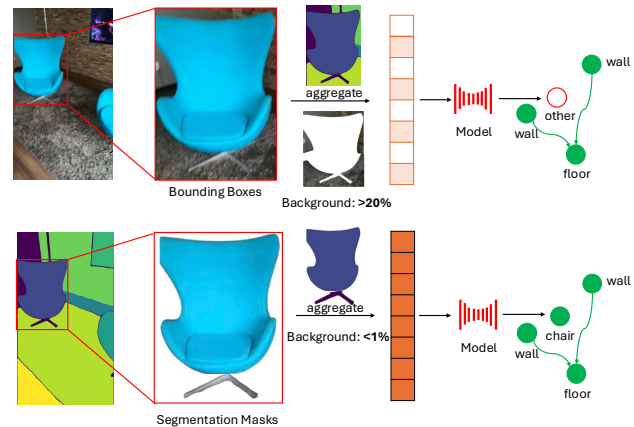


Figure 1. Existing multi-view RGB methods aggregate features using bounding boxes, introducing background noise that hinders accuracy. We use pretrained segmentation masks reduce this interference and thus improving prediction accuracy.

ships that extend beyond the visible image plane, allowing a holistic understanding of complex environments.

Many approaches [1, 28, 29] rely on ground-truth 3D LiDAR point clouds as input. However, LiDAR sensors are resource-intensive and lack inherent semantic richness. Recent works [31, 32] have shifted toward using semantically rich multi-view RGB images, which offer a balance between computational efficiency and improved semantic fidelity. Although point cloud-based methods remain valuable for capturing geometric information, our approach focuses on leveraging multi-view RGB images. By incorporating pretrained depth estimators, we generate pseudo point-based geometry to estimate the 3D semantic scene graph, reducing reliance on LiDAR while preserving spatial and structural details.

A key challenge in using only multi-view RGB images is ensuring that node features remain free from distractors. As shown in Fig. 1, prior approaches that rely on entity detectors to extract 2D bounding box proposals or regions of interest often fail to guarantee robustness since distract-

tors within the bounding box are often mistakenly aggregated. Ensuring robust initial multi-view features is crucial without ground truth point-based geometry for error correction. No explicit mechanism exists to correct incorrectly bounded objects or misclassified rare classes that are not well trained on the model. To alleviate this problem, a mechanism should exist to refine estimates using prior knowledge instead of depending solely on implicit interactions between nodes and edges to achieve confident and accurate predictions.

In this paper, we propose a framework to enhance the robustness of the model. First, we introduce a *masked feature initialization* (MFI), which leverages segmentation masks to aggregate image features instead of relying on the bounding box proposals. This reduces background noise, which results in cleaner multi-view image features. Next, we design a robust *residual spatial neighbor graph neural network* (RSN-GNN) to encode spatial information into node features. This network filters highly activated regions from neighboring nodes, refining target node features for improved predicate estimation. Finally, we propose a *confidence rescoring* (CR) module, which refines object and predicate estimates using an inverse softmax-weighted contribution of neighboring node-to-node and node-to-edge co-occurrence counts. By integrating this explicit inductive bias, our approach improves the accuracy of the prediction, particularly in low-confidence scenarios. Extensive experimental results on the benchmark 3RScan dataset show the competitiveness of our proposed approach compared to existing 2D and 3D approaches.

Our main contributions can be summarized as follows:

- We introduce a masked feature initialization to enhance the robustness of node features by reducing background distractors, yielding cleaner multi-view image features.
- We design a novel GCN architecture that integrates highly activated neighboring features into the target node to increase the robustness of the edge features.
- We propose a new refinement module to explicitly refine predictions based on statistical prior knowledge.
- Experiments show that we outperform previous state-of-the-art approaches on the 3RScan dataset, particularly on metrics that deal with low-tail imbalanced classes.

## 2. Related Work

**2D Semantic Scene Graph.** 2D semantic scene graph prediction is typically categorized into two-step and one-step approaches. The two-step approach first detects objects and then classifies their relationships [2, 4, 35, 40]. In contrast, the one-step approach jointly infers object and relationship classes [33]. Xu et al. [33] pioneers the problem of scene graph generation and they tried to solve this problem via iterative message passing. Baier et al. [2] first showed

how semantic models can be improved by incorporating triplet frequencies. Zellers et al. [40] analyzes the usefulness of statistical co-occurrences for the Visual Genome dataset and concluded that such statistical priors serves as strong regularization for the task. Chen et al. [4] formalizes the first approach (KERN) to incorporate the statistical prior directly into graph neural networks. Sharifzadeh et al. proposes Schemata to assimilate image-based relational prior knowledge into the representations within the neural network [24]. Compared to previous methods, our method combines predictions from the prediction head with statistical co-occurrence of the current node and neighbors to determine the final class using a confidence score. Unlike Sonogashira et al. [25] which uses hard thresholding, our approach uses adaptive relationship and node refinements.

**3D Semantic Scene Graph.** Methods for generating a 3D semantic scene graph can be categorized into two types based on input data. The first relies on 3D inputs such as ground-truth 3D geometry to learn contextual information [28, 30, 41]. The second relies on multi-view RGB images with some methods incorporating depth information [31] while others do not [32]. Our approach follows the latter, leveraging the estimated depth from off-the-shelf depth estimators [36] to enhance the information provided by multi-view RGB images. Wald et al. [28] first proposed 3RScan: a richly annotated 3D scene graph dataset and a method that focuses on pairwise relationships to predict the 3D scene graphs. Wu et al. [31] uses SLAM to obtain dense point representations and a novel graph convolutional network (GCN) based aggregation function to enhance 3D scene graph prediction. Zhang et al. [41] proposes an edge-oriented GCN to incorporate multi-dimensional edge features for explicit inter-node relationship modeling. Zhang et al. [42] integrates prior commonsense knowledge by learning meta-embeddings only from the class labels with their graphical structures to avoid perceptual errors. Wang et al. [30] incorporated additional information by distilling knowledge learned by the multi-modal oracle model to the 3D model. Wu et al. [32] introduces a novel entity association method to obtain the 3D entities from the predicted 2D entities and a geometric gate to incorporate geometric information to multi-view image features. Feng et al. [8] uses ConceptNet external knowledge base to accumulate both contextualized visual content and textual facts to form a 3D spatial multimodal knowledge graph. Our method leverages on the explicit statistical priors similar to 2D scene graph method KERN and Schemata while guiding low confidence predictions made by the initial node and edge predictors using both visual and geometric cues.

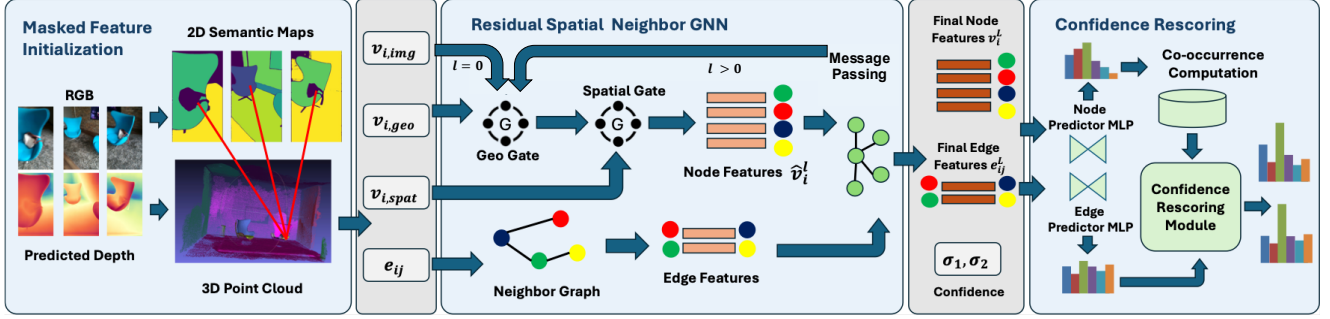


Figure 2. The architecture of our proposed framework consists of three main components: 1) A masked feature initialization step leverages backprojected semantic masks to aggregate multi-view image, geometric, and spatial features. 2) A message passing step utilizes our residual spatial neighbor GNN module to refine node and edge features through feature interaction. 3) A confidence rescoring step adjusts logits based on summary statistics of object and predicate co-occurrence from the training dataset.

### 3. Our Methodology

**Problem Definition.** Given either a set of 3D point clouds  $P$  or multi-view RGB images  $I$  of a 3D scene as input, the goal is to estimate the 3D scene graph denoted as:

$$G = (O, R), \quad (1)$$

where

- $N$ : Number of multi-view input images;
- $O = \{o_m\}_{m=1}^M$ : All objects in the scene;
- $R = \{r_{k \rightarrow m}\}_{m,k=1}^{M,M}$ : Relationships between all objects;
- $M = |O|$ : Number of objects in the scene.

Furthermore, we refer to  $o_m$  as the object class for node instance  $m$ ,  $r_{k \rightarrow m}$  as the predicate class for the edge connecting nodes  $k$  and  $m$ . Alternatively, the 3D scene graph can be defined as a set of triplets given by:  $\{o_k, r_{k \rightarrow m}, o_m\} := \{\text{node, predicate, node}\}$ .

**Overview.** Fig. 2 shows an illustration of the architecture of our proposed framework. The input to our framework is a set of  $N$  multi-view RGB images  $\{I_i\}_{i=1}^N$  of the 3D scene. Since the focus of our work is not on multi-view 3D reconstruction, we recover the 3D point cloud of the scene by running an off-the-shelf RGB-D SLAM system [26] with the metric depth maps of the input images  $\{\hat{D}_i\}_{i=1}^N$  obtained from the pretrained Depth Anything [36] model.

From the predicted 3D point cloud, we initialize multi-view image features using segmentation masks, extract geometric features with a point-based feature extractor, and derive spatial features via a linear layer. These multi-view image features serve as the initial input to our RSN-GNN module. At the GCN initialization, we enhance node features by integrating geometric and spatial features through geometric and spatial gates. We additionally enhance edge features by integrating max pooled neighboring nodes features. These enhanced node and edge features undergo message passing, with the integration process repeating at each

layer. After  $k$  iterations, the final node and edge features are processed by predictors to generate initial estimates, which are then refined using precomputed co-occurrence statistics via our CR module, producing the final scene graph output.

We elaborate on the feature initialization step in Sec. 3.1, the message passing step in Sec. 3.2, and the confidence rescoring step in Sec. 3.3.

#### 3.1. Masked Feature Initialization (MFI)

We initialize the node features of the scene graph by combining multi-view image representations from a pretrained image encoder with PointNet features derived from the predicted points in the RGB-D SLAM reconstruction step that utilizes predicted depth. To eliminate background clutter, we employ the Segment Anything Model (SAM) [14] to produce precise 2D semantic masks for each object. For each node, we backproject the associated points onto the 2D instance mask to identify its corresponding region. We then generate object proposals from these instance masks, resize them to  $224 \times 224$ , and extract their multi-view image features using the pretrained encoder. Finally, we apply these semantic masks to the multi-view image features before aggregating them into our node representations.

Our method to get the object proposals, *i.e.* the nodes of the scene graph offers two key advantages over relying solely on the entity detector from SceneGraphFusion and JointSSG [31, 32]. First, the entity detector often captures only part of an object. Although SceneGraphFusion propagates the label for this segment, the feature extractor can operate only on the detected portion. In contrast, our method employs more complete object masks, enabling the feature extractor to derive features from the entire object. Second, we exclude extraneous background features by applying these masks before feature aggregation. In SceneGraphFusion and JointSSG, the bounding boxes of the pretrained entity detector frequently include background regions, which can dilute the quality of the features and hinder

the performance. Our method is closer to the open vocabulary method ConceptGraphs [10] which uses SAM masks but with the aggregation method provided by JointSSG.

Formally, we compute a corresponding multi-view image feature  $\mathbf{v}_{i,img}$ , a geometric feature  $\mathbf{v}_{i,geo}$ , and a spatial feature  $\mathbf{v}_{i,spat}$  for each node  $v_i$ . We aggregate features only for nodes that are visible in the covisibility graph<sup>1</sup>  $G_v$  denoted by  $\mathbf{1}(v_i \in I_k)$  and have corresponding pixels within the semantic mask provided by SAM. The aggregated multi-view image feature  $\mathbf{v}_{i,img}$  is given by:

$$\mathbf{v}_i^0 = \frac{\sum_{k=1}^N \mathbf{1}(v_i \in I_k) f(\eta_k \odot x_{i,k})}{|k \in G_v|}, \quad (2)$$

which also serves as the initial node feature.  $f(\cdot)$  is the image feature encoder and  $\eta_k$  refers to the 2D semantic mask.  $x_{i,k}$  refers to the cropped image patch containing the node  $i$  from the  $k$ -th image.  $\odot$  refers to the element-wise multiplication operator. The aggregation is the arithmetic mean of the node features for each image in the covisibility graph.

The geometric feature  $\mathbf{v}_{i,geo}$  is extracted using a vanilla PointNet [21], with inputs derived from the 3D instance masks generated during the RGB-D SLAM step. For the spatial features, we employ a simple linear layer to transform the attributes of each 3D bounding box: dimensions  $\mathbf{b}_i$ , volume  $s_i$ , length  $l_i$ , and the ratio of its x- and y-axes  $r_i$  into the spatial feature as:

$$\mathbf{v}_{i,spat} = g_s([\mathbf{b}_i, s_i, l_i, r_i]), \quad (3)$$

where  $g_s$  refers to a set of learnable weights for spatial features.  $[\cdot]$  refers to the concatenation of the inputs within the operator.

### 3.2. Residual Spatial Neighbor GNN Module

Following a similar design to the learnable geometric gate in JointSSG [32], we integrate both geometric and spatial features into the initial node feature, which originally consists only of multi-view image features. The spatial features serve as additional information for individual nodes, analogous to the embeddings used for token encoding in Transformer architectures [27]. We define the geometric gate as:

$$\hat{\mathbf{v}}_i^l = \mathbf{v}_i^l + \sigma(\mathbf{w}_g^T [\mathbf{v}_i^l, \mathbf{v}_{i,geo}]) \cdot \sigma(\mathbf{v}_{i,geo}), \quad (4)$$

where  $\mathbf{w}_g$  refers to a set of learnable weights,  $\sigma$  refers to the sigmoid operator and  $l$  refers to the layer of the node feature. Furthermore, we define the spatial gate as:

$$\hat{\mathbf{v}}_i^l = \hat{\mathbf{v}}_i^l + \sigma(\mathbf{w}_s^T [\hat{\mathbf{v}}_i^l, \mathbf{v}_{i,spat}]) \cdot \sigma(\mathbf{v}_{i,spat}), \quad (5)$$

where  $\mathbf{w}_s$  is a set of learnable weights for spatial features.

<sup>1</sup>The covisibility graph is obtained as intermediate output after running the RGB-D SLAM system. It refers to a bipartite graph where each image  $I_k$  is connected with the specific node instance  $v_i$ .

Inspired by the residual connections in ResNet [11], we introduce a mechanism to integrate highly activated neighboring features into the target node feature using max pooling during the message-passing process for edge features. This approach implicitly encodes neighboring information, improving edge prediction for scene graph estimation. Specifically, we augment the edge messages in the GCN by integrating max-pooled neighboring features before message passing as follows:

$$\tilde{\mathbf{v}}_i^l = \hat{\mathbf{v}}_i^l + \max([\hat{\mathbf{v}}_j^l]_{j \in n(i)}), \quad (6)$$

where  $\max(\cdot)$  refers to the max pooling operator and  $n(i)$  refers to the neighboring nodes of node  $i$ . Max pooling allows highly activated regions in the different neighboring node features to be fused into the target node feature via the residual connection.

We apply max-pooled neighboring features to the calculation of edge messages without using them in node message updates. Since node features are highly susceptible to noise and spurious correlations from other nodes [16], incorporating max-pooled neighboring features into node message computations can introduce confusion instead of improving performance. The edge message  $m_{i \rightarrow j}$  is thus defined as:

$$m_{i \rightarrow j} = g_e([\tilde{\mathbf{v}}_i^l, \mathbf{e}_{ij}^l, \tilde{\mathbf{v}}_j^l]), \quad (7)$$

where  $g_e(\cdot)$  is an MLP and  $\mathbf{e}_{ij}^l$  is the edge descriptor used in [32] containing relative node properties such as centroid displacement, relative bounding box size difference, and a relative pose descriptor.

### 3.3. Confidence Rescoring (CR) Module

As discussed in Sec. 1, modern scene graph methods depend heavily on the entity detector and image feature extractor to produce an accurate initial object prediction. However, there is no mechanism to correct a target bounding box prediction that is initially wrong. To address this limitation, we leverage relationships with neighboring nodes, allowing us to mitigate the impact of incorrect initial predictions by upweighting contributions from neighboring node statistics to refine target object classification. Fig. 3 illustrates the logits of the final prediction after applying our confidence rescoring module. This module incorporates prior knowledge to improve prediction confidence beyond that of the original node and edge multi-layer perceptron (MLP) predictor.

The probability of predicting node  $i$  given its neighboring nodes  $n(i)$  is approximated by the output of our node predictor  $h_v(\cdot)$  as:

$$\begin{aligned} P(o_i | \{o_j : j \in n(i)\}) &= \frac{1}{Z_i} \psi_i(o_i) \prod_{j \in n(i)} \psi_{ij}(o_i, o_j) \\ &\approx \tau(h_v(\mathbf{v}^L)), \end{aligned} \quad (8)$$

where  $\psi_i(o_i)$  refers to the potential associated with node  $i$  and  $\psi_{ij}(o_i, o_j)$  refers to the potential associated with the edge between node  $i$  and its neighbor node  $j$ . The potentials are implicitly learned through the GCN unlike classical methods, which explicitly learn these potentials.  $h_v(\cdot)$  is a small MLP that predicts the class of each node and  $\tau$  refers to the softmax operator.  $L$  refers to the last layer of the GCN message passing output. The partition function given by:

$$Z_i = \sum_{o_i} \psi_i(o_i) \prod_{j \in n(i)} \psi_{ij}(o_i, o_j) \quad (9)$$

ensures the probabilities sum to 1. A similar equation can be obtained for the probability of predicting edge  $ij$  given its neighboring nodes  $i$  and  $j$ , *i.e.*:

$$P(r_{i \rightarrow j} | o_i, o_j) = \frac{1}{Z_{ij}} \psi_{ij}(o_i, o_j, r_{i \rightarrow j}) \approx \tau(h_e(\mathbf{e}^L)), \quad (10)$$

where  $h_e(\cdot)$  is a small MLP which predicts the class of each edge, and the partition function to normalize the probabilities is given by:

$$Z_{ij} = \sum_{r_{i \rightarrow j}} \psi_{ij}(o_i, o_j, r_{i \rightarrow j}). \quad (11)$$

We begin by computing summary statistics from the training dataset to node-to-node co-occurrences denoted as  $c_{ij}(o_i, o_j)$ , and node-to-edge co-occurrences denoted as  $d_{ij}(o_i, e_j)$ . Subsequently, we determine the prediction confidence of the model, where  $\alpha_v$  represents node confidence and  $\alpha_e$  represents edge confidence. Next, we obtain the node  $\alpha_v$  and edge  $\alpha_e$  confidences as:

$$\alpha_v = \max_{\mathbf{v}} \tau(h_v(\mathbf{v}^L)), \quad \alpha_e = \max_{\mathbf{e}} \tau(h_e(\mathbf{e}^L)). \quad (12)$$

We first estimate the marginal probability of a specific instantiation of node  $i$ , conditioned on node  $j$  using node-to-node co-occurrences, and on edge  $ij$  using node-to-edge co-occurrences defined as:

$$c(o_i | o_j) = [c(o_i = c_k | o_j = c_l)]_{k=0, l=0}^{C, C}, \quad \text{where} \quad (13)$$

$$c(o_i = c_k | o_j = c_l) = \frac{c(o_i = c_k, o_j = c_l)}{\sum_{k=0}^C c(o_i = c_k, o_j = c_l)}.$$

Here,  $C$  refers to the number of object classes and  $c_k$  refers to the  $k$ -th class in the set of object classes. We then integrate the summary statistics via an inverse softmax operation denoted as  $\gamma(\cdot)$  into the node probability in Eq. 14 to get:

$$P(o_i | \{o_j : j \in n(i)\}) = \tau(\alpha_v \cdot h_v(\mathbf{v}_i) + (1 - \alpha_v) \cdot \sum_{j \in n(i)} \alpha_j \cdot \gamma(\mathbf{c}(o_i | o_j))). \quad (14)$$

This formulation is chosen because directly adding the estimated marginal probability to the predictions of the model after softmax would disregard the required normalization, resulting in incompatible terms.

Similarly, we compute the estimated marginal probability of an edge  $ij$  instantiation conditioned on nodes  $i$  and  $j$  using node-to-edge co-occurrence statistics for predicate estimation:

$$\mathbf{d}(r_{i \rightarrow j} | o_j) = [d(r_{i \rightarrow j} = e_k | o_j = c_l)]_{k=0, l=0}^{P, C}, \quad \text{where} \quad (15)$$

$$d(r_{i \rightarrow j} = e_k | o_j = c_l) = \frac{d(r_{i \rightarrow j} = e_k, o_j = c_l)}{\sum_{k=0}^P d(r_{i \rightarrow j} = e_k, o_j = c_l)}.$$

Here,  $P$  refers to the total number of predicate classes and  $e_k$  refers to the  $k$ -th class in the set of predicate class. The refined logits for predicate estimation are then given by:

$$P(r_{i \rightarrow j} | o_i, o_j) = \tau(\alpha_{ij} \cdot h_e(\mathbf{e}_{ij}) + (1 - \alpha_{ij}) \cdot (\alpha_i \cdot \gamma(\mathbf{d}(r_{i \rightarrow j} | o_i))) \cdot (\alpha_j \cdot \gamma(\mathbf{d}(r_{i \rightarrow j} | o_j)))). \quad (16)$$

**Remarks.** As shown in Fig. 3, the original node probability for the correct class is already relatively high. However, our CR module further enhances the predicted probability as evidenced by the upward shift in the box plot from ‘‘Original’’ to ‘‘After CR’’ in the red boxes, with a notable increase in the lower quartile. Moreover, a larger proportion of low node probability instances are flagged as outliers as shown by the greater number of red circles at ‘‘After CR’’ compared to ‘‘Original’’. Most notable gains occur when original logits are low since CR increases the probability of the correct class by down-weighting misclassified classes  $c'$  via lower conditional probabilities  $P(c' | c)$  given conditioned class  $c$ . This targeted adjustment raises suppressed logits for rare classes, thus improving prediction performance for under-represented categories.

## 4. Experiments

### 4.1. Experimental Setup

**Datasets.** We evaluate on the 3RScan dataset following the same training and test split from [28]. The image quality of the dataset is rather poor as each scene constitutes a variable number of short videos spliced together with different motion trajectories. As the images are obtained from consumer-level cameras, the images possess a low frame rate of 10 Hz, image blur, and image jitter from sudden camera motion which makes this dataset challenging.

**Implementation Details.** We utilize the codebase provided from the 3DSSG GitHub repository<sup>2</sup> to reproduce the prior

<sup>2</sup><https://github.com/ShunChengWu/3DSSG>

Method	bath.	bed	bkskf	cab.	chair	cntr.	curt.	desk	door	floor	ofurn	pic.	refri.	show.	sink	sofa	table	toil.	wall	wind.	Mean.
IMP	0.000	<b>1.000</b>	0.000	0.341	0.467	0.000	<b>0.606</b>	0.200	<b>0.528</b>	0.900	0.125	0.143	0.000	0.000	0.333	0.450	0.525	<b>0.800</b>	0.622	0.125	0.358
VGIM	0.500	0.000	0.000	0.415	0.543	0.083	0.545	0.000	0.389	<b>0.920</b>	0.232	0.086	0.000	0.000	0.381	0.450	0.574	<b>0.800</b>	0.644	0.167	0.336
3DSSG	0.000	0.667	0.000	0.207	0.348	0.167	0.576	0.200	0.361	0.780	0.125	0.057	0.000	0.000	0.333	0.500	0.279	0.000	0.467	0.125	0.260
SGFN	0.500	0.333	0.000	0.293	<b>0.685</b>	0.250	0.515	0.200	0.444	0.880	0.232	0.171	0.000	0.000	<b>0.429</b>	0.500	0.508	0.400	<b>0.707</b>	0.125	0.359
JointSSG	<b>1.000</b>	<b>1.000</b>	0.000	0.476	0.674	0.250	0.576	0.200	0.500	0.900	0.107	0.143	0.000	0.000	<b>0.429</b>	0.550	0.541	<b>0.800</b>	0.680	0.250	0.454
Ours	<b>1.000</b>	<b>1.000</b>	<b>0.333</b>	<b>0.622</b>	0.630	<b>0.333</b>	0.545	<b>0.600</b>	<b>0.528</b>	0.880	<b>0.304</b>	<b>0.171</b>	<b>0.667</b>	<b>0.667</b>	0.333	<b>0.600</b>	<b>0.623</b>	<b>0.800</b>	0.693	<b>0.375</b>	<b>0.605</b>

Table 1. Comparison with state-of-the-art methods on the 3RScan dataset for each object class. The **Best** results are highlighted.

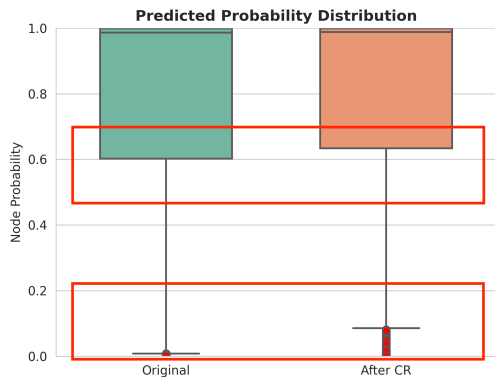


Figure 3. Box plot for predicted node probability for correct class after our CR module. The lower quartile visibly moves upwards due to an increase in predicted node probability for correct class.

methods. For JointSSG<sup>‡</sup> [32], we replace their image feature extractor from a pretrained ResNet18 to a pretrained DINOv2 with a ViT-L backbone to ensure fair comparison.

Following [32], we use PointNet [21] as the point encoder and employ two message-passing layers. For RGB-D SLAM reconstruction, we utilize the method from [26] and our predicted depth to generate predicted points. Multi-view feature extraction is performed using a pretrained DINOv2 [18] with a ViT-L backbone. To obtain object segmentation masks, we use a pretrained SAM predictor with a text prompt containing the class set of 20 NYU object classes following the subset defined in [28]. Node-to-node and node-to-edge co-occurrences are precomputed on the training set. We train the model on a single RTX 3090 GPU for two days with early stopping.

**Evaluation Metrics.** Following [29] and JointSSG[32], we 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. We map all predictions on the estimated segments to the ground truth segments for a fair comparison between the different methods.

## 4.2. Quantitative Results

We present the per-class recall for 20 object classes in the 3RScan dataset in Tab. 1, demonstrating that our method significantly outperforms others across most object cate-

gories. Notably, we are the only approach that achieves a recall above 0% for all object classes, indicating that we **robustly learn all classes** instead of just a subset. Our method tends to underperform slightly compared to the other methods on majority classes such as *floor* and *wall* classes in return, although not by a significant margin. The main experimental results for the top-1 recall and the mean recall are shown in Tab. 2. Importantly, our approach significantly outperforms others in mean recall (mRecall) for both objects and predicates, suggesting that it handles class imbalance issues [29] more effectively than competing methods. Additionally, our method surpasses all baselines in relationship and object estimation metrics while achieving comparable performance in predicate estimation. More impressively, our method with the weaker ResNet18 multi-view image features already surpasses all baselines in relationship and object estimation metrics except for predicate estimation. Our method with ResNet18 image features even manages to surpass the performance of JointSSG with DINOv2 multi-view image features by non-trivial margins. This shows that our improved performance is not purely due to the more powerful DINOv2 features. Refer to our supplementary material for more quantitative results.

## 4.3. Qualitative Results

We show qualitative results comparing our method with other competitive methods in Fig. 4. Since the scene contains a large number of objects, we display only a subset with fewer objects and predicates for better readability. The correct prediction of the *None* predicate class is omitted in the visualization due to its prevalence in the dataset. As seen in Fig. 4, our method can distinguish the relatively more ambiguous wall which the other methods fail to distinguish. The shower curtain is tricky to classify since it is a subset of a curtain that is another class of objects present in the dataset. Understandably, all methods fail to correctly classify this object. Both our method and JointSSG attain good performance on the predicate class compared to SGFN as both methods only misclassify the predicate between the curtain object and the wall.

We additionally evaluate on the entire scene for scan 43b8cae1 without comparison to other methods. Our method can recognize rare classes present in the scene, which are more challenging to learn. As seen in Fig. 5, our approach successfully classifies the rare class bookshelf instance denoted by *bkskf* which no other method managed

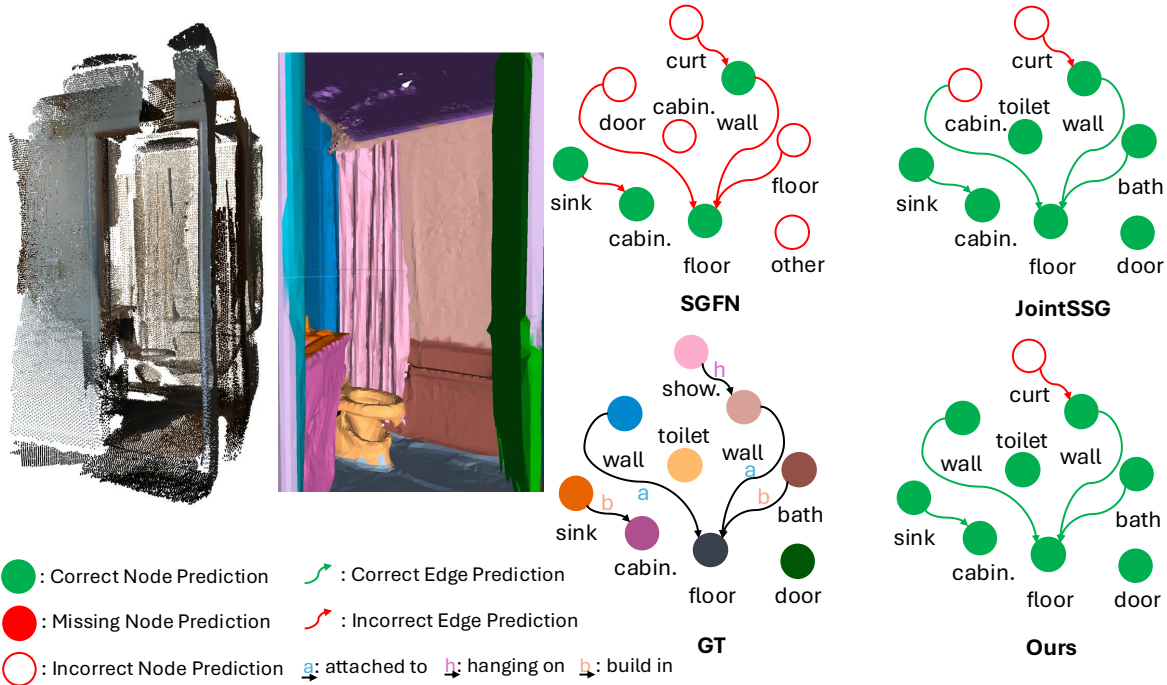


Figure 4. The qualitative results on 3RScan dataset of previous works and our proposed framework on scan 4d3d82b0. SGFN also fails to classify majority of non-background object classes except the sink and cabinet. It also fails to classify all predicate classes. JointSSG performs better but misclassifies the shower curtain as a curtain and the ambiguous blue wall as a cabinet. Our method correctly classifies the wall but misidentifies the shower curtain as a curtain.

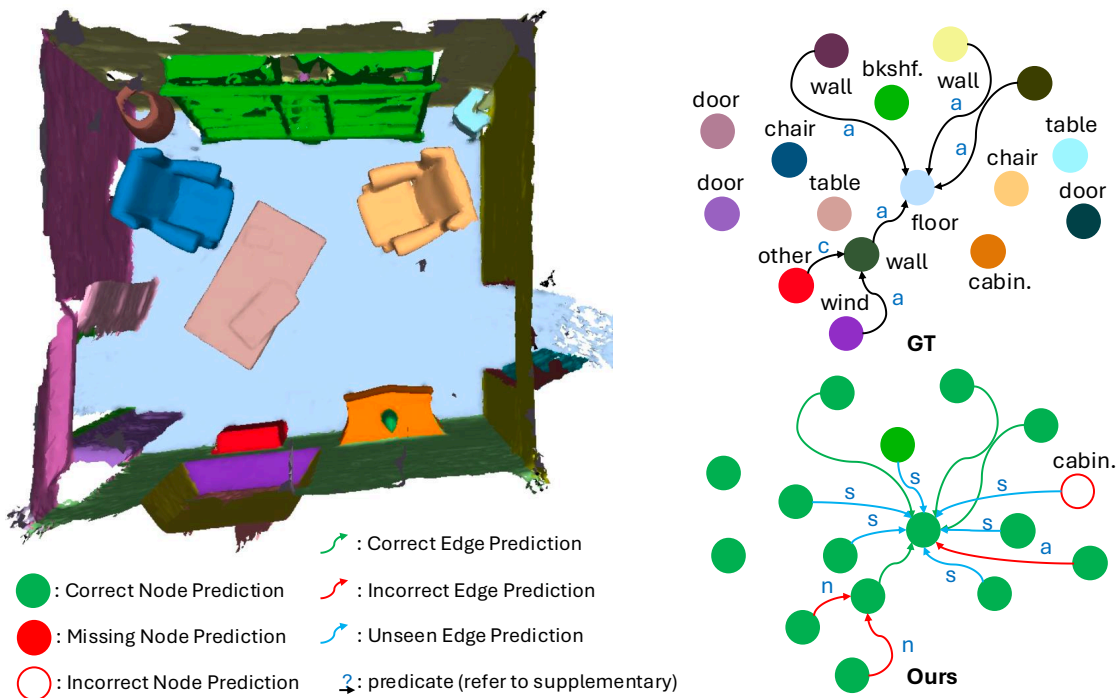


Figure 5. The qualitative results on 3RScan dataset of our proposed framework on the entire scene of scan 43b8cae1. Our method can correctly classify most objects in the scene except for the light blue dress table at the top right of the scan. Our method can also correctly classify most predicates.

Method	Recall%			mRecall%	
	Rel	Obj.	Pred.	Obj.	Pred
IMP [32]	25.8	51.8	90.4	30.0	23.0
VGfM [32]	28.3	53.3	90.7	31.6	24.4
3DSSG [32]	17.5	41.4	88.2	31.9	26.6
SGFN [32]	31.4	56.7	89.6	38.3	30.5
JointSSG [32]	34.1	58.1	89.9	43.0	33.3
IMP	25.4	47.7	89.9	35.8	19.9
VGfM	27.4	50.1	<b>90.6</b>	33.6	22.3
3DSSG	13.0	35.6	86.8	26.0	24.3
SGFN	29.9	52.2	89.5	35.9	27.1
JointSSG	32.9	54.5	88.7	45.4	35.2
JointSSG $\ddagger$	34.3	56.6	86.5	52.9	36.2
Ours $\dagger$	38.6	59.3	89.1	57.4	36.8
Ours	<b>40.5</b>	<b>61.8</b>	90.4	<b>60.5</b>	<b>39.2</b>

Table 2. Comparison with state-of-the-art methods on the 3RScan dataset with 20 object classes and 8 predicate classes. The top group of results are reported in JointSSG [32]. The middle group of results are reproduced via 3DSSG GitHub repository. JointSSG $\ddagger$  refers to JointSSG but with DINOv2 multi-view image features. Ours $\dagger$  uses ResNet18 multi-view image features; Ours uses the DINOv2 multi-view image features instead. The **Best** and **Second Best** results are highlighted, respectively.

to predict based on the results shown in Tab. 1. Our method also predicts multiple reasonable predicates of *standing on* between various objects and *floor*. Additional qualitative results can be found in the supplementary material.

#### 4.4. Ablation Studies

We analyze the effects of our feature initialization, robust neighbor residuals, and confidence rescoring module in Tab. 3. Our confidence rescoring (CR) module significantly improves performance across all metrics, indicating that it effectively leverages statistical priors to address class imbalance issues as noted in [32]. Apart from predicate mRecall, incorporating our masked feature initialization further improves performance across all metrics. This outcome is expected since masking primarily aims to refine multi-view image features, which do not directly contribute to edge feature improvements. The combination of the masked feature initialization and the CR module act as a force multiplier since the CR module benefits from cleaner and more accurate neighboring node features. Additionally, our robust neighbor residual gate maintains comparable performance across most metrics except for predicate mRecall, which benefits from additional neighbor information to better handle class imbalance in predicate estimation.

We additionally study the effects of using our CR module during inference in Tab. 4. The addition of the CR module with the prior knowledge of the dataset is akin to calculating statistics during transductive setting of inference instead

Method	Conf.	Feat.	Neigh.	Recall%			mRecall%	
				Rel.	Obj.	Pred.	Obj.	Pred.
JointSSG	×	×	×	32.9	54.5	88.2	45.4	35.2
+ Conf.	✓	×	×	37.9	57.2	89.6	54.1	36.3
+ Feat.	✓	✓	×	<b>41.1</b>	61.3	<b>90.8</b>	<b>61.3</b>	35.8
Ours	✓	✓	✓	40.5	<b>61.8</b>	90.4	60.5	<b>39.2</b>

Table 3. Ablation study on the 3RScan dataset. Conf. refers to the usage of summary statistic priors for confidence rescoring via the CR module. Feat. refers to the usage of DINOv2 image features with refinement from SAM semantic masks via MFI. Neigh. refers to the usage of RSN-GNN. The **Best** results are shown in bold.

Method	CR.	Recall%			mRecall%	
		Rel.	Obj.	Pred.	Obj.	Pred.
Ours	×	<b>41.1</b>	61.1	<b>90.6</b>	60.2	35.7
Ours	✓	40.5	<b>61.8</b>	90.4	<b>60.5</b>	<b>39.2</b>

Table 4. Ablation study on the 3RScan dataset. CR. refers to the usage of summary statistic priors for confidence rescoring via the CR module during inference. The **Best** results are shown in bold.

of a true overfitting to the dataset. Interestingly, the removal of the CR module during inference does not significantly reduce performance. It even leads to a slight increase in predicate Recall and overall relationship triplet Recall. The slight increase in predicate Recall, coupled with a drastic decrease in predicate mRecall, suggests that removing the CR module benefits the predicate classes that appear frequently in the dataset, including the *None* class, with the trade off of decreasing performance on the tail classes. Nonetheless, the use of the CR module during evaluation benefits object Recall, object mRecall, and predicate mRecall.

**Limitations.** The effectiveness of initial feature aggregation relies on the quality of the segmentation mask. Although the CR module performs well when there is no significant domain shift, the statistical prior can negatively impact performance when applied to datasets with a drastically different training data distribution. Similarly, the prior becomes less useful on datasets with very limited data.

## 5. Conclusion

We introduce a straightforward yet effective approach for enhancing 3D semantic scene graph estimation through our MFI, RSN-GNN, and CR modules. Specifically, our masked feature initialization (MFI) enhances node features by filtering out background distractors and leveraging more than just part-based image features. The RSN-GNN module strengthens edge features by integrating highly activated neighboring features into the target node feature, improving robustness. Finally, we introduce a refinement module that explicitly refines predictions using statistical prior knowledge, further enhancing performance.

## Acknowledgments

This research work is supported by the Agency for Science, Technology and Research (A\*STAR) under its MTC Programmatic Funds (Grant No. M23L7b0021), and the Tier 2 grant MOE-T2EP20124-0015 from the Singapore Ministry of Education. We also thank Shun-Cheng Wu for the helpful discussions on running the SLAM code and reproducing the results in the 3DSSG repository.

## References

- [1] Iro Armeni, Zhi-Yang He, JunYoung Gwak, Amir R Zamir, Martin Fischer, Jitendra Malik, and Silvio Savarese. 3d scene graph: A structure for unified semantics, 3d space, and camera. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5664–5673, 2019. 1
- [2] Stephan Baier, Yunpu Ma, and Volker Tresp. Improving visual relationship detection using semantic modeling of scene descriptions. In *The Semantic Web–ISWC 2017: 16th International Semantic Web Conference, Vienna, Austria, October 21–25, 2017, Proceedings, Part I 16*, pages 53–68. Springer, 2017. 2
- [3] Lichang Chen, Guosheng Lin, Shijie Wang, and Qingyao Wu. Graph edit distance reward: Learning to edit scene graph. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIX 16*, pages 539–554. Springer, 2020. 1
- [4] Tianshui Chen, Weihao Yu, Riquan Chen, and Liang Lin. Knowledge-embedded routing network for scene graph generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6163–6171, 2019. 2
- [5] Helisa Dhamo, Azade Farshad, Iro Laina, Nassir Navab, Gregory D Hager, Federico Tombari, and Christian Rupprecht. Semantic image manipulation using scene graphs. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5213–5222, 2020. 1
- [6] Helisa Dhamo, Fabian Manhardt, Nassir Navab, and Federico Tombari. Graph-to-3d: End-to-end generation and manipulation of 3d scenes using scene graphs. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16352–16361, 2021. 1
- [7] Jessica D’souza, PK Aleema, S Dhanyashree, Clita Fernandes, KM Kavitha, and Chandra Naik. Knowledge-based scene graph generation in medical field. In *2023 IEEE International Conference on Distributed Computing, VLSI, Electrical Circuits and Robotics (DISCOVER)*, pages 232–237. IEEE, 2023. 1
- [8] Mingtao Feng, Haoran Hou, Liang Zhang, Zijie Wu, Yulan Guo, and Ajmal Mian. 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*, pages 9182–9191, 2023. 2
- [9] Gege Gao, Weiyang Liu, Anpei Chen, Andreas Geiger, and Bernhard Schölkopf. Graphdreamer: Compositional 3d scene synthesis from scene graphs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21295–21304, 2024. 1
- [10] Qiao Gu, Ali Kuwajerwala, Sacha Morin, Krishna Murthy Jatavallabhula, Bipasha Sen, Aditya Agarwal, Corban Rivera, William Paul, Kirsty Ellis, Rama Chellappa, et al. Conceptgraphs: Open-vocabulary 3d scene graphs for perception and planning. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 5021–5028. IEEE, 2024. 4
- [11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 4
- [12] Felix Holm, Ghazal Ghazaei, Tobias Czempiel, Ege Özsoy, Stefan Saur, and Nassir Navab. Dynamic scene graph representation for surgical video. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 81–87, 2023. 1
- [13] Justin Johnson, Ranjay Krishna, Michael Stark, Li-Jia Li, David Shamma, Michael Bernstein, and Li Fei-Fei. Image retrieval using scene graphs. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3668–3678, 2015. 1
- [14] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023. 3
- [15] Chen Lin, Shuai Zheng, Zhizhe Liu, Youru Li, Zhenfeng Zhu, and Yao Zhao. Sgt: Scene graph-guided transformer for surgical report generation. In *International conference on medical image computing and computer-assisted intervention*, pages 507–518. Springer, 2022. 1
- [16] Xin Lin, Changxing Ding, Jing Zhang, Yibing Zhan, and Dacheng Tao. Ru-net: Regularized unrolling network for scene graph generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19457–19466, 2022. 4
- [17] Kien Nguyen, Subarna Tripathi, Bang Du, Tanaya Guha, and Truong Q Nguyen. In defense of scene graphs for image captioning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1407–1416, 2021. 1
- [18] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *Transactions on Machine Learning Research Journal*, pages 1–31, 2024. 6
- [19] Julian Ost, Fahim Mannan, Nils Thuerey, Julian Knodt, and Felix Heide. Neural scene graphs for dynamic scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2856–2865, 2021. 1
- [20] Ege Özsoy, Evin Pınar Örnek, Ulrich Eck, Tobias Czempiel, Federico Tombari, and Nassir Navab. 4d-or: Semantic scene graphs for or domain modeling. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 475–485. Springer, 2022. 1

- [21] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 652–660, 2017. 4, 6
- [22] Brigit Schroeder and Subarna Tripathi. Structured query-based image retrieval using scene graphs. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 178–179, 2020. 1
- [23] Sebastian Schuster, Ranjay Krishna, Angel Chang, Li Fei-Fei, and Christopher D Manning. Generating semantically precise scene graphs from textual descriptions for improved image retrieval. In *Proceedings of the fourth workshop on vision and language*, pages 70–80, 2015. 1
- [24] Sahand Sharifzadeh, Sina Moayed Baharlou, and Volker Tresp. Classification by attention: Scene graph classification with prior knowledge. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 5025–5033, 2021. 2
- [25] Motoharu Sonogashira, Masaaki Iiyama, and Yasutomo Kawanishi. Towards open-set scene graph generation with unknown objects. *IEEE Access*, 10:11574–11583, 2022. 2
- [26] Keisuke Tateno, Federico Tombari, and Nassir Navab. Real-time and scalable incremental segmentation on dense slam. In *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 4465–4472. IEEE, 2015. 3, 6
- [27] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 4
- [28] Johanna Wald, Helisa Dhamo, Nassir Navab, and Federico Tombari. Learning 3d semantic scene graphs from 3d indoor reconstructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3961–3970, 2020. 1, 2, 5, 6
- [29] Johanna Wald, Nassir Navab, and Federico Tombari. Learning 3d semantic scene graphs with instance embeddings. *International Journal of Computer Vision*, 130(3):630–651, 2022. 1, 6
- [30] Ziqin Wang, Bowen Cheng, Lichen Zhao, Dong Xu, Yang Tang, and Lu Sheng. Vl-sat: Visual-linguistic semantics assisted training for 3d semantic scene graph prediction in point cloud. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 21560–21569, 2023. 2
- [31] Shun-Cheng Wu, Johanna Wald, Keisuke Tateno, Nassir Navab, and Federico Tombari. Scenegrphfusion: Incremental 3d scene graph prediction from rgb-d sequences. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7515–7525, 2021. 1, 2, 3
- [32] Shun-Cheng Wu, Keisuke Tateno, Nassir Navab, and Federico Tombari. Incremental 3d semantic scene graph prediction from rgb sequences. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5064–5074, 2023. 1, 2, 3, 4, 6, 8
- [33] Danfei Xu, Yuke Zhu, Christopher B Choy, and Li Fei-Fei. Scene graph generation by iterative message passing. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5410–5419, 2017. 2
- [34] Ning Xu, An-An Liu, Jing Liu, Weizhi Nie, and Yuting Su. Scene graph captioner: Image captioning based on structural visual representation. *Journal of Visual Communication and Image Representation*, 58:477–485, 2019. 1
- [35] Jianwei Yang, Jiasen Lu, Stefan Lee, Dhruv Batra, and Devi Parikh. Graph r-cnn for scene graph generation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018. 2
- [36] Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything: Unleashing the power of large-scale unlabeled data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10371–10381, 2024. 2, 3
- [37] Xu Yang, Kaihua Tang, Hanwang Zhang, and Jianfei Cai. Auto-encoding scene graphs for image captioning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10685–10694, 2019. 1
- [38] Sangwoong Yoon, Woo Young Kang, Sungwook Jeon, SeongEun Lee, Changjin Han, Jonghun Park, and Eun-Sol Kim. Image-to-image retrieval by learning similarity between scene graphs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 10718–10726, 2021. 1
- [39] Kun Yuan, Manasi Kattel, Joël L Lavanchy, Nassir Navab, Vinkle Srivastav, and Nicolas Padoy. Advancing surgical vqa with scene graph knowledge. *International Journal of Computer Assisted Radiology and Surgery*, 19(7):1409–1417, 2024. 1
- [40] Rowan Zellers, Mark Yatskar, Sam Thomson, and Yejin Choi. Neural motifs: Scene graph parsing with global context. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5831–5840, 2018. 2
- [41] Chaoyi Zhang, Jianhui Yu, Yang Song, and Weidong Cai. Exploiting edge-oriented reasoning for 3d point-based scene graph analysis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9705–9715, 2021. 2
- [42] Shoulong Zhang, Aimin Hao, Hong Qin, et al. Knowledge-inspired 3d scene graph prediction in point cloud. *Advances in Neural Information Processing Systems*, 34:18620–18632, 2021. 2
- [43] Zhongping Zhang, Huiwen He, Bryan A Plummer, Zhenyu Liao, and Huayan Wang. Complex scene image editing by scene graph comprehension. In *BMVC*, 2023. 1