3D Spatial Multimodal Knowledge Accumulation for Scene Graph Prediction in Point Cloud

Mingtao Feng[1*] Haoran Hou[1*] Liang Zhang[1†] Zijie Wu[2†] Yulan Guo[3] Ajmal Mian[4]
[1]Xidian University, [2]Hunan University, [3]Sun Yat-Sen University, [4]The University of Western Australia

Abstract

In-depth understanding of a 3D scene not only involves locating/recognizing individual objects, but also requires to infer the relationships and interactions among them. However, since 3D scenes contain partially scanned objects with physical connections, dense placement, changing sizes, and a wide variety of challenging relationships, existing methods perform quite poorly with limited training samples. In this work, we find that the inherently hierarchical structures of physical space in 3D scenes aid in the automatic association of semantic and spatial arrangements, specifying clear patterns and leading to less ambiguous predictions. Thus, they well meet the challenges due to the rich variations within scene categories. To achieve this, we explicitly unify these structural cues of 3D physical spaces into deep neural networks to facilitate scene graph prediction. Specifically, we exploit an external knowledge base as a baseline to accumulate both contextualized visual content and textual facts to form a 3D spatial multimodal knowledge graph. Moreover, we propose a knowledge-enabled scene graph prediction module benefiting from the 3D spatial knowledge to effectively regularize semantic space of relationships. Extensive experiments demonstrate the superiority of the proposed method over current state-of-the-art competitors. Our code is available at https://github.com/HHrEtvP/SMKA.

1. Introduction

In recent years, much success has been achieved on 3D point cloud scene understanding such as semantic segmentation [9, 11, 15, 16, 21, 28, 29, 49] and object detection [10, 22, 25, 27, 43]. However, the 3D world is not only defined by objects but also by the relationships between objects. A 3D scene graph can abstract the environment as a graph where nodes represent objects and edges characterize the relationships between object pairs, which has already been recognized in recent seminal works [1, 30, 37, 38, 41, 46]. However, relationship graphs predicted by current methods are far from satisfactory due to the noisy, cluttered and partial nature of real 3D scans. Moreover, these data-driven methods treat sophisticated relationships in 3D space independently for classification using the geometric features proximity or fit, and are ignorant of commonsense or other useful 3D spatial cues beyond visual information. 3D objects in real scenes commonly have strongly structured regularities [33, 39], whose semantic and spatial arrangements follow clear patterns, but still exhibit rich structural variations even within the scene category.

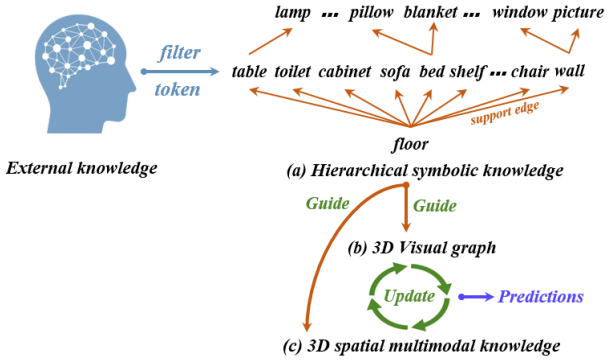

Figure 1. A brief overview of our method.

The key observation is that 3D scene structures are inherently hierarchical [20]. By definition, an instance can have multiple supports, *lamps* are standing on a *table*, *chairs* are supported by the *floor* and only the *floor* does not have any support, and it is unlikely that a *pillow* is supporting a *couch*. Although relationships themselves cast no light on the human eyes, a growing body of works [14, 31] suggest that even very complex relationship information is reasoned hierarchically and systemically according to the role of the prefrontal cortex. Relationships, such as support, can be extracted rapidly, are hard to ignore, and influence other relationships in the perceptual process. For example, a *TV* and a *sofa* are related since they together serve the function of 'watching *TV*', but these two objects can be far apart in a scene. Relationships of this kind are much more difficult, if not possible, to infer based on geometric analysis alone. The model can relate the *table* easily which supports the *TV* and use the *table* as a bridge to predict the 'front' relationship with *sofa*, where *table* and *sofa* are all supported by the *floor* and relationships within them is intuitive.

*Equal contribution
†Corresponding author

The underlying hierarchical structures in 3D scenes are label free and reliable, and can hence play an essential role in scene understanding at no additional cost. Existing 3D scene graph prediction models [1, 30, 37, 38, 41, 46] are oblivious to the underlying structures in the point cloud scenes. The question is *how to take this prior knowledge into consideration to make the 3D scene graph achieve higher accuracy?* KISG [47] proposes a graph auto-encoder to learn a closed set and ground truth prior knowledge from relationship triplets in data for 3D scene graph prediction. Although KISG [47] takes note of knowledge, it captures relevant prior knowledge from text-only ground truth labels, which merely contain facts expressed by label descriptions while lacking complex but indispensable multimodal knowledge for 3D scene graph prediction. In addition, noises contained in the manually annotated labels are easily included in the knowledge base and affects the prediction of relationships.

To address the above problems, we show that the implicit hierarchical structure correlations between object pairs and their relationships can be explicitly represented by a knowledge base. As shown in Fig. 1, we propose a 3D spatial multimodal knowledge accumulation module to explicitly merge the hierarchical structures of 3D scenes into the network to strengthen the 3D scene graph prediction process. Firstly, we filter the external commonsense knowledge base, classify the hierarchical tokens for each node, and add new support edges to form the hierarchical symbolic knowledge graph for 3D scenes. Secondly, we retrieve the hierarchical token from the reconstructed symbolic knowledge graph for object instances in 3D scenes to build a visual graph, and extract contextual features for nodes and edges using a region-aware graph network. Finally, to bridge the heterogeneous gap between the symbolic knowledge and visual information, we propose a graph reasoning network to correlate 3D spatial visual contents of scenes with textual facts. Conditioned on the learned vision-relevant 3D spatial multimodal knowledge, we incorporate this network into the relationships prediction stage as extra guidance, which can effectively regularize the distribution of possible relationships of object pairs and thus make the predictions less ambiguous.

Our main contributions are: 1) We are the first to explicitly unify the regular patterns of 3D physical spaces with the deep architecture to facilitate 3D scene graph prediction. 2) We propose a hierarchical symbolic knowledge construction module that exploits extra knowledge as the baseline to admit the hierarchical structure cues of 3D scene. 3) We introduce a knowledge-guided visual context encoding module to construct hierarchical visual graph and learn the contextualized features by a region-aware graph network. 4) We propose a 3D spatial multimodal knowledge accumulation module to regularize the semantic space of relationship prediction. Results show that the learned knowledge and proposed modules consistently boost 3D scene graph prediction performance.

2. Related Work

2D Image-based Scene Graph Generation. Scene graph was first proposed for image retrieval [17], and subsequently received increasing attention in the vision community to produce graphical abstractions of images. Mainstream approaches [5, 36, 42, 44, 45] follow a two-step pipeline that first detects objects followed by classification of the relationship for each object pair. However, research on scene graphs has focused primarily on 2D images, ignoring 3D spatial characteristics such as position and geometry, and with limited spatial coverage. Our proposed method extends 2D scene graphs to 3D spaces, where the scene representation, network architecture and training mechanism all have to be altered in fundamental ways to meet the challenges arising from learning 3D scene structures and relationships. More detailed discussions can be found in the survey [4].

Knowledge Representation has been extensively studied to incorporate prior knowledge, e.g. DBPedia [2], ConceptNet [35], WordNet [24], VisualGenome [19] and hasPart [3], to aid numerous vision tasks [23]. Gao et al. [12] incorporated commonsense knowledge to learn the internal-external correlations among room and object entities for an agent to take proper decisions at each viewpoint. Zhang et al. [48] addressed the explainability of visual reasoning by introducing the explicit integration of external knowledge. Ding et al. [8] extracted the multimodal knowledge triplet to boost the performance of visual question answering. Chen et al. [6] constructed the prior knowledge of statistical correlations between object pairs and their relationships to address the issue of the uneven distribution over different relationships. Although previous studies have taken notice of knowledge in different vision tasks, they only implicitly mine the extra knowledge base or count the frequency of relationship pairs in datasets to strengthen the iterative message propagation between relationships and objects while ignoring the intrinsic properties of the data.

Scene Graph Prediction in Point Clouds. With the recently proposed 3DSSG datasets containing 3D scene graph annotations [37], the community started to explore semantic relationship prediction in 3D real world data. SGPN [37, 38] is the first work to build a 3D scene graph using both objects and their interrelations as graph nodes. It then performs message propagation using graph convolutional networks. Kimera [30] proposed a 3D dynamic scene graph that captures metric and semantic aspects of a dynamic environment, where nodes represent spatial concepts at different levels of abstraction, and edges represent spatial-temporal relations among the nodes. EdgeGCN [46] exploits multi-dimensional edge features for explicit relationship modeling and explores two associated twinning interaction mechanisms for the independent evolution of scene graph representations. Wu et al. [41] proposed a method to incrementally build semantic scene graphs from a 3D environment given a sequence of

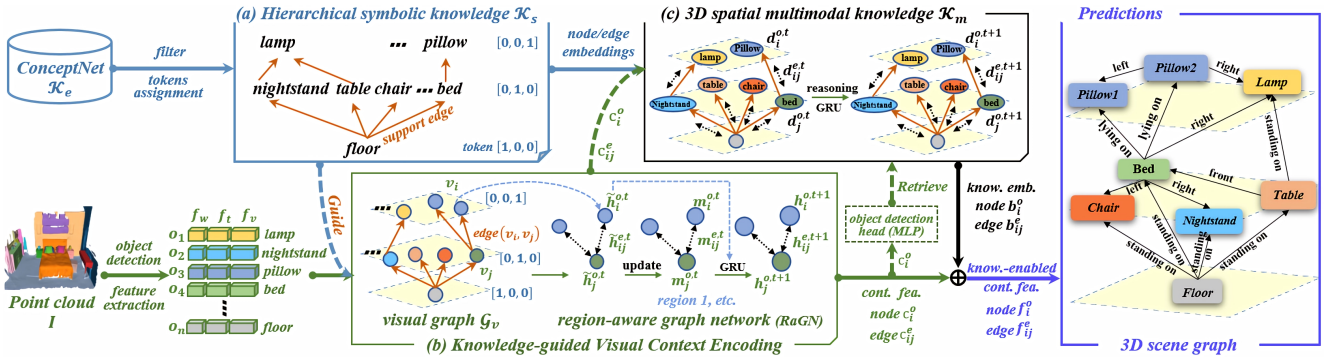

Figure 2. Method pipeline. (a) A hierarchical symbolic knowledge is firstly reconstructed to exploit external knowledge as the baseline and admit the hierarchical structure cues of 3D scene. (b) We then build a hierarchical visual graph and learn the contextualized features by the region-aware graph network. (c) Finally, a 3D spatial multimodal knowledge is accumulated to strengthen relationship predictions.

RGB-D frames. KISG [47] uses the ground truth relationship triplets in the dataset to extract the prior knowledge and then fuses it in the scene graph prediction stage. One limitation of KISG [47] is that its relevant prior knowledge depends on the text-only dataset label while ignoring hierarchical and indispensable structures in the 3D scene for visual understanding. Our method differentiates itself from these related studies by exploring the 3D implicit structure pattern and introducing 3D spatial multimodal knowledge, which enables our model to predict relationships more accurately.

3. Methodology

Problem Formulation: The goal of 3D scene graph generation is to describe a given 3D point cloud scene $\mathcal{I}$ with a semantic scene graph $\mathcal{G} = \{\mathcal{V}, \mathcal{R}\}$, where $\mathcal{V}$ and $\mathcal{R}$ represent instance object nodes and their inner relationship edges respectively. $\mathcal{G}$ forms a structured representation of the semantic content of the 3D scene. The nodes $\mathcal{V}$ consist of a set of objects $O = \{o_1, o_2, \cdots, o_n\}$ with object o_i assigned to a certain class label C, a corresponding set of bounding boxes $B = \{b_1, b_2, \cdots, b_n\}$ with $b_i \in \mathbb{R}^6$, and a set of relationship edges $\mathcal{R} = \{r_1, r_2, \cdots, r_n\}$ with each r_i represents a predicate between a pair of objects. Our proposed model can be decomposed as:

$$P(\mathcal{G}|\mathcal{I}) = P(\mathcal{K}_s|\mathcal{I})P(\mathcal{G}_v|\mathcal{K}_s, \mathcal{I})P(\mathcal{R}, \mathcal{K}_m|\mathcal{G}_v, \mathcal{K}_s, \mathcal{I}) \tag{1}$$

In this equation, the component $P(\mathcal{K}_s|\mathcal{I})$ collects all the symbolic entities from the datasets, filters the extra knowledge bases, and combines the hierarchical structure patterns of 3D scenes to construct the hierarchical symbolic knowledge $\mathcal{K}_s$. The component $P(\mathcal{G}_v|\mathcal{K}_s, \mathcal{I})$ builds visual graphs for scenes under the guidance of knowledge $\mathcal{K}_s$, where contextual features for each node are extracted. Conditioned on the knowledge $\mathcal{K}_s$ and visual graph $\mathcal{G}_v$, the component $P(\mathcal{R}, \mathcal{K}_m|\mathcal{G}_v, \mathcal{K}_s, \mathcal{I})$ accumulates the 3D spatial multimodal knowledge by correlating the knowledge $\mathcal{K}_s$ with visual content and predicts relationships simultaneously. Fig. 2 illustrates the overall pipeline of the proposed model.

3.1. Hierarchical Symbolic Knowledge Initialization

Unlike KISG [47], we do not use a closed set or ground truth relationship triplets from labels to learn prior knowledge. Hence, we must make an additional choice of what knowledge sources to use and how to clean them. Prior knowledge of object classes can be reliable predictors of the likelihoods of physical support relationships. For instance, it is unlikely that a *cup* is supported by a *wall* while *tables* are almost always supported by the *floor*. Therefore, given a set of objects, we can classify each object based on whether it is directly supported by the *floor*. The result is a three-layer hierarchical structure about objects in the 3D scene. In particular, the first layer only contains the *floor* since it does not have any support. The second layer contains objects directly supported by the *floor*, e.g. *bed*, *table*, and *sofa*. The third layer contains the remaining objects usually supported by objects in the second layer, e.g. *pillow*, *cup*, and *cushion*.

To exploit the regular structure patterns in 3D spaces and construct the scene graph hierarchically, we construct a hierarchical symbolic knowledge graph to guide the 3D spatial knowledge reasoning. Knowledge sources, such as ConceptNet [35] and DBPedia [2], are a valuable tool containing commonsense knowledge about the real world. In this work, we use ConceptNet as our external knowledge base which gives us more spatial relationships and common pairwise objects. While ConceptNet contains very useful information, it also includes some knowledge that is irrelevant to our model. To mitigate this issue, we limit the ConceptNet to commmon object categories in 3D point cloud scenes. We collect object categories from two widely-used 3D point cloud datasets, SUNRGBD [34] and Scannet [7], and then include edges that only include these objects. After filtering, we have a total of about 5,000 edges and 760 nodes.

We denote the external knowledge graph as $\mathcal{K}_e = \{\mathcal{V}_e, \mathcal{E}_e\}$ where $\mathcal{V}_e$ and $\mathcal{E}_e$ represent nodes and edges respectively. To merge the hierarchical structures in 3D spaces into the external knowledge graph and construct the hierarchical

symbolic knowledge graph $\mathcal{K}_s$, we first use a pre-trained multi-layer perceptron (MLP) to classify the hierarchical tokens for each node in the external knowledge graph to distinguish the discrepancy among different layers of nodes. The hierarchical token of each node denotes its corresponding layer in the hierarchical structure. Each node is then initialized as the concatenation of its trainable hierarchical token and the word2vec (GloVe [26]) representation of the object category. Since the hierarchical structure of 3D spaces is built based on the physical support relationships between objects, we add additional edges representing support relationships between nodes to the external knowledge graph $\mathcal{K}_e$. Specifically, we define a new edge type: given two nodes s_i and s_j, we connect s_i to s_j using a support edge to represent the physical support relationship between s_i and s_j. By definition, each node in the hierarchical structure is supported by the node in neighboring layers. Therefore, we add a support edge between two correlated nodes in neighboring layers. Each edge is initialized as the trainable GloVe representation of its edge type. Finally, we formulate the updated external knowledge graph as hierarchical symbolic knowledge graph $\mathcal{K}_s$. Additional details can be found in supplementary.

3.2. Knowledge-guided Visual Context Encoding

As shown in Fig. 2, taking a scene point cloud with object instance annotations as input, we build a hierarchical visual graph $\mathcal{G}_v = \{\mathcal{V}_v, \mathcal{E}_v\}$ where $\mathcal{V}_v$ and $\mathcal{E}_v$ denotes object instances and edges of object pairs respectively, under the guidance of the hierarchical symbolic knowledge graph $\mathcal{K}_s$. Then, a region-aware graph network is employed to propagate node messages through the visual graph $\mathcal{G}_v$ to learn the contextualized feature representation.

Visual graph construction. We use Point Cloud Transformer [13] to extract spatial-aware visual features f_v for each object instance. To encode the spatial features f_t of each bounding box, we use an MLP to lift the parameters of each bounding box (i.e., center and size) to feature space. We assign the semantic features f_w for each object using an embedding table initialized by GloVe [26]. Each node in the visual graph is initialized as the concatenation of features f_v, f_t and f_w. To capture the implicit structure of the point cloud scene, we route each node in the visual graph $\mathcal{G}_v$ into its corresponding layer according to the hierarchical tokens in hierarchical symbolic knowledge graph $\mathcal{K}_s$. Then, we complete the edge set $\mathcal{E}_v$ of visual graph $\mathcal{G}_v$ by extracting potential physical relationships between nodes in the adjacent layers. Specifically, we add an edge representing physical support relationship between node pair in the visual graph $\mathcal{G}_v$ if a support edge also exists between the corresponding nodes in the hierarchical symbolic knowledge graph $\mathcal{K}_s$. Similar to [46], we model the spatial interactions between node pairs and encode the initial edge embedding for node pairs using an MLP.

Contextualized features encoding. Objects sharing the same physical support are correlated since they have similar functional role in the environment and are generally in close proximity to each other. For instance, both *pillow* and *clothes* are usually supported by a *bed*. Therefore, we propose a region-aware graph network to jointly highlight the interrelated regions of each node in the visual graph $\mathcal{G}_v$ and encode the hierarchical contexts of the input scene.

Given the initial representations of nodes and edges in the visual graph $\mathcal{G}_v$, the region-aware graph network iteratively updates the hidden state $\mathbf{h}_i^{o,t}$ of each node v_i and $\mathbf{h}_{ij}^{e,t}$ of each edge (v_i, v_j) at each time step t via message passing. Since the contextual regions around each node in the visual graph $\mathcal{G}_v$ can be defined as other nodes sharing the same physical support with it, each node first gathers information from nodes within the same contextual region to enrich its current hidden state before propagating messages along the edges in the visual graph $\mathcal{G}_v$. Specifically, the enriched hidden state $\tilde{\mathbf{h}}_i^{o,t}$ of each node is:

$$\tilde{\mathbf{h}}_i^{o,t} = \mathbf{h}_i^{o,t} + \sum_{j \in N_r(i)} \psi(\mathbf{h}_j^{o,t}) \tag{2}$$

$N_r(i)$ contains nodes that share the same level support with node v_i and ψ is a feed forward network for non-linear transformation. For edge (v_i, v_j), its enriched hidden state $\tilde{\mathbf{h}}_{ij}^{e,t}$ is computed by:

$$\tilde{\mathbf{h}}_{ij}^{e,t} = \mathbf{h}_{ij}^{e,t} + \sum_{k \in N_r(i)} \psi(\mathbf{h}_k^{o,t}) + \sum_{s \in N_r(j)} \psi(\mathbf{h}_s^{o,t}) \tag{3}$$

After the feature representation enhancements, the message passing of nodes and edges can be formulated as:

$$\mathbf{h}_i^{o,t+1} = GRU(\tilde{\mathbf{h}}_i^{o,t}, \mathbf{m}_i^{o,t}) \tag{4}$$

$$\mathbf{h}_{ij}^{e,t+1} = GRU(\tilde{\mathbf{h}}_{ij}^{e,t}, \mathbf{m}_{ij}^{e,t}) \tag{5}$$

where $\mathbf{m}_i^{o,t}$ and $\mathbf{m}_{ij}^{e,t}$ are the incoming messages for updating each node and edge. The calculation of the message for each node is:

$$\mathbf{m}_i^{o,t} = \sum_{j \in N_v(i)} (\varphi_n(\tilde{\mathbf{h}}_j^{o,t}) + \varphi_e(\tilde{\mathbf{h}}_{ij}^{e,t})) \tag{6}$$

where $\mathcal{N}_v(i)$ denotes the neighbor nodes of v_i in the visual graph $\mathcal{G}_v$, φ_n and φ_e are two non-linear transformation for associated nodes and edges. For each edge, we transform the hidden state of subject and object node by two MLPs before fusing them to obtain the message:

$$\mathbf{m}_{ij}^{e,t} = \varphi_s(\tilde{\mathbf{h}}_i^{o,t}) + \varphi_o(\tilde{\mathbf{h}}_j^{o,t}) \tag{7}$$

We take the final hidden states of nodes and edges as the contextual feature $\mathbf{c}_i^o$ for each node $v_i \in \mathcal{V}_v$ and $\mathbf{c}_{ij}^e$ for each edge $(v_i, v_j) \in \mathcal{E}_v$.

3.3. Spatial Multimodal Knowledge Accumulation

Though our hierarchical symbolic knowledge graph $\mathcal{K}_s$ can provide high-quality knowledge about the hierarchical structures of point cloud scene, this information is largely limited to symbolic knowledge that can only be explicitly expressed by text-relevant labels for relationship triplets. Therefore, we propose a novel schema to accumulate 3D spatial multimodal knowledge $\mathcal{K}_m$ progressively from the visual context via a graph reasoning network. We then incorporate the learned multimodal knowledge $\mathcal{K}_m$ and the contextual features to predict the possible relationships.

Reasoning on knowledge graph. Since the contextual features encode the implicit hierarchical structure patterns in 3D spaces, we design a graph reasoning network which utilizes the visual contextual features and textual facts from the hierarchical symbolic knowledge graph $\mathcal{K}_s$ to accumulate 3D spatial multimodal knowledge $\mathcal{K}_m$ by aligning the entities in the symbolic knowledge graph with related visual contextual features.

The graph reasoning network generates context for 3D spatial multimodal knowledge $\mathcal{K}_m$, which is in the form of embeddings that capture the regular structure patterns in 3D scenes for each node and edge in the hierarchical symbolic knowledge graph $\mathcal{K}_s$. Given the contextual features of nodes and edges in visual graph $\mathcal{G}_v$, each node and edge in the graph reasoning network receives three inputs: (1) the trainable node or edge embedding in the hierarchical symbolic knowledge graph $\mathcal{K}_s$, (2) a $0/1$ indicator of whether this node or edge appears in the visual graph $\mathcal{G}_v$, (3) the contextual feature $\mathbf{c}_i^o$ and $\mathbf{c}_{ij}^e$ in the visual graph $\mathcal{G}_v$ corresponding to this node or edge, missing nodes and edges are padded with zero vectors. The graph reasoning network uses message passing to perform reasoning on hierarchical symbolic knowledge graph $\mathcal{K}_s$. Specifically, at each time step t, to calculate the hidden states $\mathbf{d}_i^{o,t}$ for all nodes $s_i \in \mathcal{V}_s$ and $\mathbf{d}_{ij}^{e,t}$ for all edges $(s_i, s_j) \in \mathcal{E}_s$, each node and edge first gather messages from their neighbors through the graph structure then update their hidden states:

$$\mathbf{d}_i^{o,t+1} = GRU(\mathbf{d}_i^{o,t}, \mathbf{m}_i^{o,t}), \tag{8}$$

$$\mathbf{d}_{ij}^{e,t+1} = GRU(\mathbf{d}_{ij}^{e,t}, \mathbf{m}_{ij}^{e,t}), \tag{9}$$

where $\mathbf{m}_i^{o,t}$ and $\mathbf{m}_{ij}^{e,t}$ are the incoming messages for nodes and edges. The incoming message for each node is

$$\mathbf{m}_i^{o,t} = \sum_{j \in N_k(i)} (\varphi_n(\mathbf{d}_j^{o,t}) + \varphi_e(\mathbf{d}_{ij}^{e,t})), \tag{10}$$

where $N_k(i)$ denotes the neighbor nodes of node s_i in the knowledge graph $\mathcal{K}_s$. Similar to Eq. (7), the incoming message for each edge is

$$\mathbf{m}_{ij}^{e,t} = \varphi_s(\mathbf{d}_i^{o,t}) + \varphi_o(\mathbf{d}_j^{o,t}). \tag{11}$$

We take the sum of the stacked hidden states as the 3D spatial multimodal knowledge embedding $\mathbf{b}_i^o$ for all nodes and $\mathbf{b}_{ij}^e$ for all edges in the symbolic knowledge graph $\mathcal{K}_s$.

Knowledge-enabled Scene Graph Prediction. To incorporate the 3D spatial multimodal knowledge $\mathcal{K}_m$ into scene graph inference, we propose fusing the multimodal knowledge embedding with the contextual features in the visual graph to facilitate 3D scene graph prediction. Towards this goal, we utilize an MLP as object detection head to predict confident initial class guesses given the contextual node features. We then select the three most confident multimodal knowledge embeddings for each node. For edges in the visual graph, we select the three most confident object categories for the subject and object node based on the initial guesses. We then retrieve the multimodal knowledge embedding using the predicted subject and object categories. Since the multimodal knowledge embedding and the contextual features are in different feature spaces, we transform them by two MLPs φ_b and φ_c respectively before fusing them. For each node in the visual graph, we fuse the retrieved multimodal knowledge embedding $\{\mathbf{b}^k\}_{k=1,2,3}$ and the contextual node feature $\mathbf{c}_i^o$ to obtain the knowledge-enabled contextual feature $\mathbf{f}_i^o$:

$$\mathbf{f}_i^o = \phi(\varphi_c(\mathbf{c}_i^o) + \varphi_b(\sum_{k=1}^{3} \mathbf{b}^k)). \tag{12}$$

For each edge in the visual graph, the multimodal knowledge embedding is fused with its contextual feature in the same way as the node.

Equipped with the 3D spatial multimodal knowledge-enabled contextual features $\mathbf{f}_i^o$ for nodes and $\mathbf{f}_{ij}^e$ for edges in the visual graph, we generate the scene graph by decoding the contextual features using a standard graph convolution network (GCN) [18]. We assume that each object pair can have a relationship (including none) and fully connect them as a graph where relationships are represented as edges. Each node is initialized by its contextual node feature $\mathbf{f}_i^o$, and each edge is initialized either by the contextual edge feature $\mathbf{f}_{ij}^e$ or the contextual features of its subject and object nodes if the edge is not presented in the visual graph. The last part of the GCN consists of two detection heads for object and relationship classification. The object detection head takes the decoded node features as input to predict the object classification possibilities. The relationship prediction head first fuses the decoded subject and object node features with the decoded edge features, then predicts a discrete distribution over all possible relationship classes.

Loss Function. We adopt the standard cross entropy loss for object and relationship classification in our model. Since the contextual node feature $\mathbf{c}_i^o$ is used to predict the initial class guesses, we use a cross entropy loss $\mathcal{L}_{init}^o$ for the initial detection. For the final prediction, we use two cross entropy

Methods	PredCls		SGCls		SGDet	
	R@50/100	mR@50/100	R@50/100	mR@50/100	R@50/100	mR@50/100
3D+IMP [42]	48.15 / 48.72	21.56 / 21.85	17.41 / 17.89	9.06 / 9.23	24.54 / 24.57	21.71 / 21.72
3D+MOTIFS [45]	52.43 / 53.37	24.35 / 24.52	18.34 / 18.57	9.74 / 9.86	26.58 / 26.59	24.12 / 24.17
3D+VCTree [36]	53.12 / 54.38	24.75 / 24.91	19.93 / 20.24	10.34 / 10.55	27.58 / 27.62	24.92 / 24.94
3D+KERN [6]	54.74 / 56.53	25.21 / 25.83	21.41 / 21.78	11.02 / 11.36	27.75 / 27.78	24.03 / 24.05
3D+Schemata [32]	58.13 / 59.11	42.11 / 42.83	28.72 / 28.97	26.72 / 27.05	28.12 / 28.13	25.29 / 25.30
3D+HetH [40]	58.24 / 58.75	42.53 / 42.74	28.83 / 29.05	26.68 / 26.85	28.17 / 28.18	25.31 / 25.32
Ours	**68.32 / 69.49**	**66.54 / 66.92**	**31.50 / 31.64**	**30.29 / 30.56**	**29.41 / 29.44**	**25.35 / 25.36**

Table 1. Comparison with state-of-the-art 2D scene graph prediction methods re-implemented to work on 3DSSG dataset.

Methods	PredCls		SGCls		SGDet	
	R@50/100	mR@50/100	R@50/100	mR@50/100	R@50/100	mR@50/100
SGPN [37]	57.71 / 58.05	38.12 / 38.67	28.39 / 28.74	22.23 / 22.57	- / -	- / -
EdgeGCN [46]	58.42 / 59.11	38.84 / 39.35	28.58 / 28.93	22.67 / 23.33	- / -	- / -
KISG [47]	64.47 / 64.93	63.19 / 63.52	29.46 / 29.65	28.20 / 28.64	- / -	- / -
Ours	**68.32 / 69.49**	**66.54 / 66.92**	**31.50 / 31.64**	**30.29 / 30.56**	**29.41 / 29.44**	**25.35 / 25.36**

Table 2. Comparison with 3D scene graph prediction methods on the 3DSSG dataset.

losses $\mathcal{L}^o_{final}$ and $\mathcal{L}^r_{final}$ for the object and relationship classification:

$$\mathcal{L}_{final} = w_o \mathcal{L}^o_{final} + w_r \mathcal{L}^r_{final} \tag{13}$$

where w_o and w_r are the weights for object and relation loss. In our experiment, we set w_o to 0.75 and w_r to 1. Our final loss function can be formulated as:

$$\mathcal{L} = \mathcal{L}^o_{init} + \mathcal{L}_{final} \tag{14}$$

4. Experiments

4.1. Experimental Configuration

We evaluate our model on 3DSSG dataset [37]. Following [47], we select 160 object categories and 27 relationship classes for detection. We compare our model with others in three standard tasks proposed in [42]. (1) Predicate Classification (PredCls): Given the ground truth 3D bounding boxes and their corresponding semantic labels, our model classifies the relationship between each object pair. (2) Scene Graph Classification (SGCls): Given the ground truth 3D bounding boxes, our model predicts the relationships as well as the object categories jointly. (3) Scene Graph Generation (SGDet): Given the raw point cloud, our model detects 3D objects, their semantic information, as well as their relationships in an end-to-end manner. Following existing 2D and 3D scene graph generation works, we adopt the constrained evaluation metric recall@K (R@K) and mean recall@K (mR@K).

Our model is implemented in PyTorch, and trained using one NVIDIA GTX TITAN X GPU for 40 epochs with the ADAM optimizer. We use an initial learning rate of 0.0001, weight decay of 0.5, and mini-batch of 4. After 15, 25, and 40 epochs, we multiply the learning rate by 0.1. We adopt VoteNet [27] as the 3D object detection backbone to generate an initial set of 256 object candidates in the SGDet task. The Point Cloud Transformer is pre-trained on the 3DSSG dataset using the same settings in [13].

4.2. Comparison to State-of-the-Art

We first compare our model with the following state-of-art 2D image scene graph generation models, modified to fit the 3DSSG dataset: IMP [42], MOTIFS [45] and VCTree [36] which creatively devise various message passing methods for improving graph representations. KERN [6], Schemata [32], and HetH [40] incorporate statistical priors and learning-based commonsense knowledge into the scene graph prediction. Therefore, we include these models to illustrate the superiority of the 3D spatial multimodal knowledge about the implicit hierarchical structure correlations between object pairs in the 3D scene.

Our results in Tab. 1 lead to a few key observations: (1) Our model consistently outperforms all the existing approaches on all metrics and achieve 3.57% boost on mR@50 in SGCls task and 10.08% boost on R@50 in PredCls task. This indicates that leveraging regular patterns of 3D physical spaces is beneficial for scene graph prediction. (2) Our model outperforms traditional message passing model IMP and MOTIFS. Furthermore, our method achieves considerable improvement when compared to VCTree. (3) Compared to Schemata, our model achieves an improvement of 2.78% and 10.19% on R@50 in SGCls and PredCls, suggesting that our multimodal knowledge embedding is a better approach compared to the class-level prototypical representations learned from perceptual outputs in Schemata. (4) Compared with KERN and HetH, our proposed hierarchical structure of 3D spaces is superior to the graph structure they adopted to represent the input as our model outperforms them with a significant margin. (5) The performance has been saturated in the SGDet task. This is mainly because object detection performance on this dataset is a bottleneck that limits the performance.

We also compare the performance of our model with the state-of-the-art 3D point cloud-based scene graph predic-

Methods	R@50/100	mR@50/100
Knowledge $\mathcal{K}_s$		
w/o Hierarchical Tokens	30.47 / 30.67	28.94 / 29.19
w/o Support Edge	30.55 / 30.74	29.17 / 29.47
w/o Both	28.41 / 28.47	27.13 / 27.52
Visual Context Encoding		
$\mathcal{G}_v$ replaced w/ $\mathcal{G}_{fc}$	28.17 / 28.32	26.28 / 26.29
w/o RaGN	26.43 / 26.57	24.23 / 24.36
RaGN replaced w/ GCN	31.03 / 31.21	29.67 / 29.88
Knowledge $\mathcal{K}_m$		
w/o $\mathbf{b}_i^o$ and $\mathbf{b}_{ij}^e$	26.27 / 26.35	22.93 / 23.18
w/o $\mathbf{c}_i^o$ and $\mathbf{c}_{ij}^e$ as input	28.14 / 28.31	25.05 / 25.31
Ours	**31.50 / 31.64**	**30.29 / 30.56**

Table 3. Quantitative results of different module configurations on the SGCls task.

tion models to demonstrate the effectiveness of 3D spatial multimodal knowledge. We include several existing works such as SGPN [37], EdgeGCN [46] and KISG [47] since they all report competitive results. SGPN and EdgeGCN exploit multi-dimensional edge features for explicit relationship modeling whereas KISG learns a group of class-dependent prototypical representations for each semantic class. As shown in Tab. 2, our model dominantly surpasses all methods. Benefiting from the hierarchical structure of 3D spaces, our model is able to reason complex relationship hierarchically and systematically. Compared to SGPN and EdgeGCN, our model improves the R@50 by 2.92% and 9.90% in SGCls and PredCls tasks. We can also see that our method outperforms KISG by 2.04% on R@50 in SGCls. KISG captures class-related priors in the scene from text-only ground truth labels. Such knowledge cannot efficiently represent diverse relationships and complex 3D environments. In contrast, our model extracts indispensable 3D spatial multimodal knowledge which benefits the scene graph prediction.

4.3. Ablation Study

We only report the performance results in the Recall and mean Recall metrics on the SGCls task for ablation studies. The results are shown in Tab. 3.

Hierarchical symbolic knowledge. We first look at the hierarchical symbolic knowledge graph $\mathcal{K}_s$ to investigate its effectiveness. Specifically, we find that using ConceptNet without classifying the hierarchical tokens or adding support edges leads to sub-optimal performance. Furthermore, using ConceptNet without any augmentation drops the performance significantly, indicating that both the hierarchical tokens and support edges are crucial elements of the hierarchical structures in 3D scene.

Knowledge-guided visual context encoding. Next, we analyse the knowledge-guided visual context encoding module. We can see that replacing the hierarchical visual graph $\mathcal{G}_v$ with a fully-connected graph $\mathcal{G}_{fc}$ decreases the performance by a margin of 3.33% on R@50, indicating that the hierarchical structure is superior to a plain fully-connected graph in terms of modeling context. Furthermore, removing the subsequent region-aware graph network (RaGN) and directly fusing the multimodal knowledge embedding with the initial representation of each node and edge in the visual graph negatively impacts the performance on all metrics. Replacing the region-aware graph network with a standard graph convolution network also hurts the performance.

Variants	PredCls R@50	PredCls mR@50	SGCls R@50	SGCls mR@50
$\mathcal{G}_r$	62.74	58.25	28.17	27.28
$\mathcal{G}_t$	**68.41**	**66.59**	**31.59**	**30.35**
$\mathcal{G}_v$(original)	68.32	66.54	31.50	30.29

Table 4. Comparison of different variants of the visual graph.

Methods	Head	Body	Tail
SGPN [37]	39.42	23.64	13.03
EdgeGCN [46]	39.51	23.85	13.15
KISG [37]	40.36	24.56	13.61
Ours	**44.23**	**26.27**	**14.73**

Table 5. The R@50 metric of biased relationship prediction on the SGCls task.

3D spatial multimodal knowledge accumulation. Lastly, we examine the accumulated multimodal knowledge $\mathcal{K}_m$ to learn about how $\mathcal{K}_m$ and rest of the model interact. We first see how much of the improvement comes from the 3D spatial multimodal knowledge $\mathcal{K}_m$. As shown in Tab. 3, the multimodal knowledge embedding significantly improves the R@50 and mR@50 by 5.23% and 7.36% respectively. In addition, dropping the contextual feature input $\mathbf{c}_i^o$ for nodes and $\mathbf{c}_{ij}^e$ for edges in the graph reasoning network decreases the performance by a margin of 3.36% and 5.24% on R@50 and mR@50 in SGCls. This drop in performance indicates that the contextual feature plays a pivotal role in bridging the heterogeneous gap between the symbolic knowledge and visual information.

4.4. Further Analysis

Analysis on the hierarchical structure of 3D spaces. To validate the potential of the hierarchical visual graph $\mathcal{G}_v$ in capturing the inherent hierarchical structure of a 3D scene, we design two visual graph variants and compare them to the hierarchical visual $\mathcal{G}_v$: (1) Instead of using the hierarchical symbolic knowledge graph $\mathcal{K}_s$, we build a ground truth graph $\mathcal{G}_t$ based on the ground truth labels for support relations. In particular, each edge in $\mathcal{G}_t$ represents the ground truth support relationship of the input scene. (2) We also design a randomly connected graph $\mathcal{G}_r$, where we keep all of the nodes the same but randomize the edges that connect them. As shown in Tab. 4, both $\mathcal{G}_v$ and $\mathcal{G}_t$ outperform $\mathcal{G}_r$

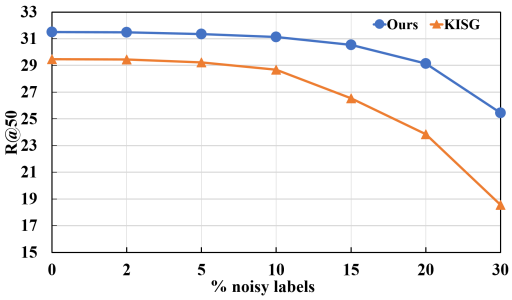

Figure 3. Comparison of our model and KISG on the SGCls task when trained with noisy labels.

with a significant margin on all metrics. More importantly, we observe that $\mathcal{G}_t$ and $\mathcal{G}_v$ perform mostly similar while $\mathcal{G}_t$ slightly outperforms $\mathcal{G}_v$. The results confirm that the hierarchical visual graph $\mathcal{G}_v$ is one of the more optimal ways of extracting the hierarchical structure patterns of 3D spaces.

Robustness of 3D spatial multimodal knowledge. Additionally, we investigate the robustness of the 3D spatial multimodal knowledge $\mathcal{K}_m$ by training our model with noisy labels. Specifically, we add different proportions of noises into the 3DSSG training set by replacing part of ground truth relationships with the randomly selected wrong relationships for input scenes. The performance of our model and KISG [47] on the SGCls task is reported in Fig. 3. We can see that, the performance of KISG decreases drastically while ours decreases slowly with increasing noise rate. Under the 30% noise rate condition, our model improves the R@50 metric by about 6.89% over KISG, which indicates that our model achieves improved robustness over KISG. The main reason is that KISG captures relevant prior knowledge from text-only ground truth labels and noises contained in the labels are easily included in their knowledge base and affects the prediction of relationships. Different with KISG, our model leverages the inherently hierarchical structures of 3D scenes and accumulates multimodal knowledge which is both label free and reliable.

Long-tail analysis. We also investigate how our model performs on the long-tail part of the dataset. To do this, we order all the relationships based on the frequency of each relationship category occurring in triplets. We select the 5 most common relationship categories as the head, the 5 least common relationship categories as the tail, and the rest of the categories as the body. Tab. 5 reports the R@50 metric on each long-tail category groups of our model. Moreover, our model achieves best performance when evaluating the R@50 metric on the tail relationship categories, which shows that our model has the ability to mitigate the effect of sample imbalance. The main reason is that the hierarchical structures can be extracted accurately which influence other relationships in the prediction process.

4.5. Qualitative Results

We visualize intermediate results in Fig. 4(a-c). We can see that both the hierarchical visual graph $\mathcal{G}_v$ and 3D scene graph $\mathcal{G}$ are well constructed. However, our model incorrectly classifies the relationship between *Window1* and *Floor*. This is mainly because our model fails to extract discriminative features for *Window1* as there are few points within its bounding box. The token of *Window1* is classified incorrectly in the second layer while it should be in the third layer. We provide more visualization samples in the supplementary.

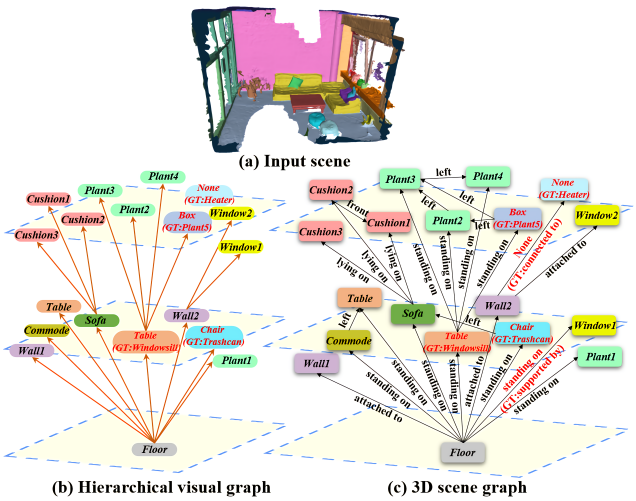

Figure 4. Visualizations of our predicted scene graph on 3DSSG dataset. Red indicates the misclassified objects or relationships.

5. Conclusion

We proposed a method for 3D scene graph prediction from raw point clouds. Our method explores the regular patterns of 3D physical spaces into the deep network to facilitate 3D scene graph prediction. Hierarchical symbolic knowledge is first reconstructed via exploiting external knowledge as the baseline to admit the hierarchical structure cues of a 3D scene. A knowledge-guided visual context encoding module then builds a hierarchical visual graph and learns the contextualized features by a region-aware graph network. Finally, a 3D spatial multimodal knowledge accumulation module is proposed to regularize the semantic space of relationship prediction. Extensive experiments on the 3DSSG dataset show that our method outperforms existing state-of-the-art and can mitigate the effect of data imbalance and label noises. In the future, we plan to exploit the attributes of 3D objects to build richer knowledge graphs to improve the prediction performances of attribute-focused relationships, such as *same symmetric as* and *same texture as*.

6. Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grant 62003253, Grant 61973106, Grant U2013203, Grant U21A20482 and Grant U20A20185. Professor Ajmal Mian is the recipient of an Australian Research Council Future Fellowship Award (project number FT210100268) funded by the Australian Government.

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

[2] Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. Dbpedia: A nucleus for a web of open data. In *The semantic web*, pages 722–735. 2007. 2, 3

[3] Sumithra Bhakthavatsalam, Kyle Richardson, Niket Tandon, and Peter Clark. Do dogs have whiskers? a new knowledge base of haspart relations. *arXiv preprint arXiv:2006.07510*, 2020. 2

[4] Xiaojun Chang, Pengzhen Ren, Pengfei Xu, Zhihui Li, Xiaojiang Chen, and Alexander G Hauptmann. A comprehensive survey of scene graphs: Generation and application. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021. 2

[5] Long Chen, Hanwang Zhang, Jun Xiao, Xiangnan He, Shiliang Pu, and Shih-Fu Chang. Counterfactual critic multi-agent training for scene graph generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4613–4623, 2019. 2

[6] 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, 6

[7] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5828–5839, 2017. 3

[8] Yang Ding, Jing Yu, Bang Liu, Yue Hu, Mingxin Cui, and Qi Wu. Mukea: Multimodal knowledge extraction and accumulation for knowledge-based visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5089–5098, 2022. 2

[9] Zijin Du, Hailiang Ye, and Feilong Cao. A novel local-global graph convolutional method for point cloud semantic segmentation. *IEEE Transactions on Neural Networks and Learning Systems*, 2022. 1

[10] Mingtao Feng, Syed Zulqarnain Gilani, Yaonan Wang, Liang Zhang, and Ajmal Mian. Relation graph network for 3d object detection in point clouds. *IEEE Transactions on Image Processing*, 30:92–107, 2020. 1

[11] Mingtao Feng, Liang Zhang, Xuefei Lin, Syed Zulqarnain Gilani, and Ajmal Mian. Point attention network for semantic segmentation of 3d point clouds. *Pattern Recognition*, 107:107446, 2020. 1

[12] Chen Gao, Jinyu Chen, Si Liu, Luting Wang, Qiong Zhang, and Qi Wu. Room-and-object aware knowledge reasoning for remote embodied referring expression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3064–3073, 2021. 2

[13] Meng-Hao Guo, Jun-Xiong Cai, Zheng-Ning Liu, Tai-Jiang Mu, Ralph R Martin, and Shi-Min Hu. Pct: Point cloud transformer. *Computational Visual Media*, 7(2):187–199, 2021. 4, 6

[14] Alon Hafri and Chaz Firestone. The perception of relations. *Trends in Cognitive Sciences*, 2021. 1

[15] Qingyong Hu, Bo Yang, Linhai Xie, Stefano Rosa, Yulan Guo, Zhihua Wang, Niki Trigoni, and Andrew Markham. Randla-net: Efficient semantic segmentation of large-scale point clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11108–11117, 2020. 1

[16] Maximilian Jaritz, Tuan-Hung Vu, Raoul De Charette, Émilie Wirbel, and Patrick Pérez. Cross-modal learning for domain adaptation in 3d semantic segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. 1

[17] 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. 2

[18] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016. 5

[19] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123(1):32–73, 2017. 2

[20] Manyi Li, Akshay Gadi Patil, Kai Xu, Siddhartha Chaudhuri, Owais Khan, Ariel Shamir, Changhe Tu, Baoquan Chen, Daniel Cohen-Or, and Hao Zhang. Grains: Generative recursive autoencoders for indoor scenes. *ACM Transactions on Graphics (TOG)*, 38(2):1–16, 2019. 1

[21] Mengtian Li, Yuan Xie, Yunhang Shen, Bo Ke, Ruizhi Qiao, Bo Ren, Shaohui Lin, and Lizhuang Ma. Hybridcr: Weakly-supervised 3d point cloud semantic segmentation via hybrid contrastive regularization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14930–14939, 2022. 1

[22] Ze Liu, Zheng Zhang, Yue Cao, Han Hu, and Xin Tong. Group-free 3d object detection via transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2949–2958, 2021. 1

[23] Kenneth Marino, Xinlei Chen, Devi Parikh, Abhinav Gupta, and Marcus Rohrbach. Krisp: Integrating implicit and symbolic knowledge for open-domain knowledge-based vqa. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14111–14121, 2021. 2

[24] George A Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995. 2

[25] Xuran Pan, Zhuofan Xia, Shiji Song, Li Erran Li, and Gao Huang. 3d object detection with pointformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7463–7472, 2021. 1

[26] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In

Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), pages 1532–1543, 2014. 4

[27] Charles R Qi, Or Litany, Kaiming He, and Leonidas J Guibas. Deep hough voting for 3d object detection in point clouds. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9277–9286, 2019. 1, 6

[28] 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. 1

[29] Charles R Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++ deep hierarchical feature learning on point sets in a metric space. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 5105–5114, 2017. 1

[30] Antoni Rosinol, Andrew Violette, Marcus Abate, Nathan Hughes, Yun Chang, Jingnan Shi, Arjun Gupta, and Luca Carlone. Kimera: From slam to spatial perception with 3d dynamic scene graphs. *The International Journal of Robotics Research*, 40(12-14):1510–1546, 2021. 1, 2

[31] Morteza Sarafyazd and Mehrdad Jazayeri. Hierarchical reasoning by neural circuits in the frontal cortex. *Science*, 364(6441), 2019. 1

[32] 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*, volume 35, pages 5025–5033, 2021. 6

[33] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *European conference on computer vision*, pages 746–760. Springer, 2012. 1

[34] Shuran Song, Samuel P Lichtenberg, and Jianxiong Xiao. Sun rgb-d: A rgb-d scene understanding benchmark suite. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 567–576, 2015. 3

[35] Robyn Speer, Joshua Chin, and Catherine Havasi. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Thirty-first AAAI conference on artificial intelligence*, 2017. 2, 3

[36] Kaihua Tang, Hanwang Zhang, Baoyuan Wu, Wenhan Luo, and Wei Liu. Learning to compose dynamic tree structures for visual contexts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6619–6628, 2019. 2, 6

[37] 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, 6, 7

[38] Johanna Wald, Nassir Navab, and Federico Tombari. Learning 3d semantic scene graphs with instance embeddings. *International Journal of Computer Vision*, pages 1–22, 2022. 1, 2

[39] Kai Wang, Yu-An Lin, Ben Weissmann, Manolis Savva, Angel X Chang, and Daniel Ritchie. Planit: Planning and instantiating indoor scenes with relation graph and spatial prior networks. *ACM Transactions on Graphics (TOG)*, 38(4):1–15, 2019. 1

[40] Wenbin Wang, Ruiping Wang, Shiguang Shan, and Xilin Chen. Sketching image gist: Human-mimetic hierarchical scene graph generation. In *European Conference on Computer Vision*, pages 222–239. Springer, 2020. 6

[41] Shun-Cheng Wu, Johanna Wald, Keisuke Tateno, Nassir Navab, and Federico Tombari. Scenegraphfusion: 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

[42] 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, 6

[43] Qiangeng Xu, Yiqi Zhong, and Ulrich Neumann. Behind the curtain: Learning occluded shapes for 3d object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 2893–2901, 2022. 1

[44] Jianwei Yang, Jiasen Lu, Stefan Lee, Dhruv Batra, and Devi Parikh. Graph r-cnn for scene graph generation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 670–685, 2018. 2

[45] 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, 6

[46] 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. 1, 2, 4, 6, 7

[47] Shoulong Zhang, Aimin Hao, Hong Qin, et al. Knowledge-inspired 3d scene graph prediction in point cloud. *Advances in Neural Information Processing Systems*, 34, 2021. 2, 3, 6, 7, 8

[48] Yifeng Zhang, Ming Jiang, and Qi Zhao. Explicit knowledge incorporation for visual reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1356–1365, 2021. 2

[49] Na Zhao, Tat-Seng Chua, and Gim Hee Lee. Few-shot 3d point cloud semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8873–8882, 2021. 1