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

Mingtao Feng\textsuperscript{*1} Haoran Hou\textsuperscript{*1} Liang Zhang\textsuperscript{1†} Zijie Wu\textsuperscript{†2} Yulan Guo\textsuperscript{3} Ajmal Mian\textsuperscript{4}
\textsuperscript{1}Xidian University, \textsuperscript{2}Hunan University, \textsuperscript{3}Sun Yat-Sen University, \textsuperscript{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 \url{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.

The key observation is that 3D scene structures are inherently hierarchical [20]. By definition, an instance can have multiple supports, \textit{lamps} are standing on a \textit{table}, \textit{chairs} are supported by the \textit{floor} and only the \textit{floor} does not have any support, and it is unlikely that a \textit{pillow} is supporting a \textit{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 systematically 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 \textit{TV} and a \textit{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 \textit{table} easily which supports the \textit{TV} and use the \textit{table} as a bridge to predict the ‘front’ relationship with \textit{sofa}, where \textit{table} and \textit{sofa} are all supported by the \textit{floor} and relationships within them is intuitive.
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
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

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 common 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 $K_e = \{V_e, E_e\}$ where $V_e$ and $E_e$ represent nodes and edges respectively. To merge the hierarchical structures in 3D spaces into the external knowledge graph and construct the hierarchical
We assign the semantic features 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 $K_s$. 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 $K_s$. 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 $G_v = \{V_v, E_v\}$ where $V_v$ and $E_v$ denote object instances and edges of object pairs respectively, under the guidance of the hierarchical symbolic knowledge graph $K_s$. Then, a region-aware graph network is employed to propagate node messages through the visual graph $G_v$, 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 $F_v$ 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 $F_s$ 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 $G_v$ into its corresponding layer according to the hierarchical tokens in hierarchical symbolic knowledge graph $K_s$. Then, we complete the edge set $E_v$ of visual graph $G_v$ 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 $G_v$ if a support edge also exists between the corresponding nodes in the hierarchical symbolic knowledge graph $K_s$.

**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 $G_v$ and encode the hierarchical contexts of the input scene.

Given the initial representations of nodes and edges in the visual graph $G_v$, the region-aware graph network iteratively updates the hidden state $h_{i,t}^{o,t}$ of each node $v_i$ and $h_{ij,t}^{e,t}$ 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 $G_v$, 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 $G_v$. Specifically, the enriched hidden state $\tilde{h}_{i,t}^{o,t}$ of each node is:

$$\tilde{h}_{i,t}^{o,t} = h_{i,t}^{o,t} + \sum_{j \in N_r(i)} \psi(h_{j,t}^{o,t})$$

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

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

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

$$h_{i,t}^{o,t+1} = GRU(\tilde{h}_{i,t}^{o,t}, m_{i,t}^{o,t})$$

$$h_{ij,t}^{e,t+1} = GRU(\tilde{h}_{ij,t}^{e,t}, m_{ij,t}^{e,t})$$

where $m_{i,t}^{o,t}$ and $m_{ij,t}^{e,t}$ are the incoming messages for updating each node and edge. The calculation of the message for each node is:

$$m_{i,t}^{o,t} = \sum_{j \in N_r(i)} \left( \varphi_n(h_{j,t}^{o,t}) + \varphi_e(h_{ij,t}^{e,t}) \right)$$

where $N_r(i)$ denotes the neighbor nodes of $v_i$ in the visual graph $G_v$, $\varphi_n$ and $\varphi_e$ 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:

$$m_{ij,t}^{e,t} = \varphi_s(\tilde{h}_{ij,t}^{e,t}) + \varphi_o(\tilde{h}_{ij,t}^{o,t})$$

We take the final hidden states of nodes and edges as the contextual feature $c_i^t$ for each node $v_i \in V_v$ and $c_{ij}^t$ for each edge $(v_i, v_j) \in E_v$. 

9185
3.3. Spatial Multimodal Knowledge Accumulation

Though our hierarchical symbolic knowledge graph \( K_s \) 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 \( K_m \) progressively from the visual context via a graph reasoning network. We then incorporate the learned multimodal knowledge \( K_m \) 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 \( K_s \) to accumulate 3D spatial multimodal knowledge \( K_m \) 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 \( K_m \), 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 \( K_s \). Given the contextual features of nodes and edges in visual graph \( G_v \), 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 \( K_s \), (2) a 0/1 indicator of whether this node or edge appears in the visual graph \( G_v \), (3) the contextual feature \( c_i^o \) and \( c_{ij}^e \) in the visual graph \( G_v \) 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 \( K_s \). Specifically, at each time step \( t \), to calculate the hidden states \( d_i^{o,t} \) for all nodes \( s_i \in V_s \) and \( d_{ij}^{e,t} \) for all edges \( (s_i, s_j) \in E_s \), each node and edge first gather messages from their neighbors through the graph structure then update their hidden states:

\[
d_i^{o,t+1} = GRU(d_i^{o,t}, m_i^{o,t}), \tag{8}
\]

\[
d_{ij}^{e,t+1} = GRU(d_{ij}^{e,t}, m_{ij}^{e,t}), \tag{9}
\]

where \( m_i^{o,t} \) and \( m_{ij}^{e,t} \) are the incoming messages for nodes and edges. The incoming message for each node is

\[
m_i^{o,t} = \sum_{j \in N_h(i)} (\varphi_n(d_j^{o,t}) + \varphi_e(d_{ij}^{e,t})), \tag{10}
\]

where \( N_h(i) \) denotes the neighbor nodes of node \( s_i \) in the knowledge graph \( K_s \). Similar to Eq. (7), the incoming message for each edge is

\[
m_{ij}^{e,t} = \varphi_s(d_i^{o,t}) + \varphi_o(d_j^{o,t}). \tag{11}
\]

We take the sum of the stacked hidden states as the 3D spatial multimodal knowledge embedding \( b_i^o \) for all nodes and \( b_{ij}^e \) for all edges in the symbolic knowledge graph \( K_s \).

**Knowledge-enabled Scene Graph Prediction.** To incorporate the 3D spatial multimodal knowledge \( K_m \) 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 \( \varphi_b \) and \( \varphi_c \) respectively before fusing them. For each node in the visual graph, we fuse the retrieved multimodal knowledge embedding \( \{b_k^e\}_{k=1,2,3} \) and the contextual node feature \( c_i^o \) to obtain the knowledge-enabled contextual feature \( f_i^o \):

\[
f_i^o = \phi(\varphi_c(c_i^o) + \varphi_b(\sum_{k=1}^{3} b_k^e)). \tag{12}
\]

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 \( f_i^o \) for nodes and \( f_{ij}^e \) 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 \( f_i^o \), and each edge is initialized either by the contextual edge feature \( f_{ij}^e \) 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 \( c_i^o \) is used to predict the initial class guesses, we use a cross entropy loss \( L_{\text{init}}^o \) for the initial detection. For the final prediction, we use two cross entropy...
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-
The results are shown in Tab. 3.

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

<table>
<thead>
<tr>
<th>Methods</th>
<th>R@50/100</th>
<th>mR@50/100</th>
</tr>
</thead>
<tbody>
<tr>
<td>Knowledge $K_s$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>w/o Hierarchical Tokens</td>
<td>30.47 / 30.67</td>
<td>28.94 / 29.19</td>
</tr>
<tr>
<td>w/o Support Edge</td>
<td>30.55 / 30.74</td>
<td>29.17 / 29.47</td>
</tr>
<tr>
<td>w/o Both</td>
<td>28.41 / 28.47</td>
<td>27.13 / 27.52</td>
</tr>
<tr>
<td>Visual Context Encoding</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$G_v$ replaced w/ $G_{fc}$</td>
<td>28.17 / 28.32</td>
<td>26.28 / 26.29</td>
</tr>
<tr>
<td>w/o RaGN</td>
<td>26.43 / 26.57</td>
<td>24.23 / 24.36</td>
</tr>
<tr>
<td>RaGN replaced w/ GCN</td>
<td>31.03 / 31.21</td>
<td>29.67 / 29.88</td>
</tr>
<tr>
<td>Knowledge $K_m$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>w/o $b_{ij}$ and $b_{ij}'$</td>
<td>26.27 / 26.35</td>
<td>22.93 / 23.18</td>
</tr>
<tr>
<td>w/o $e_{ij}'$ and $e_{ij}$ as input</td>
<td>28.14 / 28.31</td>
<td>25.05 / 25.31</td>
</tr>
<tr>
<td>Ours</td>
<td>31.50 / 31.64</td>
<td>30.29 / 30.56</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Methods</th>
<th>PredClS</th>
<th>SGCIS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>R@50</td>
<td>mR@50</td>
</tr>
<tr>
<td>$G_v$</td>
<td>62.74</td>
<td>58.25</td>
</tr>
<tr>
<td>$G_t$</td>
<td>68.41</td>
<td>66.59</td>
</tr>
<tr>
<td>$G_v$(original)</td>
<td>68.32</td>
<td>66.54</td>
</tr>
</tbody>
</table>

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

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 $K_s$ 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 mod-

3D spatial multimodal knowledge accumulation. Lastly, we examine the accumulated multimodal knowledge $K_m$ to learn about how $K_m$ and rest of the model interact. We first see how much of the improvement comes from the 3D spatial multimodal knowledge $K_m$. 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 $e_{ij}'$ 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.

Further Analysis

Analysis on the hierarchical structure of 3D spaces. To validate the potential of the hierarchical visual graph $G_v$ in capturing the inherent hierarchical structure of a 3D scene, we design two visual graph variants and compare them to the hierarchical visual $G_v$: (1) Instead of using the hierarchical symbolic knowledge graph $K_s$, we build a ground truth graph $G_t$ based on the ground truth labels for support relations. In particular, each edge in $G_t$ represents the ground truth support relationship of the input scene. (2) We also design a randomly connected graph $G_r$, where we keep all of the nodes the same but randomize the edges that connect them. As shown in Tab. 4, both $G_v$ and $G_t$ outperform $G_r$. We can see that replacing the hierarchical visual graph $G_v$ with a fully-connected graph $G_{fc}$ 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.

3D spatial multimodal knowledge accumulation. Lastly, we examine the accumulated multimodal knowledge $K_m$ to learn about how $K_m$ and rest of the model interact. We first see how much of the improvement comes from the 3D spatial multimodal knowledge $K_m$. 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 $e_{ij}'$ for nodes and $e_{ij}$ 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 $G_v$ in capturing the inherent hierarchical structure of a 3D scene, we design two visual graph variants and compare them to the hierarchical visual $G_v$: (1) Instead of using the hierarchical symbolic knowledge graph $K_s$, we build a ground truth graph $G_t$ based on the ground truth labels for support relations. In particular, each edge in $G_t$ represents the ground truth support relationship of the input scene. (2) We also design a randomly connected graph $G_r$, where we keep all of the nodes the same but randomize the edges that connect them. As shown in Tab. 4, both $G_v$ and $G_t$ outperform $G_r$. We can see that replacing the hierarchical visual graph $G_v$ with a fully-connected graph $G_{fc}$ 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.
with a significant margin on all metrics. More importantly, we observe that $G_t$ and $G_v$ perform mostly similar while $G_t$ slightly outperforms $G_v$. The results confirm that the hierarchical visual graph $G_v$ 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 $K_m$ 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 $G_v$ and 3D scene graphs $G$ 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.

### 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


