3D Spatial Multimodal Knowledge Accumulation for Scene Graph Prediction in Point Cloud (supplementary material)

1. Methodology Additional Details

1.1. Data Preparation

In this section, we introduce the data preparation for the Point Cloud Transformer and scene graph prediction model training. Each input scene is a point cloud of $40,000$ points randomly sub-sampled from a 3D scan. In addition to XYZ coordinates, we also include the RGB value of each point to facilitate visual feature encoding. For each object instance in the input scene, we sample the point number to 512 using farthest point sampling (FPS). We randomly scale the points by Uniform$[0.9; 1.1]$. Since the 3DSSG [6] dataset contains proximity relationships such as *left* and *right*, which rely on the heading angle of each object, we do not perform rotation augmentation for the dataset.

1.2. Knowledge Graph Construction

Hierarchical Tokens. We employ a Multi-layer Perceptron (MLP) to classify the hierarchical token for each node s_i in the external knowledge graph $\mathcal{K}_e$ to distinguish the discrepancy between different layers of nodes. Specifically, we first map each node s_i into word embedding vector $\mathbf{u}_i$ using GloVe [3] and then apply the MLP to $\mathbf{u}_i$ to learn the node representation $\mathbf{x}_i$. A fully connected layer and softmax function are then applied to $\mathbf{x}_i$, resulting in the hierarchical token $\mathbf{z}_i \in \mathbb{R}^3$. The three values in $\mathbf{z}_i$ correspond to the probabilities that node s_i belongs to the three layers in the hierarchical structures, i.e., the first, second and third layer. The hierarchical token $\mathbf{z}_i$ is then projected to an embedding vector using a randomly initialized trainable embedding table $\mathbf{W}_l \in \mathbb{R}^{3\times 200}$. Since the layer annotation of each object category is not directly available from existing datasets, we use the manually annotated labels for relationships in the 3DSSG dataset to determine the layer of each object category in the dataset and obtain their layer annotations. We then use the layer annotations as the training data to train the MLP before integrating it into our model.

Support Edges. One of the crucial elements of the hierarchical structures in 3D scenes is the proposed support edges in the hierarchical symbolic knowledge graph $\mathcal{K}_s$. Recall that, we add a support edge between two correlated nodes in neighboring layers to represent the physical support relationships between nodes. Specifically, two nodes in the knowledge graph $\mathcal{K}_s$ are considered correlated when there exists a two-hop path between them in the knowledge graph $\mathcal{K}_s$. Any existing knowledge edge between two nodes is replaced by the newly-added support edge.

1.3. Visual Graph Construction

In the hierarchical visual graph construction stage, for each detected object instance in the input scene, the visual feature f_v encoded by Point Cloud Transformer [2] is a 1024 dimensional feature vector while 256 is the dimension of spatial feature f_t. We set the dimension of semantic feature f_w to 200. The object feature becomes 1480 dimensional after concatenating the visual feature f_v, spatial feature f_t and semantic feature f_w.

2. Extensive Experiments on 3DSSG Dataset

Quantitative results and comparison. We report both the recall@K (R@K) and mean recall@K (mR@K) without graph constraint for comprehensive comparison with several existing state-of-the-art 2D and 3D scene graph prediction methods. Unconstrained recall and mean recall omit the graph constraint that merely one relationship is obtained for a given object pair, so that multiple relationships can be obtained, leading to higher values. Tab. 1 summarizes the performances of re-implemented 2D scene graph generation models on the 3DSSG dataset. We observe that the performance improvement is considerable when compared to the constrained recall and mean recall in the main text, this is partly because unconstrained recall and mean recall involves all the 27 relation scores (including *none*) of each subject-object pair in the recall ranking. As shown, our model achieves the best performance compared with other 2D scene graph generation models when evaluated on the 3DSSG dataset. Our model improves the R@50 of 2.17% and 12.97% in SGCls and PredCls tasks compared to other methods. This highlights the advantage of exploiting hierarchical structures of 3D spaces, pursuing a thorough 3D space understanding for scene graph prediction. It is also worth noting that our model achieves higher mean recall performance in both SGCls and PredCls tasks compared to re-implemented 2D image-based scene graph generation

Methods	PredCls		SGCls		SGDet	
	R@50/100	mR@50/100	R@50/100	mR@50/100	R@50/100	mR@50/100
3D+IMP [8]	57.44 / 59.37	28.38 / 31.28	18.30 / 18.51	11.71 / 12.48	25.32 / 25.84	22.74 / 23.66
3D+MOTIFS [9]	60.16 / 63.89	31.17 / 34.53	19.48 / 20.11	12.51 / 14.29	27.62 / 27.63	25.26 / 25.85
3D+VCTree [5]	61.37 / 65.19	34.71 / 39.36	22.41 / 24.16	18.17 / 21.64	29.37 / 30.42	26.63 / 26.82
3D+KERN [1]	64.15 / 76.54	36.33 / 40.27	24.19 / 27.03	19.40 / 22.39	29.81 / 31.03	26.62 / 26.93
3D+Schemata [4]	66.71 / 79.34	53.57 / 58.41	33.13 / 35.88	28.34 / 31.35	30.54 / 31.85	28.41 / 28.83
3D+HetH [7]	66.85 / 79.69	53.44 / 58.37	33.25 / 35.62	28.13 / 31.74	30.37 / 31.68	28.31 / 28.72
Ours	**79.82 / 89.62**	**78.34 / 87.39**	**35.42 / 37.72**	**34.47 / 37.15**	**31.35 / 32.49**	**29.14 / 29.57**

Table 1. Comparison with state-of-the-art 2D scene graph prediction methods, without graph constraint.

Methods	PredCls		SGCls		SGDet	
	R@50/100	mR@50/100	R@50/100	mR@50/100	R@50/100	mR@50/100
SGPN [6]	69.56 / 82.04	49.75 / 54.38	32.41 / 35.14	30.27 / 33.43	- / -	- / -
EdgeGCN [10]	76.43 / 84.52	72.79 / 83.47	32.61 / 34.77	31.86 / 34.58	- / -	- / -
KISG [11]	78.25 / 88.17	77.48 / 86.24	34.13 / 36.26	34.02 / 36.13	- / -	- / -
Ours	**79.82 / 89.62**	**78.34 / 87.39**	**35.42 / 37.72**	**34.47 / 37.15**	**31.35 / 32.49**	**29.14 / 29.57**

Table 2. Comparison with 3D scene graph prediction methods on the 3DSSG dataset, without graph constraint.

models, suggesting that our model can improve the prediction of less frequent relationships. We also compare the unconstrained recall and mean recall with several state-of-the-art 3D point-based scene graph prediction models. The results are reported in Tab. 2. We can see that our model consistently achieves competing performances against existing methods. The hierarchical structures of 3D spaces allows our model to reason multiple relationships hierarchically and systematically, leading to more fine-grained multi output relationships.

Evaluation on individual relationships. To illustrate the performance gain of each individual relationship category brought by the hierarchical structure of 3D spaces, we compare the performance of each relationship category on R@5 of our model and KISG [11]. As shown in Fig. 1, our model improves the R@5 on most relationships. The improvement is obvious for proximity relationships, including *left*, *right*, *front* and *behind*. The class-related priors adopted in KISG can not efficiently represent the proximity relationships since proximity relationships depend heavily on the visual content of the specific point cloud scene. In contrast, the hierarchical structure correlations between object pairs allows our model to infer complex proximity relationships based on not only geometric analysis but also the regular structure patterns of the scene graph. We can also see that our model achieves an improvement of 0.86% on *same as*. This is partly because the hierarchical structure patterns of 3D scenes also improves object classification. In addition, our model significantly improves the performance of certain tail relationships, such as *part of* and *lying in*, and has less and acceptable damage to the performance of the head relationships.

3. Additional Ablations

In this section, we show the results of the final sets of ablations.

Settings	R@50/100	mR@50/100
One-hop	30.17 / 30.45	28.56 / 28.74
Three-hop	29.84 / 30.15	28.38 / 28.56
Two-hop(original)	**31.50 / 31.64**	**30.29 / 30.56**

Table 3. Quantitative results of different support edges in the hierarchical symbolic knowledge graph $\mathcal{K}_s$ on the SGCls task.

Firstly, we conduct additional ablation studies on the impact of adding support edges between different nodes in the hierarchical symbolic knowledge graph $\mathcal{K}_s$. We use two settings here: one-hop and three-hop. One-hop means we add support edges between nodes in adjacent layers when there exists one-hop paths between the nodes. Three-hop means we add support edges between nodes in adjacent layers if at least one three-hop path exist between the nodes. As shown in Tab. 3, we can see that both the one-hop and three-hop setting negatively impact the performance of our model compared to the original two-hop setting. The reason is that one-hop setting ignores potential physical support relationships and three-hop setting includes many irrelevant edges that do not represent physical support relationships between nodes.

Variant	R@50/100	mR@50/100
Same layer	28.51 / 28.78	24.53 / 24.77
Same support(original)	**31.50 / 31.64**	**30.29 / 30.56**

Table 4. Quantitative results of different contextual region definition on the SGCls task.

Next, we examine the contextual regions around each node in the hierarchical visual graph $\mathcal{G}_v$. In our work, the contextual regions around each node in the hierarchical visual graph $\mathcal{G}_v$ are defined as other nodes sharing the same physical support with it. To demonstrate the efficacy of this

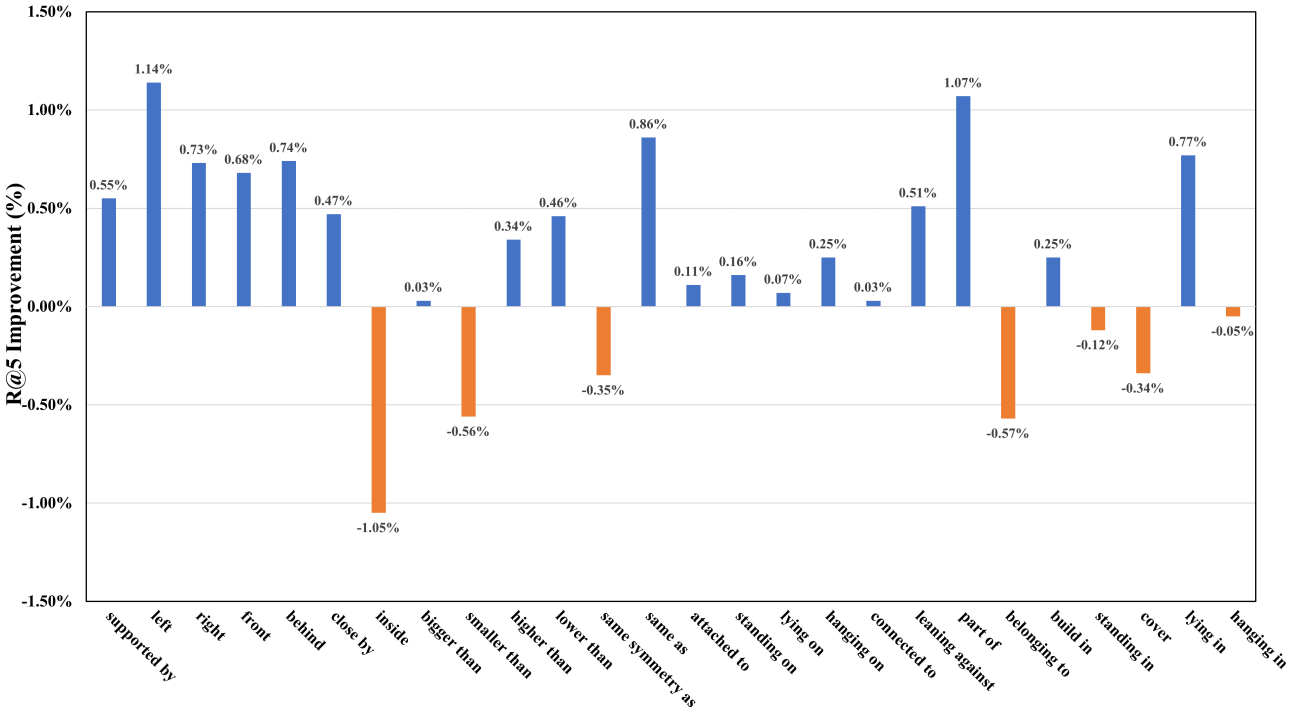

Figure 1. Improvement in PredCls task of our model for R@5 in comparison with KISG.

design, we devise a variant of our model, which defines the contextual regions around each node as the nodes in the same layer instead of the nodes sharing the same physical support. The results are shown in Tab. 4. We observe that adopting this variant of the model drops the R@50 and mR@50 in SGCls by 2.99% and 5.76% respectively. It shows that objects sharing the same physical support are closely-correlated, which helps the scene graph generation.

Methods	R@50/100	mR@50/100
w/o embedding	27.35 / 28.02	24.15 / 24.28
w/o $0/1$ indicator	30.83 / 30.97	29.86 / 29.99
w/o $\mathbf{c}_i^o$	28.77 / 29.21	25.53 / 25.68
w/o $\mathbf{c}_{ij}^e$	30.12 / 30.36	28.77 / 28.94
Ours	**31.50 / 31.64**	**30.29 / 30.56**

Table 5. Performance comparison of different input configuration of the graph reasoning network on the SGCls task.

Finally, we look at the inputs to the graph reasoning network in the 3D spatial multimodal knowledge accumulation module. Each node and edge in the graph reasoning network receives three inputs, including the node or edge embedding from the hierarchical symbolic knowledge graph $\mathcal{K}_s$, the $0/1$ indicator, and the contextual feature $\mathbf{c}_i^o$ for nodes or $\mathbf{c}_{ij}^e$ for edges. As shown in Tab. 5, removing the embedding from the hierarchical symbolic knowledge graph $\mathcal{K}_s$ drastically decreases the R@50 by a margin of 4.15%, indicating that symbolic knowledge is essential for progressively accumulating multimodal knowledge $\mathcal{K}_m$. We also ablate the $0/1$ indicator and find that this part makes the least difference, dropping the performance less than 1%. Additionally, removing the contextual feature input $\mathbf{c}_i^o$ only for nodes in the graph reasoning network decreases the performance significantly, much more than the affect of removing the contextual feature input $\mathbf{c}_{ij}^e$ only for edges.

4. Additional Qualitative Results

We visualize two more sample generated by our model in Fig. 2. The first sample is demonstrated through Fig. 2(a)-(c). Specifically, (a) shows the input scene, (b) shows the hierarchical visual graph $\mathcal{G}_v$, and (c) shows the predicted 3D scene graph $\mathcal{G}$. We can see that the hierarchical visual graph in (b) accurately describes the hierarchical structure of the input scene in (a) based on the physical support relationships between objects. Even though objects such as *desk* and *monitor* are misclassified, their hierarchical tokens are correctly classified, which demonstrates the effectiveness of the hierarchical symbolic knowledge graph $\mathcal{K}_s$. As we can see in (c), our model successfully predicts most relationships in the input scene. It is mainly because our model utilizes both the hierarchical structure of the input scene and the 3D spatial multimodal knowledge $\mathcal{K}_m$ to reason complex relationships hierarchically and systematically. Proximity relationships, such as *left* and *front*, are difficult to predict due to the complex and diverse spatial arrangement of 3D objects. Our model incorrectly classifies some proximity relationships while still achieves many correct predictions

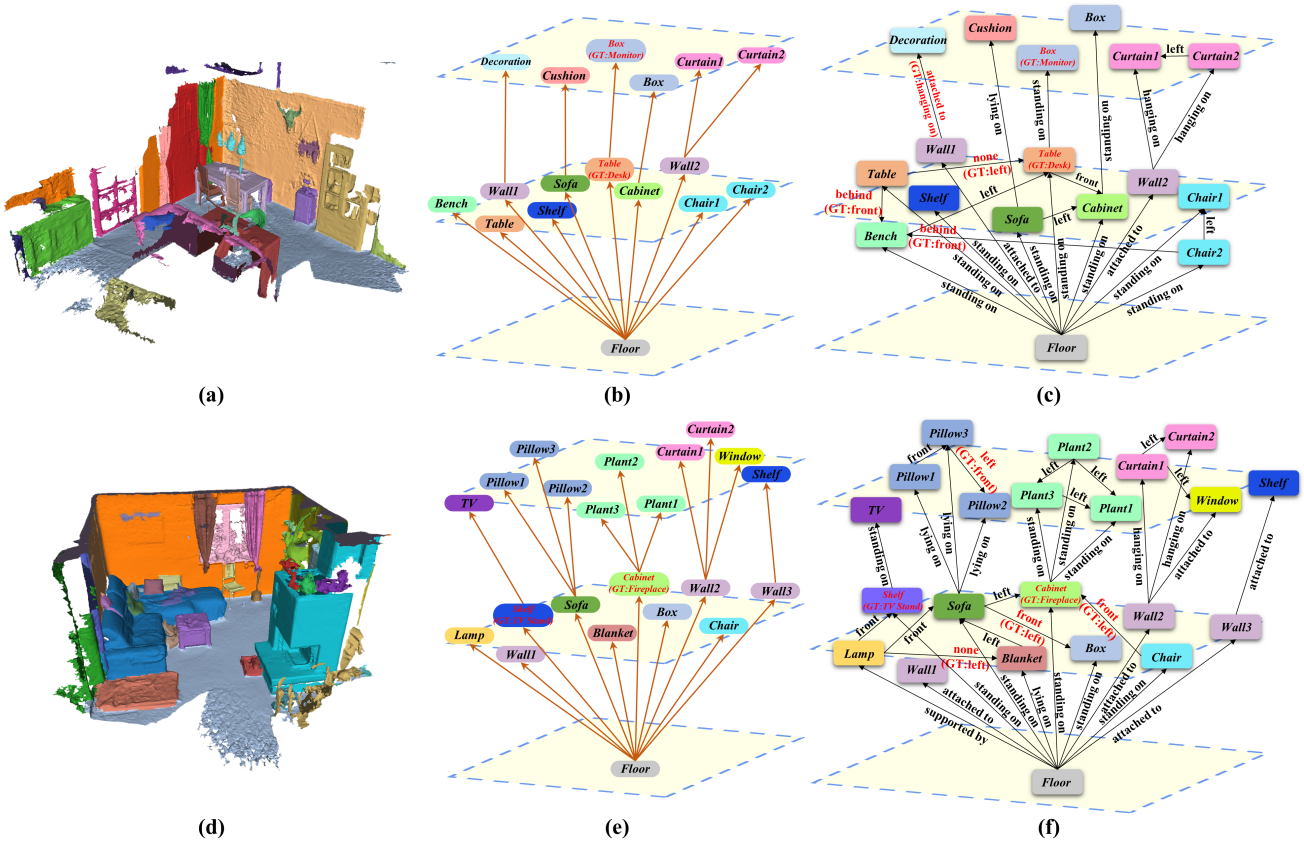

Figure 2. More visualization samples of our predicted scene graph on 3DSSG dataset. Red indicates the misclassified objects or relationships.

with the proposed method.

Fig. 2(d)-(f) shows a more challenging sample as there are more objects in the 3D scene. We can see that the predictions of objects and relationships are relatively stable compared with the first example, proving that our model possesses excellent generalization capabilities. Our model struggles to predict the correct object category when the visual and semantic information of the object are ambiguous as our model incorrectly classifies the *tv stand* and *fireplace* as *shelf* and *cabinet* respectively. The hierarchical visual graph in (e) is not affected much when there are incorrect object predictions. In (f), we can see that our model identifies the *curtains* and *plants* and the associated relationships. These objects share the same physical support and thus are closely-correlated. The result proves that exploiting the contextual regions around each objects, which are defined as the objects sharing the same physical support, is beneficial for scene graph generation.

References

[1] 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

[2] 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. 1

[3] 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. 1

[4] 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. 2

[5] 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] 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

[7] 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. 2

[8] 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

[9] 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

[10] Chaoyi Zhang, Jianhui Yu, Yang Song, and Weidong Cai. Exploiting edge-oriented reasoning for 3d point-based scene graph analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9705–9715, 2021. 2

[11] 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