


Exploring Contextual Relationships in 3D Cloud Points by Semantic Knowledge Mining: Supplementary Material

Lianggangxu Chen,¹  Jiale Lu,¹ Yiqing Cai,¹ Changbo Wang^{1,†} and Gaoqi He^{2,†}

¹School of Computer Science and Technology, East China Normal University, Shanghai, China

² School of Computer Science and Technology, Shanghai Key Laboratory of Mental Health and Psychological Crisis Intervention, East China Normal University, Shanghai, China.

In this supplementary material, we provide additional implementation details for our method in Section 1. In Section 2, we present detailed analysis of the 3D semantic scene graph (3DSSG) dataset [WDNT20]. In Section 3, we report an analysis of the computational cost of our method. In Section 4, we show additional experimental results. More ablation studies of our method are shown in Section 5. In Section 6, we show the future directions of 3D scene graph generation.

1. Additional implementation details

We follow the same data preparation in [WDNT20] and [ZHQ*21]. For SC-GCN, we set the dimension of node representations to 1024, and perform 2 message aggregation iterations. The project function $R(\cdot)$ in Eqn. (2) is implemented by using a 1×1 convolutional layer. We train the multi-scale PointNet (MS PointNet) for entity classification with the focal loss [LGG*17] mentioned in our main paper. We trained each model three times to calculate the standard deviation. We contacted the author by email (johanna.wald@tum.de) to get the permission of 3DSSG dataset. We implemented the KISGP model [ZHQ*21], SGF model [WWT*21] and EdgeGCN model [ZYSC21] based on its released code with MIT license. The SGP model [WDNT20] have no public code for now. We reproduced them based on their papers. All models are trained on the 3DSSG dataset with the same random seeds and the same split for a fair comparison.

2. Detailed analysis of the 3DSSG dataset

The 3DSSG dataset annotates support, proximity, and comparative predicates among daily indoor objects. To gain insight into the 3DSSG dataset, we conducted an analysis of repeated predicates in 3D scene graphs between same object pairs. Over 31.5% of 3D scenes in 3DSSG dataset contain the predicates between same object pairs, and some 3D scenes even contain ten same objects, which does not occur in the dataset of the image scene graph (Visual genome) [KZG*17]. Most of the scenes in the visual genome dataset are outdoor scenes.

For the scene id (8eabc455-5af7-2f32-8606-a0bdb6c537d) in

the test set of 3DSSG dataset: it contains 10 pictures. The predicates between them are all comparative predicates, such as **higher than** and **bigger than**. The scene id (f2c76ff1-2239-29d0-87f5-8a0346584384) in the test set: it contains many pillows, and there are all proximity predicates between them (**left**, **right**). In mispredictions of SOTA [ZHQ*21], at least 34% of the results are comparative and proximity predicates. The above analysis shows that our motivation is promising, and addressing predicate prediction errors between same objects can greatly improve the accuracy of the final 3D scene graph generation results.

Figure 1 is a visualization result of entity level semantic clues in the 3DSSG dataset. We show the top 6 triples with the largest number and the top 4 triples with the least number. In particular, there is only one predicate **hanging on** between the **lamp** and the **ceiling**, thus propagating this information can better help us predict this predicate.

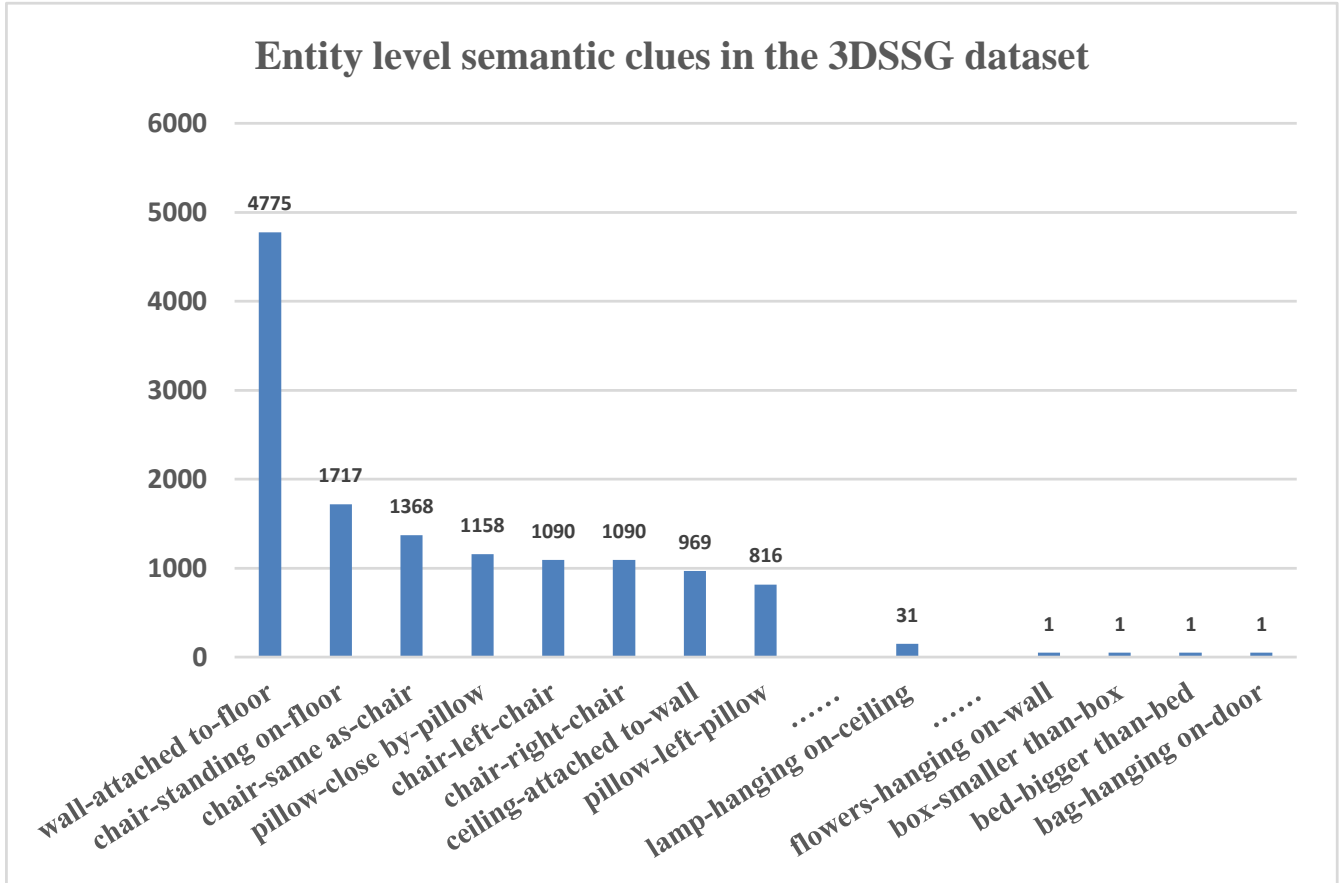
Figure 2 shows the sankey diagram results for path level semantic clues. Our path is directional, and the s on the left represents source entity, the t on the right represents target entity. Additionally, Figure 2 visualizes the strength of the entity connection. We see that the dataset provides a strong bias for predicting predicates given the entity pairs, which is fully used by our model in the main paper.

3. Train time and memory cost

We compute the training speed and memory cost of our method and compare to KISGP [ZHQ*21] using one Nvidia RTX 2080Ti GPU with 11 gigabytes of memory, and summarize the results in Table 1. In general, our results are significantly more accurate as demonstrated in the main paper by at the cost of a little increase in the train time. Another important factor is the trainable parameters. Our method has 13.7% more parameters than KISGP. One of the reasons is that our method has two types of semantic clues with graph convolutional network. In our message passing method, we both consider single neighbor and neighbor pairs jointly by 2 steps of global message passing.

Table 1: Train time and memory cost of our method compared to the KISGP model [ZHQ*21].

Model	Train time (min/epoch)	Memory (MiB)
KISGP [ZHQ*21]	22.83	1431
Our method	23.87	1627

**Figure 1:** The statistical results of entity level semantic clues in the 3DSSG dataset.

4. Additional experimental results

Can our model really understand semantic knowledge? In order to verify that semantic knowledge really improves performance through co-occurrence probability in the train set, instead of using powerful label feature representation or more clearly feature refinement by GCN, we trained our model with the processed training set. We randomly swap co-occurrence probabilities between different entities, which essentially changes the weights in the GCN. As shown in Table 2, the average recall value (with constraint) drops significantly from 0.508 to 0.488, when 25% of the training scenes are swapped, which is equivalent to adding noise in the semantic knowledge. The results drop further as more scenes are processed. The experimental result (third row) is in line with expectations: average recall drops to 0.475. In fourth row, it is even lower than

KISGP [ZHQ*21], processed scenes indicate the semantic knowledge is completely shattered and the noise is further amplified. The results demonstrate that the semantic knowledge is learned in our framework.

More evaluation metrics: In the main paper, we compare the standard scene graph generation evaluation metrics in [ZHQ*21]. [WDNT20] also proposed three other evaluation metrics in their paper (Relationship Prediction, Object Class Prediction and Predicate Prediction). We also compare these evaluation metrics in table 3. Their metrics are generally higher due to the unstrict calculation. Taking R@100 as an example, their approach is to calculate 100 possibilities for each triplet, and then see if they appear in the ground truth.

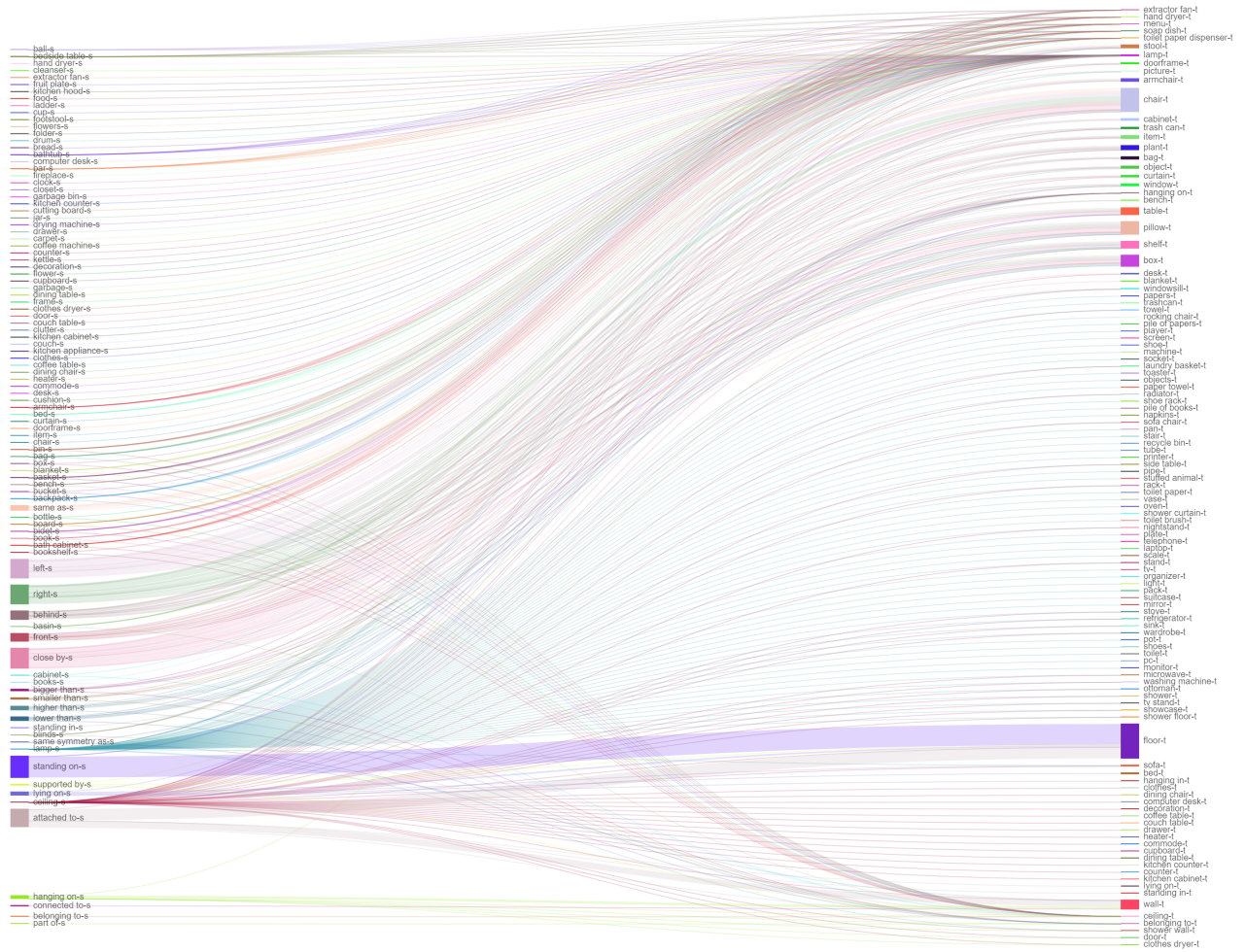


Figure 2: A bipartite mapping of path level semantic clues in the 3DSSG dataset. The width of the curve represents the statistical results of path numbers between different entity pairs.

Table 2: We shuffle the scenes in the training set to explore the sensitivity of the model to semantic clues.

Swapped scenes	Original scenes	SGCLS			PREDCLS			
		R@20	R@50	R@100	R@20	R@50	R@100	Mean
1	0	0.335	0.360	0.362	0.601	0.662	0.728	0.508
25%	75%	0.330	0.343	0.354	0.599	0.645	0.655	0.488
50%	50%	0.295	0.326	0.337	0.598	0.642	0.651	0.475
75%	25%	0.283	0.289	0.292	0.583	0.610	0.633	0.448

3D scene graph generation with failed semantic segmentation: We follow previous works to separate node sets of objects using instance segmentation results. Incorrect segmentation can lead to the following failure cases: 1)The object is incorrectly detected or not detected; 2)The predicate is ambiguous and difficult to be identified even by humans. Figure 3(a) shows the ground truth of triplet (towel, hanging in, shower). Figure 3(b) shows the classification error caused by the failed towel segmentation result.

5. Additional ablation studies

Comparison of different loss functions: As shown in Table 4, the effects of different loss functions are compared on the mean recall. The effect of focal loss depends on the focusing parameters, and the best effect is when the focusing parameter $\gamma = 2$. When the $\gamma = 0$, it is the classical BCE loss used in SGG. The result of Focal loss is better than the BCE loss, which proves its effectiveness in 3D SGG.

Table 3: Evaluation of the Relationship Prediction, Object Class Prediction and Predicate Prediction task on 3DSSG [WDNT20] dataset.

Method	Relationship Prediction		Object Class Prediction		Predicate Prediction	
	R@50	R@100	R@5	R@10	R@3	R@5
Baseline [WDNT20]	0.39	0.45	0.66	0.77	0.62	0.88
SGPN [WDNT20]	0.40	0.66	0.68	0.78	0.89	0.93
SGF [WWT*21]	0.85	0.87	0.70	0.80	0.97	0.99
EdgeGCN [ZYSC21]	0.40	0.49	0.91	0.98	0.79	0.91
KISGP [ZHQ*21]	0.86	0.92	0.99	0.98	0.72	0.86
Our method	0.87	0.89	0.99	0.98	0.97	0.99

Table 4: Ablation studies on different loss functions.

Loss	SGCLS			PREDCLS			
	mR@20	mR@50	mR@100	mR@20	mR@50	mR@100	Mean
$\gamma = 0$ (BCE-loss)	0.204	0.227	0.278	0.547	0.620	0.621	0.416
$\gamma = 2$ (Focal-loss)	0.254	0.297	0.298	0.577	0.640	0.643	0.452
$\gamma = 3$ (Focal-loss)	0.201	0.223	0.288	0.557	0.630	0.641	0.423

6. Future directions of 3D SGG

Exploration for outdoors is one of the future directions of 3D SGG. With the widespread availability of LiDARs, depth cameras and light field cameras, 3D point cloud data on outdoor scenes is becoming increasingly available and widely used in augmented and virtual reality, 3D object detection, and 3D semantic segmentation. Therefore, 3D SGG on outdoor scenes also interests all the authors. At present, there are only indoor datasets. The complex 3D SGG is also one of the future directions we are interested in. Some works of image-based complex SGG [JLY*22, KdVC*20] have used dual-hierarchy message propagation to refine the representation hierarchically and eliminate redundant information. Besides, projecting the point cloud into 2D space may be helpful for the prediction. There have been many works in 3D vision using information from 2D images to assist prediction, such as 3D semantic segmentation [NSL*21] and 3D dense captioning [YYL*22]. Incorporating knowledge of 2D images is also a possible future direction for the 3D SGG.

References

- [JLY*22] JIALE L., LIANGGANGXU C., YIQING C., HAOYUE G., CHANGHONG L., CHANGBO W., GAOQI H.: Dh-gcn: Saliency-aware complex scene graph generation using dual-hierarchy graph convolutional network. In *IEEE International Conference on Multimedia Expo* (2022). 4
- [KdVC*20] KNYAZEV B., DE VRIES H., CANGEA C., TAYLOR G. W., COURVILLE A., BELILOVSKY E.: Graph density-aware losses for novel compositions in scene graph generation. *arXiv preprint arXiv:2005.08230* (2020). 4
- [KZG*17] KRISHNA R., ZHU Y., GROTH O., JOHNSON J., HATA K., KRAVITZ J., CHEN S., KALANTIDIS Y., LI L.-J., SHAMMA D. A., ET AL.: Visual genome: Connecting language and vision using crowd-sourced dense image annotations. *International journal of computer vision* 123, 1 (2017), 32–73. 1
- [LGG*17] LIN T.-Y., GOYAL P., GIRSHICK R., HE K., DOLLÁR P.:

Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision* (2017), pp. 2980–2988. 1

- [NSL*21] NEKRASOV A., SCHULT J., LITANY O., LEIBE B., ENGELMANN F.: Mix3d: Out-of-context data augmentation for 3d scenes. In *2021 International Conference on 3D Vision (3DV)* (2021), IEEE, pp. 116–125. 4
- [WDNT20] WALD J., DHAMO H., NAVAB N., TOMBARI F.: Learning 3d semantic scene graphs from 3d indoor reconstructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2020), pp. 3961–3970. 1, 2, 4
- [WWT*21] WU S.-C., WALD J., TATENO K., NAVAB N., TOMBARI F.: Scenegrphfusion: Incremental 3d scene graph prediction from rgb-d sequences. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2021), pp. 7515–7525. 1, 4
- [YYL*22] YUAN Z., YAN X., LIAO Y., GUO Y., LI G., CUI S., LI Z.: X-trans2cap: Cross-modal knowledge transfer using transformer for 3d dense captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2022), pp. 8563–8573. 4
- [ZHQ*21] ZHANG S., HAO A., QIN H., ET AL.: Knowledge-inspired 3d scene graph prediction in point cloud. *Advances in Neural Information Processing Systems* 34 (2021). 1, 2, 4
- [ZYSC21] ZHANG C., YU J., SONG Y., CAI W.: Exploiting edge-oriented reasoning for 3d point-based scene graph analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2021), pp. 9705–9715. 1, 4

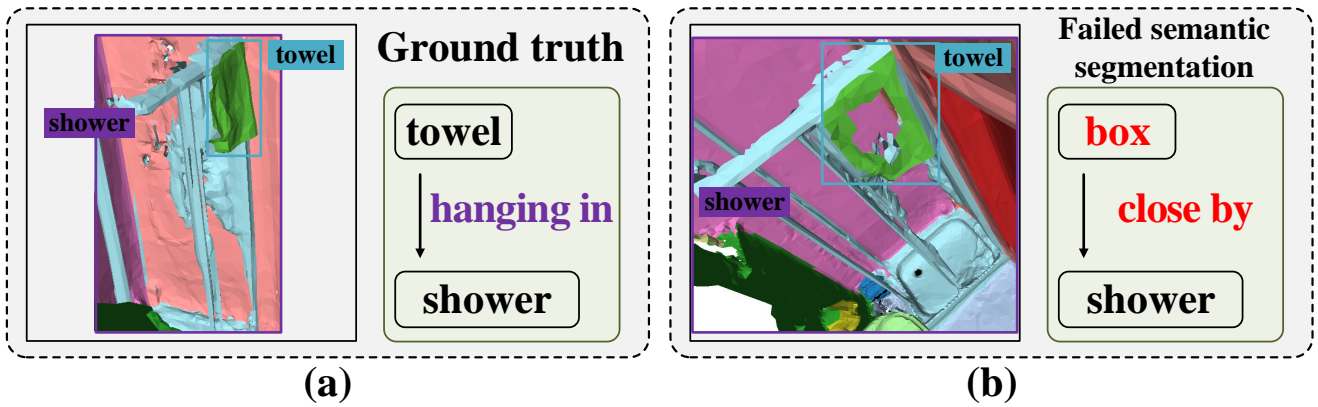


Figure 3: The most common failure cases. (a) The ground truth of triplet (towel, hanging in, shower). (b) The *towel* object is misclassified as *box*. The *hanging in* and *close by* predicates are ambiguous by failed semantic segmentation result.