Explore Contextual Information for 3D Scene Graph Generation —Supplementary Material—

***Abstract*—This supplemental material presents additional information and results supporting the main manuscript. Sec. I shows additional qualitative examples of our method and a detailed ablation study of our proposed architecture. Sec. II provides a mapping between fine-grained object categories and coarse-grained object categories.**

I. More Quantitative and Qualitative Results

In the main manuscript, we perform all evaluation methods on the 3DSSG [1] dataset. 3DSSG is composed of 3RSCAN [2] and scene graph annotations. It features 1482 scene graphs, which contain 534 classes of objects and 40 relationships. To alleviate the severe object class imbalance problem in the SGG task, SGGPoint [3] only retains 27 object classes and 16 relation classes (3DSSG-O27R16) in the 3DSSG dataset. And it also combines multi-label relationships between nodes into one relationship. This results in its network not being able to understand complex scenes with fine-grained objects and multiple relationships well. We argue that scene graphs based on coarse-grained labels and single relationships contain little scene context and are challenging to apply to subsequent high-level tasks. So we take the same 160 object categories and 26 predicate labels as in [1]. To demonstrate that our method can be applied to scenarios with different levels of complexity, we additionally provide the results of our method on 3DSSG-O27R16. As shown in Table I, our method outperforms SGGPoint on the few-category scene graph generation task.

TABLE I
Quantitative comparison with SGGPoint on 3DSSG-O27R16.

Model	Object Class Prediction		Predicate Prediction		Relationship Prediction	
	R@5	R@10	R@3	R@5	R@50	R@100
SGGPoint	90.10	97.12	81.54	82.10	54.65	55.17
Ours	**91.23**	**97.92**	**96.80**	**98.81**	**94.28**	**96.52**

In the main manuscript, we mentioned the drawback of KERN. When the object labels are predicted incorrectly, the relation prediction will be significantly affected, as wrong object labels will bring invalid co-occurrence knowledge to the process of feature information organization. The inconsistency between the final SGG results and the actual visual information also exists in other methods that simply use prior knowledge to organize environmental information. We additionally provide the results of [4], as shown in Table II, [4] advocates the use of graph auto-encoder to automatically extract class-dependent representations and topological patterns as prior knowledge to enhance the accuracy of relationship predictions. Perception and prior are naively treated as separate components, which are trained separately from different inputs (from images, triples, or label embeddings), and their predictions are usually fused in a probability space. As shown in the Table II, our method outperforms it in entity and relation prediction.

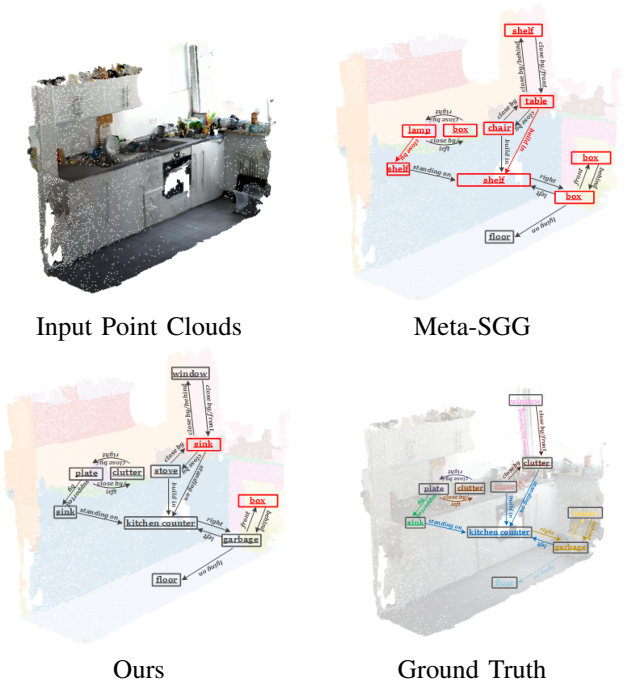

Fig. 1. **Comparisons against the Meta-SGG.** For clarity, we only show the results for a local area of the scene. We use gray boxes to indicate the entities and underlines to indicate the relations. The different colors of entity names and relation labels are the same as the colors of the class-agnostic instance labels, except that red indicates an incorrect prediction result.

TABLE II
Quantitative comparison with Meta-SGG.

Model	Object Class Prediction		Predicate Prediction		Relationship Prediction	
	R@5	R@10	R@3	R@5	R@50	R@100
Meta-SGG [4]	47.73	51.46	83.99	90.12	35.98	37.50
Ours	**73.40**	**82.59**	**89.90**	**96.10**	**61.94**	**68.24**

At the same time, we can also see the shortcomings of Meta-SGG in Figure 1. Since its graph encoder is independently trained, its prediction results are inconsistent with visual information, resulting in its prediction results only satisfying prior knowledge, and predicting a large number of entities that do not appear in the current scene, such as `table` and `shelf`.

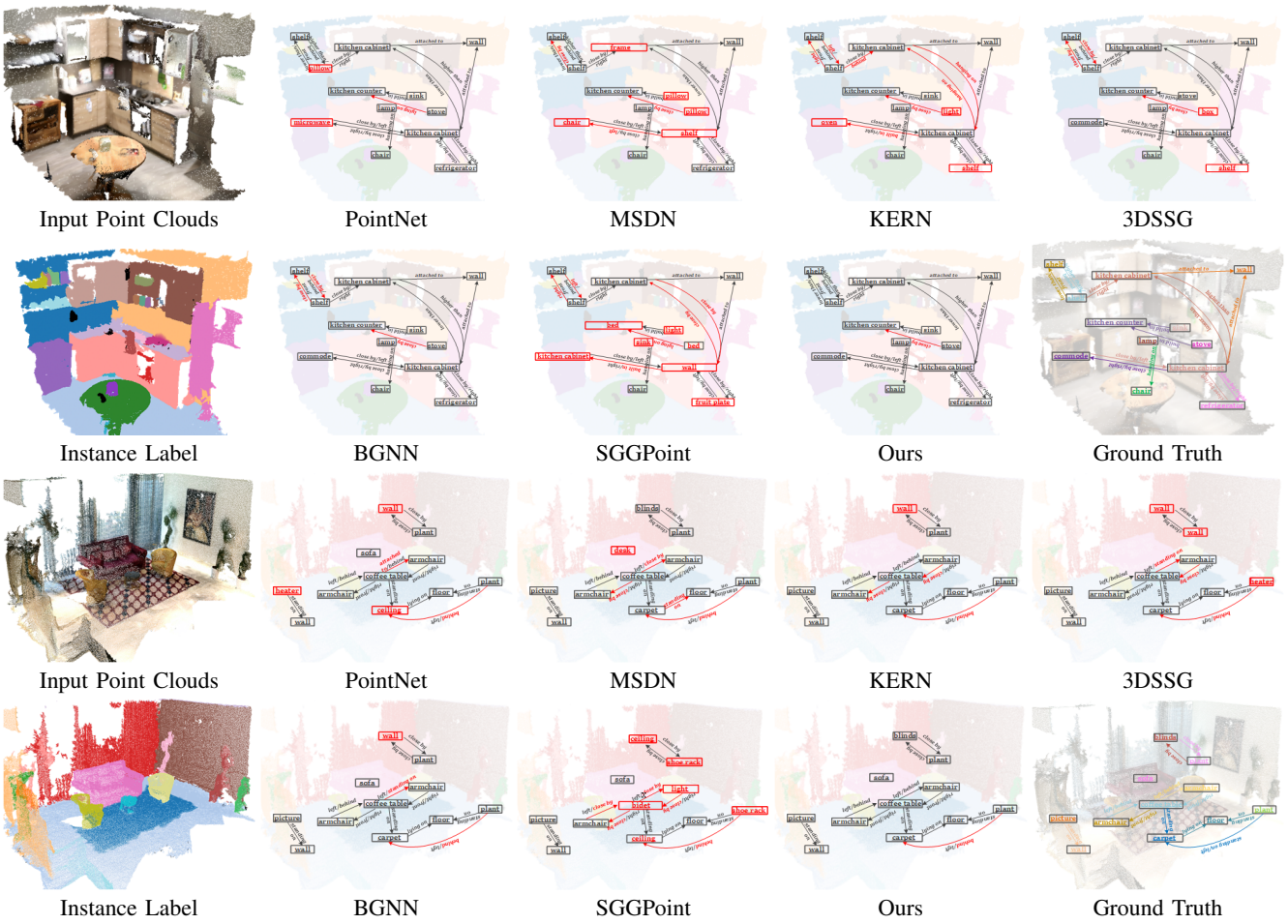

Fig. 2. **Comparisons against the state-of-the-arts.** We only show the results for a local area of the scene. We use gray boxes to indicate the entities and underlines to indicate the relations. The different colors of entity names and relation labels are the same as the colors of the class-agnostic instance labels, except that red indicates an incorrect prediction result.

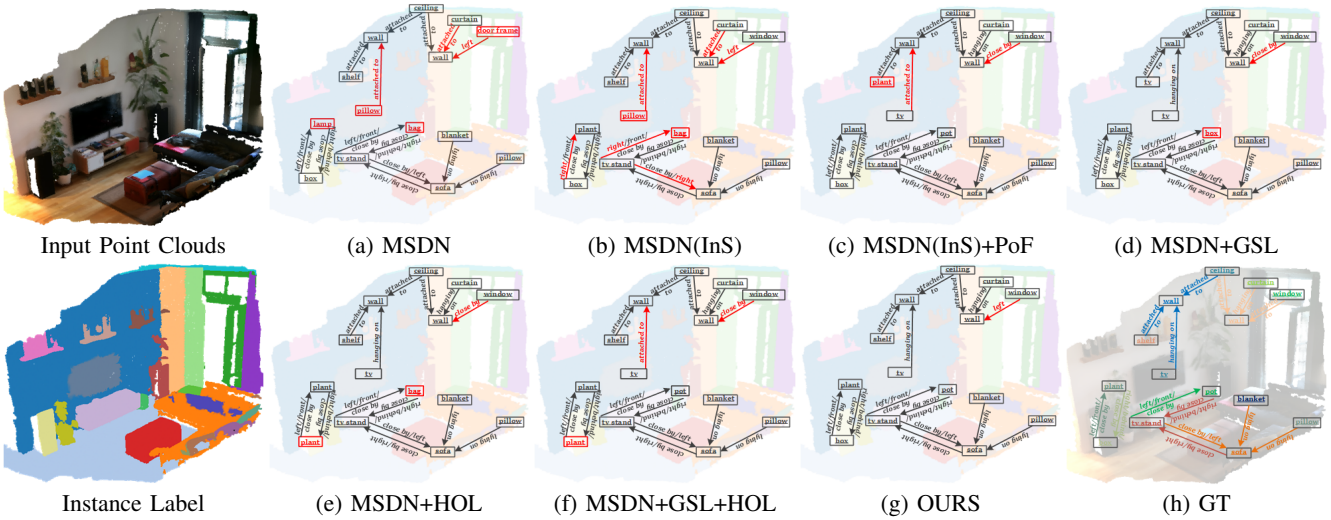

Fig. 3. **Visualization results for ablation study.** For clarity, we only show the results for a locally representative area of the scene. The rest of the annotation is the same as described above.

Input Point Clouds Ground Truth Ours

Fig. 4. **Results of small object predictions.** The desktop and windowsill have a large number of small objects placed in clutter, including `plants`, `books`, `folder` and `pack`. And the point clouds of objects such as `pc` and `heater` only contain fragmented point cloud inputs.

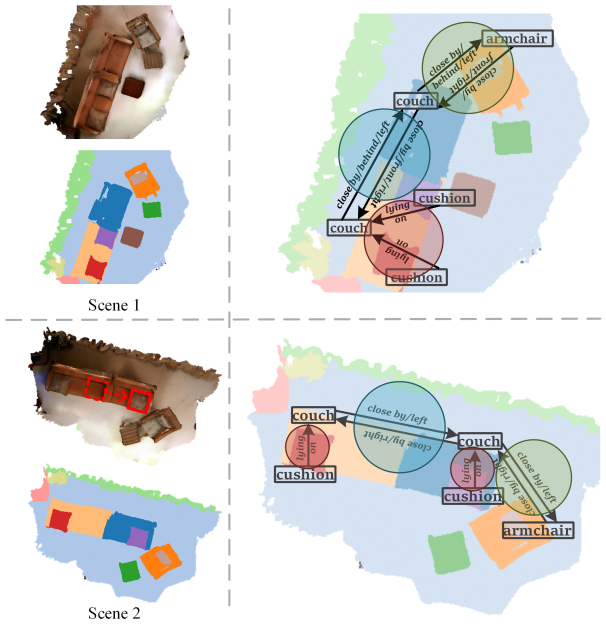

Fig. 5. **Scene translation and rotation.** We marked three areas where the relationship changes correspondingly due to scene changes. Two operations, rotation and object translation, occur simultaneously from Scene 1 to Scene 2. Due to the scene rotation, the relationship between the objects in the blue and green areas changes from `front-behind/right-left` to `right-left`. Due to object translation, the relationship entity in the red area is changed.

In main manuscript, we showed the SGG results for three categories of indoor scenes, kitchen, cafe, and study room. Here we add two more scenes, as shown in Figure 2. Our method has a higher accuracy in predicting both objects and relationships compared to the other methods. We also present a detailed ablation study of our proposed architecture in the main manuscript, where we show additional qualitative examples in Figure 3.

We additionally show the prediction of our method for objects in scenes with a large number of complex small objects in Figure 4. Our method successfully predicts the labels of most of the objects, even if they only have partial point cloud inputs, such as `pc`, `plant`, and `folder`.

Also, we test the effect of scene changes (translation and rotation) on the result of scene graph generation in Figure 5 and observe that these changes mainly affect the labels of relationship. Scene rotation causes a shift in the observation perspective of the scene graph, which results in a different perception of the corresponding positions between objects, and therefore generates different relationship labels. And the substantial position translation of objects may lead to the change of interaction objects. At the same time, we also find that the effects of small-scale scene rotation and object translation on the scene graph are not obvious, which means scene graph cannot perceive subtle environmental changes.

II. Coarse-to-fine Object Class Mapping

The 3RSCAN dataset provides multiple granularity object labels, including NYU40 [5], RIO27 [2] and RIO7 [2]. Our approach selects NYU40 as the coarse-grained object label and we show the mapping between the fine-grained object labels and the coarse-grained NYU40 object labels in Fig 6.

References

[1] 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] Johanna Wald, Armen Avetisyan, Nassir Navab, Federico Tombari, and Matthias Nießner. Rio: 3d object instance re-localization in changing indoor environments. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7658–7667, 2019. 1, 3

[3] 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

[4] Shoulong Zhang, Aimin Hao, Hong Qin, et al. Knowledge-inspired 3d scene graph prediction in point cloud. *Advances in Neural Information Processing Systems*, 34:18620–18632, 2021. 1

[5] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *Proceedings of the European Conference on Computer Vision*, pages 746–760, 2012. 3, 4

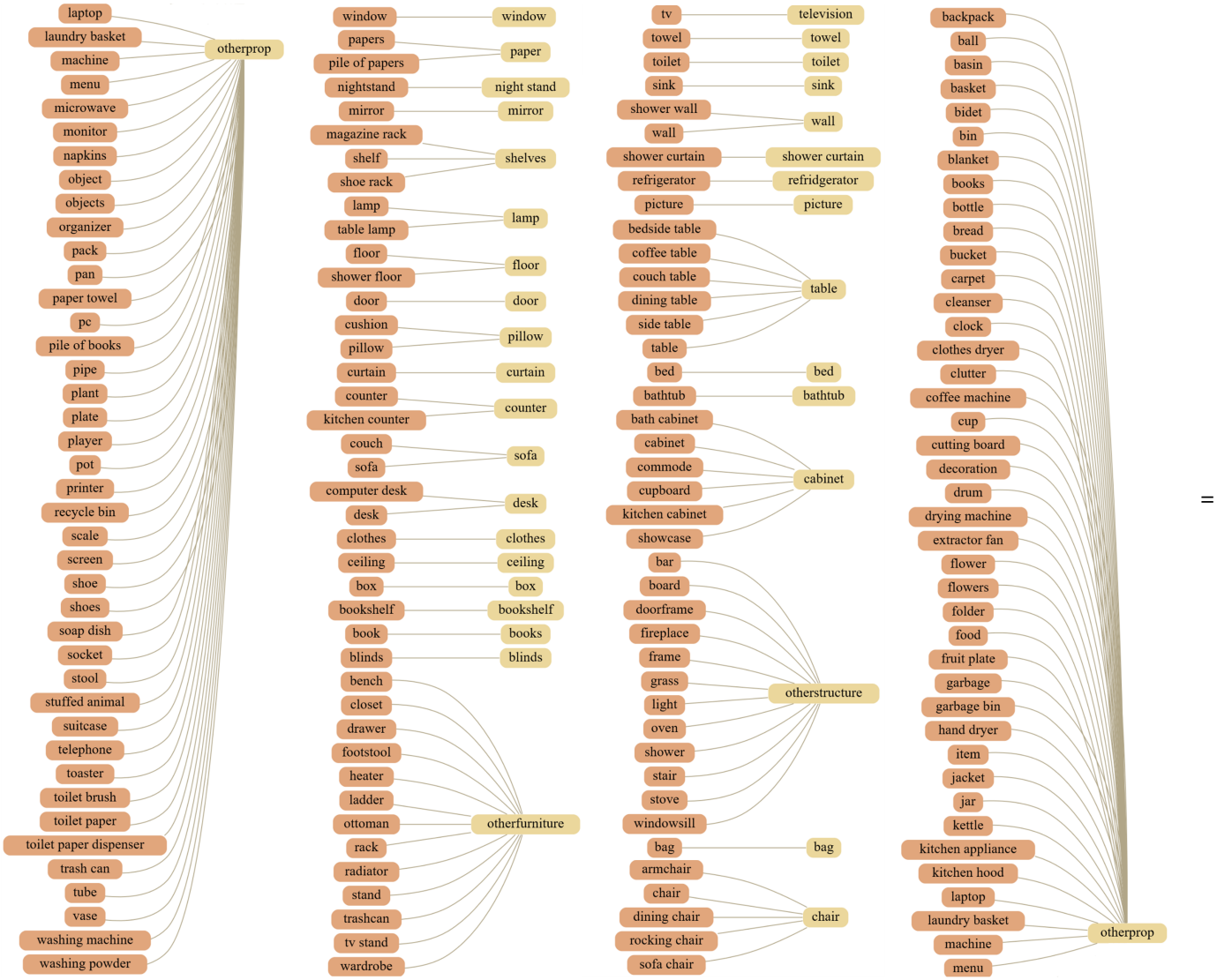

Fig. 6. **Mapping the 160 fine-grained object categories to 40 coarse-grained NYU40 [5] object categories.** The word with orange color indicates the fine-grained class categories, the word with yellow color indicates the coarse-grained NYU40 object categories.