METHOD FOR GENERATING INDOOR 3D SCENE GRAPHS BASED ON INSTANCE FEATURES AND RELATIONSHIP ENCODING

Han Du[1,2,3], Benhe Cai[1,2,3], Xiaoming Li[1,2,3], Weixi Wang[1,2,3], Shengjun Tang[1,2,3] *

[1] Research Institute for Smart Cities, School of Architecture and Urban Planning, Shenzhen University, Shenzhen, P.R. China
[2] State Key Laboratory of Subtropical Building and Urban Science, P.R. China
[3] Guangdong–Hong Kong-Macau Joint Laboratory for Smart Cities, P.R. China

KEY WORDS: Scene Understanding, Deep Learning, Point Cloud, 3D Scene Graph

ABSTRACT:

A 3D scene graph is a compact and explicit representation in scene analysis. In today’s 3D scene graph prediction methods, the feature encoding method of nodes and edges is relatively simple, which essentially hinders the network from fully learning 3D point cloud features. In this paper, we propose a 3D scene graph task framework that fully expresses node and edge features, trying to meet the requirements of fully utilizing point cloud features to achieve high-precision prediction. Experimental results show that with the help of the new representation method, the prediction performance of 3D scene graphs has been significantly improved.

1. INTRODUCTION

Scene understanding is the most essential task in computer vision. It imitates the human visual system to perceive clues in complex scenes to understand the structure and relationships of the scene(Fei-Fei et al., 2004). In recent years, with the rapid progress of digitization of the three-dimensional world, 3D scene understanding has attracted more and more research interests. Traditional 3D scene understanding focuses on the semantic and geometric properties of scene objects, including 3D target detection and recognition, instance segmentation, semantic segmentation, and shape prediction and classification. However, the semantics of contextual connections between associated objects and relationships between objects have not been much explored. Scene graph(Johnson et al., 2015) is an abstract representation that stores scene semantics, where graph nodes are scene entities and their connections are meaningful relationships between them. In the three-dimensional field, scene graphs are gradually becoming popular(Armeni et al., 2019)(Kim et al., 2019). As a representation method for advanced scene understanding, the three-dimensional scene graph abstracts objects and relationships into nodes and edges, and can effectively represent information such as structure, semantics, and contextual connections in the scene. Therefore, 3D scene graphs can be widely used in tasks such as exploration, interaction, generation and operation of 3D scenes, such as autonomous driving, robot navigation and path planning(Kim et al., 2019), AR/VR(Tahara et al., 2020), scene generation and modification(Dhamo et al., 2021), etc.

Although the research on 2D scene graphs has made great progress, the research on 3D scene graphs has become popular in recent years. Armeni et al.(Armeni et al., 2019) expressed the relationship between scenes according to the levels of buildings, rooms, objects, and cameras. Rosinol et al.(Rosinol et al., 2021) consider the real-time representation of agents and incorporates dynamics into hierarchical scene graphs. Based on the 3RScan dataset, Wald et al. extended it to a newly established benchmark for the point cloud-based indoor environment 3D scene graph dataset 3DSSG(Wald et al., 2020). Graph neural networks are widely used in reasoning of 2D scene graphs(Yang et al., 2018). In 3D data, Wald et al.(Wald et al., 2020), Wu et al.(Wu et al., 2021) use graph neural networks to predict three-dimensional semantic scene graphs from class-agnostic instance segmentation scenes. On the basis of point cloud data, Zhang et al.(Zhang et al., 2021a) designed the EdgeGCN module to bridge nodes and edges to perform inference efficiently. Zhang et al.(Zhang et al., 2021b) used prior knowledge to improve the accuracy of relationship prediction. Wang et al.(Wang et al., 2023) used images and language texts to conduct multimodal auxiliary training to further improve the scene graph prediction performance. Besides, some advanced techniques(Bae et al., 2022) for 3D scene graph which combined with SLAM or robotics. Wu et al.(Wu et al., 2021) proposed a method to incrementally generate 3D scene graph from RGB-D images. For this, they designed an attention-based graph processing mechanism that combines incrementally incoming 3D scene graph and recognizing relations not detected in previous steps. Similarly, Hughes et al.(Hughes et al., 2022) presented a real-time 3D scene graph generation method using top-down loop closure detection with a hierarchical descriptor that captures statistics across scene graph layers for optimizing the entire 3D scene graph.

Among the many 3D scene graph generation methods mentioned above, the forms of data sources are different. In armeni’s research(Armeni et al., 2019), in order to generate hierarchical results, the data sources are 3D mesh and RGB panoramic images, etc. In Kimera(Rosinol et al., 2021), in order to complete the construction of the scene graph during the slam process, rosinol et al. used RGB-D and IMU data as input data to the agent. Similarly, in the work of Hughes et al.(Hughes et al., 2022), the data source was also RGB-D and IMU data obtained by the agent in real time. In addition, the work of other researchers is mainly based on point cloud, and the common research paradigm is feature extraction of objects and relationships respectively. At the object representation level, PointNet(Qi et al., 2017) and DGCNN(Wang et al., 2019) are the main backbones. In terms of relationship representation, feature extraction is based on connecting object pairs in a relation(Wald et al., 2019)(Liu et al., 2022). The process is similar to object

* Corresponding author

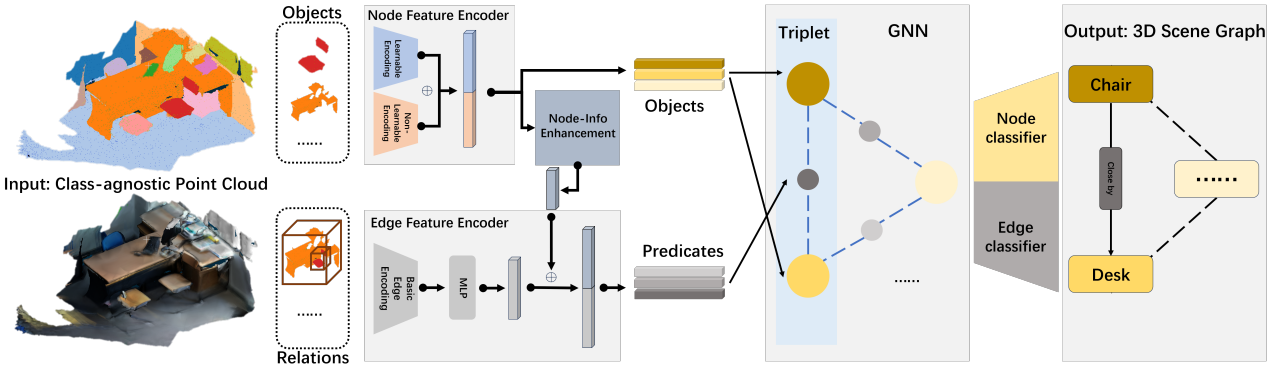

Figure 1. Overview of our proposed framework.

representation. And there are also those that directly use the attribute differences between objects as feature initialization(Wu et al., 2021)(Wang et al., 2023).

Although progress has been made in message passing between nodes of scene graphs and multi-modal auxiliary training, there is still little in-depth research on the encoding of objects and relationships in relationships triplets. Therefore, this paper mainly focuses on the encoder module of objects and relationships, and the impact of new encoding methods on scene graph prediction results was explored.

In the three-dimensional scene graph prediction method, the feature encoding method of nodes and edges is relatively simple. Usually, a single point cloud learning method is used in the network, which cannot fully represent the object point cloud features, which essentially hinders the network from fully learning the 3D point cloud features. In this paper, we propose a 3D scene graph prediction framework that fully expresses node and edge features, trying to meet the requirements of using point cloud features to achieve high-precision prediction. The main contributions of this article are as follows: 1) Using the advanced non-learnable point cloud feature extraction module Point-NN(Zhang et al., 2023) and combine it with the traditional learnable network PointNet(Qi et al., 2017) as the feature extraction module. Point-NN(Zhang et al., 2023) can effectively extract the small positional geometric features of objects and complement PointNet(Qi et al., 2017) to fully characterize objects. 2) Propose a Node-Info Enhancement module, which aims to express the difference and overall characteristics of the two nodes of the relationship in the triplet, and improve the accuracy and learnability of relationship encoding. Experimental results show that with the help of the new proposed method, the prediction performance of 3D scene graphs is improved.

2. METHODOLOGY

2.1 Overview

The main goal of this work is to generate a scene graph from an indoor scene point cloud. Given a point cloud $\boldsymbol{P}$, class-agnostic objects with $\boldsymbol{N}$ instance segmentations and semantic labels $\boldsymbol{M}$ for each object. Our goal is to predict the semantic labels of each object and the relationships between them, that is, the 3D semantic scene graph $\boldsymbol{G} = \{\boldsymbol{O}, \boldsymbol{R}\}$. The o_i in the object set $\boldsymbol{O}$ are named object instances specified by the semantic label $\boldsymbol{M}$. R_{ij} in the relation set $\boldsymbol{R}$ describes the predicate in the relation triplet $<$*subject, predicate, object* $>$, o_i represents the subject node, and o_j represents the object node.

As depicted in Figure 1, the input of the framework proposed in this paper is a scanned class-agnostic scene point cloud. Node encoders and edge encoders are used for feature encoding respectively. The fully connected graph of scene nodes and the processed features is trained by the GCN network, and then two classifiers complete the label prediction of nodes and edges to realize the construction of the scene graph.

2.2 Node Feature Encoder

The node encoder of previous three-dimensional scene graph work often uses learnable backbone networks to extract independent object and relationship-specific features, such as Pointnet(Qi et al., 2017), DGCNN(Wang et al., 2019), but such feature extraction is relatively insufficient. Therefore, we employ a learnable network Pointnet(Qi et al., 2017) $f_L(\cdot)$ and a non-learnable network Point-NN(Zhang et al., 2023) $f_N(\cdot)$ to perform feature extraction respectively as the initialization of object features. The PointNet(Qi et al., 2017) network is simple and efficient, and can be used well as an encoder tool for point cloud representation to capture the global characteristics of the entire point cloud. In order to better represent the local features of point clouds and maintain the simplicity of the overall network, we use a non-learnable network Point-NN(Zhang et al., 2023) without training. Point-NN(Zhang et al., 2023) provides a clue for non-parametric methods to understand 3D point clouds. Point-NN(Zhang et al., 2023) can capture complementary geometric knowledge to enhance existing methods. In addition, we concatenate the spatial invariant properties(Wu et al., 2021), and the node features o_i are calculated as follows:

$$o_i = cat[f_L(P_i), f_N(P_i), \sigma_i, ln(b_i), ln(v_i), ln(l_i)], \quad (1)$$

where P_i denotes point cloud of object instance i, σ_i denotes standard deviation of the position of points, the size $b = (b_x, b_y, b_z)$, the volumn $v = b_x b_y b_z$, and the maximum side length $l = max(b_x, b_y, b_z)$ of the bounding box.

2.3 Edge Feature Encoder

We first follow the same practice as SGFN(Wu et al., 2021) to encode the edge feature r_{ij}, which basically calculates the attribute difference between two objects in a relation triplet, and then edge features are encoded by projecting the concatenated differences of these attributes between two instances, via multi-layer perceptron (MLP) layers. The attributes of the object are similar to those in the Node feature Encoder:

$$e_{ij} = MLP(cat[\bar{P}_i - \bar{P}_j, \sigma_i - \sigma_j, b_i - b_j, ln(\frac{l_i}{l_j}), ln(\frac{v_i}{v_j})]), \quad (2)$$

Method	Object R@1	Predicate R@1	Relationship R@1
SGFN	50.94	78.46	25.71
SGFN with Node-info Enhancement	51.34	79.10	26.58
SGFN with Point-NN	51.52	79.15	26.63
Ours	51.12	79.55	25.75

Table 1. Evaluation of the scene graph prediction performance on 3DSSG with 160 objects and 26 predicate classes.

where e_{ij} denotes initial edge features, $\bar{P}_i$ denotes centroid of object instance i.

Node-Info Enhancement. Although the above expression calculates the difference between object attributes, it is still relatively simple to distinguish relationships. Therefore, we propose the node info enhancement module, which aims to express the difference between the two nodes and the overall characteristics of the relationship in the triplet, and improve the relationship encoding accuracy and learnability. In previous edge feature encoding, there are usually two methods. One is to extract features from the point clouds of the two objects in the relationship pair at the same time, and obtain the features of an object pair as the initialization vector of the edge(Wald et al., 2020)(Liu et al., 2022); the other is to calculate the attribute difference of the two objects in the object pair, and obtain the initialization vector(Wu et al., 2021)(Wang et al., 2023). In this article, in order to obtain richer object pair features and keep the network simple, after using the above relationship calculation method, we are inspired by the first method and consider using node features directly without repeated object pair feature extraction. Therefore, The node difference features and node features between objects are integrated into the relationship features. The specific method to calculate edge feature r_{ij} is shown in the following expression:

$$r_{ij} = cat[e_{ij}, o_i - o_j, cat[o_i, o_j]] \tag{3}$$

2.4 GNN

In this paper, we employ the same GNN structure as in SGFN(Wu et al., 2021) which propagate the features using a GNN with 2 message passing layers to enhance the features by enclosing the neighborhood information after the initial feature embedding on nodes and edges. Compared with the fixed weight of node feature update in GCN(Kipf and Welling, 2016), SGFN(Wu et al., 2021) which utilizes a Feature-wise Attention (FAT) module to pass messages between nodes and edges, and to learn the weight of nodes during update stage. Finally, the node category and edge category will be predicted by two MLP classifiers. We use the same loss functions as in SGPN(Wald et al., 2020).

3. EXPERIMENT

3.1 Setups and Implementation Details

Datasets. We use the 3DSSG dataset to conduct experiments. 3DSSG is a large scale 3D dataset that extends 3RScan(Wald et al., 2019) with semantic scene graph annotations, containing relationships, attributes and class hierarchies. In particular, it provides 1482 scene graphs containing 48k object nodes and 544k edges. A ground-truth semantic scene graph is defined by a set of tuples between nodes and edges, where nodes represent specific 3D object instances in a 3D scan. Nodes are defined by their semantics, a hierarchy of classes, and a set of properties that describe the visual and physical appearance of an object instance and its affordances. Edges are semantic relationships (predicates) between nodes(Fanfan et al., 2022). In experiments, we use the same method for data preparation and training/validation split as in 3DSSG, which includes 160 object categories and 26 relationship categories.

Metrics and Tasks. We follow the same evaluation scheme in SGPN(Wald et al., 2020) to separately report relationship, object, and predicate prediction accuracy with a top-n(Lu et al., 2016) evaluation metric.

Implementation Details. All experiments are carried out on PyTorch platform equipped with one NVIDIA A100 GPU card. The network parameter setting are similar to SGFN(Wu et al., 2021) framework. We train the network for 200 epochs, and the base learning rate is set as 0.001. *Nobj* = 160 and *Nrel* = 26 in our experiments. GNN modules are repeated for $T = 2$ times

3.2 Results of Scene Graph Prediction

Quantitative Results. As shown in Table 1, we compare our method with the current state-of-the-art approach proposed by Wu et al(Wu et al., 2021). The SGFN(Wu et al., 2021) result in the table is based on our reproduction model and is used as our baseline here. Judging from the experimental results, our method is better than the SGFN(Wu et al., 2021) method in classifying objects, predicates and relationships. It is particularly important to note that in this table we use the Recall@*top n* evaluation metric, but we only use the data of Recall@1 for comparison. This is because we believe that this can intuitively and most strictly reflect the level of the model's reasoning ability.

Qualitative Results. As shown in Fig 1, we selected three different scenes (*living room*, *study room*, *cafe*) to verify the reliability of the proposed method in predicting entities and contextual connections in different environments. In complex scenes, our prediction accuracy of entity semantics has improved. For relationships, our proposed method makes more accurate predictions of different relationship structures in stereo(*hanging on*, *supported by*, *standing on*) and plane scenes(*left*, *right*).

Ablation Study. We conducted ablation studies on the two proposed innovations. As shown in Table 1, SGFN(Wu et al., 2021) with Node-info Enhancement refers to the experimental results obtained by adding the Node-info Enhancement module to use the two objects in the relationship triplet to enhance the relationship encoding based on SGFN(Wu et al., 2021). The final prediction capabilities are better than the SGFN(Wu et al., 2021) method. In SGFN(Wu et al., 2021) with PointNN(Zhang et al., 2023), we added a non-learnable network module PointNN(Zhang et al., 2023) to SGFN(Wu et al., 2021) to enhance the ability to describe object features from more subtle aspects of the object. The final experimental results also show the ability to predict 3D scene graphs is better than SGFN(Wu et al., 2021) method. However, what is interesting is that the training of Node-info Enhancement and

PointNN(Zhang et al., 2023) may bring bias to the representation of predicates, resulting in a slight decrease in predicate prediction when using the ours method, but this degradation is not severe, which validate the effectiveness of our proposed method.

4. CONCLUSIONS

In this study, we proposed a 3D scene graph prediction method. Based on traditional learnable object point cloud feature extraction, we integrate a non-learnable network for feature enhancement. In the representation of relationships, we express object pair characteristics and object difference characteristics by calculating the characteristics of the two objects in the relationship pair. Experiments show that our proposed method improves the accuracy of reasoning. In the future, we will try to achieve better scene understanding through 3D scene graphs.

5. ACKNOWLEDGEMENTS

This work was supported in part by the Natural Science Foundation of Guangdong Province (Project No. 2024A1515030061), the State Key Laboratory of Subtropical Building and Urban Science (Project No. 2023ZB18), the Research Program of Shenzhen Science and Technology Innovation Committee grant (Project No. JCYJ20210324093012033), National Key Research and Development Program of China (Project No. 2022YFB3903700), the Research Program of Shenzhen Science and Technology Innovation Committee grant (Project No. KJZD20230923115508017).

REFERENCES

Armeni, I., He, Z.-Y., Gwak, J., Zamir, A. R., Fischer, M., Malik, J., Savarese, S., 2019. 3d scene graph: A structure for unified semantics, 3d space, and camera. *Proceedings of the IEEE/CVF international conference on computer vision*, 5664–5673.

Bae, J., Shin, D., Ko, K., Lee, J., Kim, U.-H., 2022. A survey on 3d scene graphs: Definition, generation and application. *International Conference on Robot Intelligence Technology and Applications*, Springer, 136–147.

Dhamo, H., Manhardt, F., Navab, N., Tombari, F., 2021. Graph-to-3d: End-to-end generation and manipulation of 3d scenes using scene graphs. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 16352–16361.

Fanfan, W., Feihu, Y., Weimin, S., Zhong, Z., 2022. 3D scene graph prediction from point clouds. *Virtual Reality & Intelligent Hardware*, 4(1), 76–88.

Fei-Fei, L., Koch, C., Iyer, A., Perona, P., 2004. What do we see when we glance at a scene? *Journal of Vision*, 4(8), 863–863.

Hughes, N., Chang, Y., Carlone, L., 2022. Hydra: A real-time spatial perception system for 3D scene graph construction and optimization. *arXiv preprint arXiv:2201.13360*.

Johnson, J., Krishna, R., Stark, M., Li, L.-J., Shamma, D., Bernstein, M., Fei-Fei, L., 2015. Image retrieval using scene graphs. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3668–3678.

Kim, U.-H., Park, J.-M., Song, T.-J., Kim, J.-H., 2019. 3-D scene graph: A sparse and semantic representation of physical environments for intelligent agents. *IEEE transactions on cybernetics*, 50(12), 4921–4933.

Kipf, T. N., Welling, M., 2016. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*.

Liu, Y., Long, C., Zhang, Z., Liu, B., Zhang, Q., Yin, B., Yang, X., 2022. Explore contextual information for 3d scene graph generation. *IEEE Transactions on Visualization and Computer Graphics*.

Lu, C., Krishna, R., Bernstein, M., Fei-Fei, L., 2016. Visual relationship detection with language priors. *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*, Springer, 852–869.

Qi, C. R., Su, H., Mo, K., Guibas, L. J., 2017. Pointnet: Deep learning on point sets for 3d classification and segmentation. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 652–660.

Rosinol, A., Violette, A., Abate, M., Hughes, N., Chang, Y., Shi, J., Gupta, A., Carlone, L., 2021. Kimera: From SLAM to spatial perception with 3D dynamic scene graphs. *The International Journal of Robotics Research*, 40(12-14), 1510–1546.

Tahara, T., Seno, T., Narita, G., Ishikawa, T., 2020. Retargetable ar: Context-aware augmented reality in indoor scenes based on 3d scene graph. *2020 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct)*, IEEE, 249–255.

Wald, J., Avetisyan, A., Navab, N., Tombari, F., Nießner, M., 2019. Rio: 3d object instance re-localization in changing indoor environments. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 7658–7667.

Wald, J., Dhamo, H., Navab, N., Tombari, F., 2020. Learning 3d semantic scene graphs from 3d indoor reconstructions. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3961–3970.

Wang, Y., Sun, Y., Liu, Z., Sarma, S. E., Bronstein, M. M., Solomon, J. M., 2019. Dynamic graph cnn for learning on point clouds. *ACM Transactions on Graphics (tog)*, 38(5), 1–12.

Wang, Z., Cheng, B., Zhao, L., Xu, D., Tang, Y., Sheng, L., 2023. Vl-sat: Visual-linguistic semantics assisted training for 3d semantic scene graph prediction in point cloud. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 21560–21569.

Wu, S.-C., Wald, J., Tateno, K., Navab, N., Tombari, F., 2021. Scenegraphfusion: Incremental 3d scene graph prediction from rgb-d sequences. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7515–7525.

Yang, J., Lu, J., Lee, S., Batra, D., Parikh, D., 2018. Graph r-cnn for scene graph generation. *Proceedings of the European conference on computer vision (ECCV)*, 670–685.

Zhang, C., Yu, J., Song, Y., Cai, W., 2021a. Exploiting edge-oriented reasoning for 3d point-based scene graph analysis. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9705–9715.

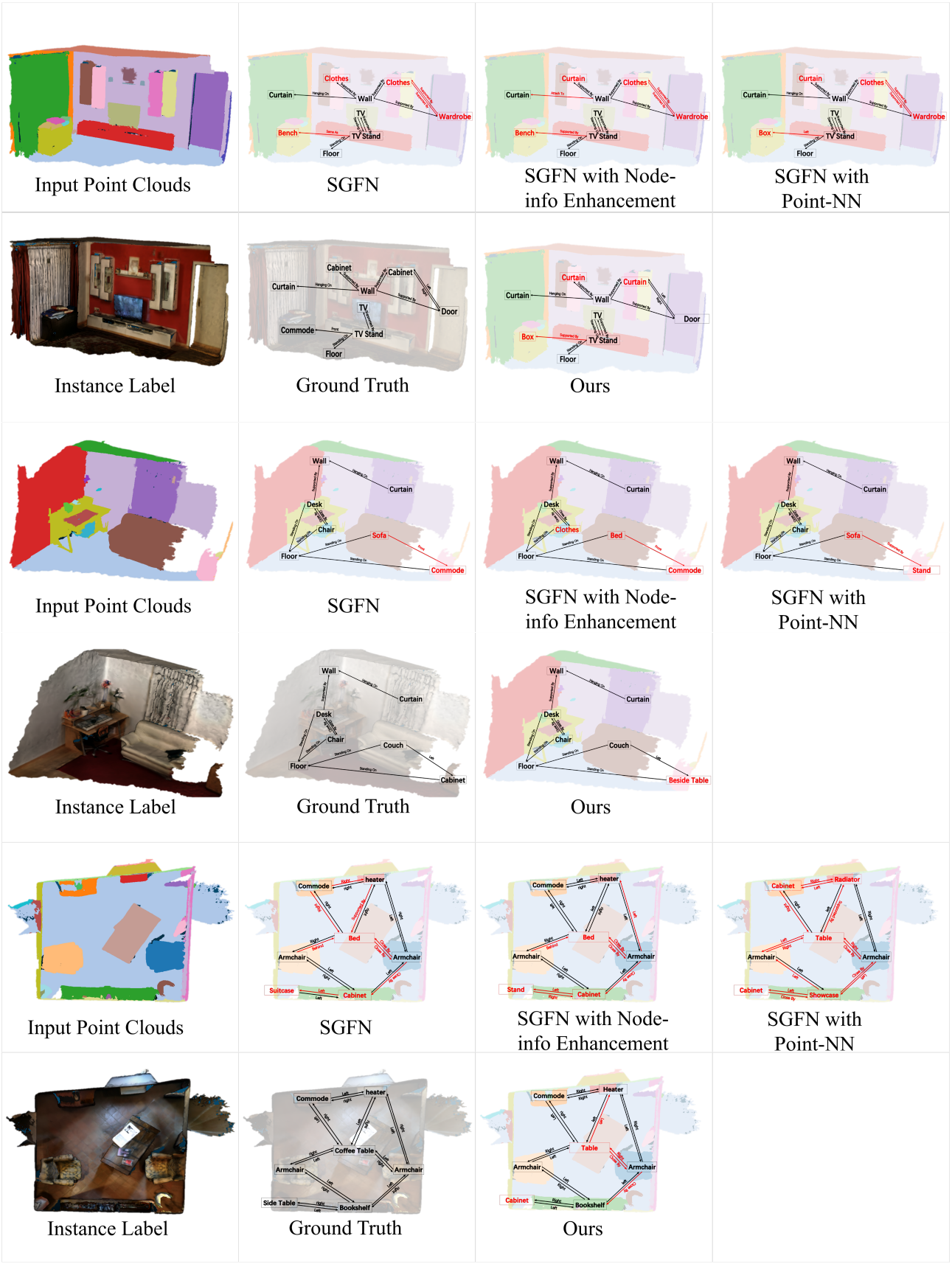

Figure 2. Visualization results. We only show the results for a local area of the scene. We use black boxes to indicate the entities and arrows to indicate the relations. Black indicates correct result and red indicates an incorrect prediction result.

Zhang, R., Wang, L., Wang, Y., Gao, P., Li, H., Shi, J., 2023. Parameter is not all you need: Starting from non-parametric networks for 3d point cloud analysis. *arXiv preprint arXiv:2303.08134*.

Zhang, S., Hao, A., Qin, H. et al., 2021b. Knowledge-inspired 3d scene graph prediction in point cloud. *Advances in Neural Information Processing Systems*, 34, 18620–18632.