3DCNN-DQN-RNN: A Deep Reinforcement Learning Framework for Semantic Parsing of Large-scale 3D Point Clouds: Supplementary Materials

(a) original 3D models (b) point clouds

Figure 1: 3D models and their point clouds. (a) 3D models of tables and its neighboring chairs. (b) the corresponding point cloud data.

A. Overview

The supplementary material provides a demo video and an additional quantitative test example to the main paper.

In Sec B, we provide a demo video to illustrate the process that an eye window localizes and detects the category table in a room of the Stanford 3D semantic parsing data set [1] (Stanford dataset). In Sec C, we give the classification result of the SUNCG point cloud data obtained by our method, and analyze the experimental result. The training dataset is composed of the Stanford dataset and 20% of the SUNCG point cloud data. The other SUNCG point cloud data is taken as the testing data. The SUNCG point cloud data, which is derived from the SUNCG dataset [2], has been described in Section 4 of the main paper. The code we use for converting SUNCG dataset into point cloud data is accessible on github (https://github.com/CKchaos/scn2pointcloud_tool). Fig. 1(a) illustrates 3D models of tables and its neighboring chairs. Fig. 1(b) illustrates their corresponding point cloud data.

To validate the performance of our method, we also obtain the classification result using the method of Armeni et al. [1] on the same training data and testing data as ours.

B. A Demo Video of the Detection Process

The attached video shows the process of an eye window detect the category table in a room of the Stanford dataset. There are mainly two types of tables in the scene. As illustrated in the video, it is noted that the eye window can envelop all of the tables in the scene. We also observe that the DQN has the ability to rapidly localize the approximate locations of the tables, and spend most of time on detecting the boundaries of the tables. In the experiment, we only classify a point into a category if the point is detected by the eye window for more than 5 times. Even the boundaries seem rough in the video, the final classification performance is still high. The eye window has also spent a lot of time around the bookcase since the bookcase resembles a table in its height and structure. We plan to integrate the context information into our method to enhance the detection performance of the eye window.

C. Experimental Results on SUNCG Point Cloud Dataset

C.1. Training data and testing data

To further validate the performance of our method, the training data comes from two different indoor point clouds. We then transfer the trained parameters of our network model to the testing data for performing the classification. Specifically, we first train the 3D CNN on the Stanford dataset. Then, in order to make the 3D CNN adapt to the new environment, based on the original parameters we use 20% of the SUNCG point cloud data to further train the 3D CNN. The training process is the same as it was on the Stanford dataset. It costs 48 hours and the training error converges to approximately 10%. The other 80% of the SUNCG point cloud data are taken as the testing data.

C.2. Comparison between our method and the method of Armeni *et al.* [1]

Table 1 lists the classification results obtained by our method and the method of Armeni *et al.* [1]. From this table, it is noted that our method outperforms the method of Armeni *et al.* [1].

	door	table	bookcase	chair	mean
N1+N2	18.53	**52.31**	**27.01**	**34.98**	**33.21**
S1	**21.84**	29.60	14.14	8.66	18.56

Table 1: Comparison of methods. N1+N2 is our method. S1 refers to the method of Armeni et al. *[1].*

C.3. Analysis of the experimental result

The classification performance of our method on the SUNCG point cloud dataset is not higher than that on the Stanford dataset. The main reasons lie in twofold.

On the one hand, the complexity and irregularity of objects in the SUNCG point cloud dataset. For instance, chairs in the Stanford 3D data set generally have two styles, and thus this similarity of structure among chairs can be easily learned by the 3D CNN. In this way, the 3D CNN produces more accurate scores for the DQN to determine the next action of the eye window. However, in the SUNCG point cloud dataset, the styles of chairs are different in different rooms. We lack enough training data with similar features to train the 3D CNN for producing an accurate score. The same cases happen to other objects. Compared with the Stanford dataset, the spatial relationship among objects in the SUNCG point cloud dataset are more complicated, and the patterns of the objects are also more difficult to be recognized.

On the other hand, the performance of the used computer hardware is limited. In order to enable the proposed method to deal with large-scale 3D point clouds, we utilize the 3D CNN with 6 layers (including convolutional and fully connected layers) to parse the point clouds. The too limited layers of the 3D CNN reduce the generalization ability of the 3D CNN. Due to the limitations of the hardware, fierce down-sampling is applied in our network model. In this situation, the feature representation is not discriminative. If a deeper 3D CNN with less down-sampling is employed to handle point clouds of complex scenes like the SUNCG point cloud dataset, we believe the classification accuracy would be greatly enhanced.

References

[1] I. Armeni, O. Sener, A. R. Zamir, H. Jiang, I. Brilakis, M. Fischer, and S. Savarese. 3d semantic parsing of large-scale indoor spaces. *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,*, 2016. 1

[2] S. Song, F. Yu, A. Zeng, A. X. Chang, M. Savva, and T. Funkhouser. Semantic scene completion from a single depth image. *Proceedings of 30th IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 1