A COMPARATIVE STUDY OF POINT CLOUDS SEMANTIC SEGMENTATION USING THREE DIFFERENT NEURAL NETWORKS ON THE RAILWAY STATION DATASET

Y. A. Lumban-Gaol[1,2], Z. Chen[1,*], M. Smit[1], X.Li[1], M. A. Erbaşu[1], E. Verbree[1], J. Balado[1], M. Meijers[1], N. van der Vaart[3]

[1] Faculty of Architecture and the Built Environment, Delft University of Technology, The Netherlands - (YustisiArdhitasariLumbanGaol, Z.Chen-26, M.Smit-6, X.Li-49, M.A.Erbasu)@student.tudelft.nl, (E.Verbree, J.BaladoFrias, B.M.Meijers)@tudelft.nl
[2] Geospatial Information Agency (BIG), Jl. Raya Jakarta-Bogor Cibinong, Indonesia
[3] Esri Nederland, The Netherlands - Nvandervaart@esri.nl

Commission III, WG III/2

KEY WORDS: Point Clouds, Deep learning, Indoor Scene, Semantic Segmentation, Railway Station

ABSTRACT:

Point cloud data have rich semantic representations and can benefit various applications towards a digital twin. However, they are unordered and anisotropically distributed, thus being unsuitable for a typical Convolutional Neural Networks (CNN) to handle. With the advance of deep learning, several neural networks claim to have solved the point cloud semantic segmentation problem. This paper evaluates three different neural networks for semantic segmentation of point clouds, namely PointNet++, PointCNN and DGCNN. A public indoor scene of the Amersfoort railway station is used as the study area. Unlike the typical indoor scenes and even more from the ubiquitous outdoor ones in currently available datasets, the station consists of objects such as the entrance gates, ticket machines, couches, and garbage cans. For the experiment, we use subsets from the data, remove the noise, evaluate the performance of the selected neural networks. The results indicate an overall accuracy of more than 90% for all the networks but vary in terms of mean class accuracy and mean Intersection over Union (IoU). The misclassification mainly occurs in the classes of couch and garbage can. Several factors that may contribute to the errors are analyzed, such as the quality of the data and the proportion of the number of points per class. The adaptability of the networks is also heavily dependent on the training location: the overall characteristics of the train station make a trained network for one location less suitable for another.

1. INTRODUCTION

Over the last few years, technology has constantly been evolving, and with the constant growth of computational power, ideas from a long time ago have resurfaced to finally show their worth completely. With this constant growth of information, the collection of data has also increased as well as the detail with which this has been captured. This gives a way to the emerging deep learning methods that can learn from the data. Recently, deep learning has been used to address multiple geospatial problems and has proven to be competent (Zhu et al., 2017, Ma et al., 2019, Ardabili et al., 2019).

The training data of deep learning are defined by the application. For indoor and outdoor applications, naturally, this implies the use of different data. This paper strives to test the state-of-the-art deep learning models in an environment that is still somewhat less unexplored, namely the indoor scene. Compared with the outdoor scene, the indoor scene is more complex to parse since it is more costumed and the variety of the indoor features surpasses that of the outdoors (Meijers et al., 2005, Pang et al., 2018). Nevertheless, it does not imply that the indoor scene is inferior to the outdoor scene. On the contrary, the indoor scene is closer to the human habitat, therefore worthy of exploring equally, if not more. We focus on point cloud data which is currently less explored than the traditional image-based machine learning.

Processing unstructured point clouds is non-trivial, and it is only recently that deep learning approaches have been proposed for tackling this task (Qi et al., 2017a, Qi et al., 2017b, Li et al., 2018, Thomas et al., 2019). These point clouds are usually obtained from LiDAR sensors mounted on a vehicle or from visual SLAM approaches; few are collected for the indoor environment. There is a lack of attention to public space where unexploited patterns may exist from the indoor scenes.

In this paper, we investigate how deep neural networks perform within the context of a public indoor environment. Specifically, we evaluate the performance with the point cloud acquired in a railway station. Compared with the existing indoor scenes (Khoshelham et al., 2017, Dai et al., 2017, Armeni et al., 2016), our scene contains more significant noise where moving objects appear. Besides, the point clouds captured by the terrestrial laser scanner exhibit varying density regarding the distance between the objects to the scanner. We extensively evaluate the performance of several deep neural network architectures for semantic segmentation on this data. An advantage that applying deep neural networks for applications, such as asset management, brings is that data does not need possibly hundreds of man-hours to be labelled. This saves a lot of time and also expense each time a scan is made.

2. RELATED WORKS

Recent advance in deep learning has boosted diverse computer vision applications. In the geospatial sector, deep learning-powered solutions contribute to the creation of the digital twin, where automatic object detection and semantic segmentation from point clouds play an important role (Zhu et al., 2017). These applications include urban planning (Urech et al., 2020),

* Corresponding author

asset management (Fang et al., 2016), public safety (Wang et al., 2015), etc.

Deep Learning on Point Clouds. Point clouds are unordered and anisotropically distributed in space. Therefore, unlike the grid data such as images or voxels, point clouds are more difficult to process efficiently in deep neural networks due to this irregularity. Volumetric CNNs (Maturana and Scherer, 2015, Wu et al., 2015, Qi et al., 2016) project the point sets into grids with uniformity. However, this type of methods often involves non-trivial projection and are often constrained by its resolution due to the sparsity of the voxel representations. Recently, there are neural networks proposed that directly consume raw point clouds. PointNet (Qi et al., 2017a) and Deep Sets (Zaheer et al., 2017) both address order invariance of input points using a symmetric function over the inputs. PointNet++ (Qi et al., 2017b) further improves the local feature aggregation by applying PointNet hierarchically over the point set. PointCNN (Li et al., 2018) applies an χ-transformation to learn the weighting of the input features and the point set permutation. Moreover, with graph structures proven to be successful in geometric learning (Battaglia et al., 2018, Zhou et al., 2020), deep neural networks utilizing graph structures are proposed (Landrieu and Simonovsky, 2018, Wang et al., 2019). Within this paper, we evaluate several state of the art solutions for semantic segmentation.

Indoor Point Cloud Application. Indoor scene semantics based on point cloud is essential for many applications, such as planning, localization and navigation services (Flikweert et al., 2019, Quintana et al., 2016). However, indoor environments pose specific challenges for point cloud semantic segmentation due to complex layout, variety of object types and occlusions (Ochmann et al., 2016, Pang et al., 2018). There are indoor point cloud datasets available that target different scenes (Khoshelham et al., 2017, Dai et al., 2017, Armeni et al., 2016). However, none of the existing datasets cover the scene of a railway station. This paper strives to study the indoor scene that is less exploited. Specifically, the lack of benchmark on railway station point cloud semantic segmentation motivates the study of this paper.

3. MATERIALS AND METHOD

3.1 Data

The study uses a LiDAR dataset at the Amersfoort Central Station. It consists of standard information, such as position (XYZ), intensity, and additional Red-Green-Blue (RGB) colour from the camera. An overview of the point cloud dataset from outside and inside the station is shown in Figure 1. The data acquisition was conducted in October 2019 with 19 different scan locations inside the station.

The raw point clouds were unlabeled and still noisy, i.e. moving objects were present. We distil the whole data first to distinguish the interesting assets that are typically inside the station. Based on the screening, we define five classes: clutter, entrance gate, couch, garbage can, and floor.

Furthermore, we subset the data into several partitions based on the planar position X and Y. We did the manual labelling to the specified classes and data cleaning to each partition to remove noise, for example, people sitting on the couch. However, it is impossible to remove the noise completely, so we still have

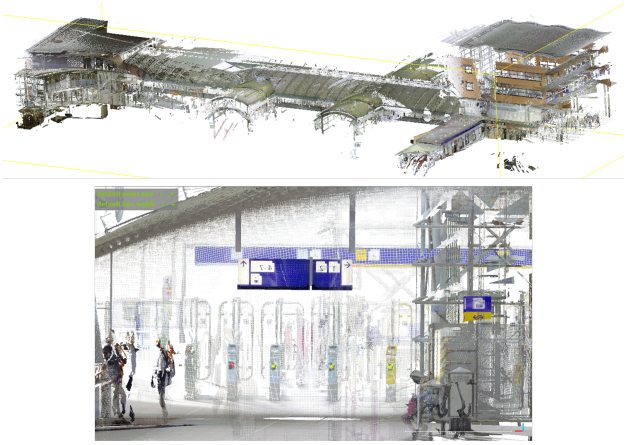

Figure 1. Top: overview of the whole points of Amersfoort station; bottom: sample of point clouds data inside the station.

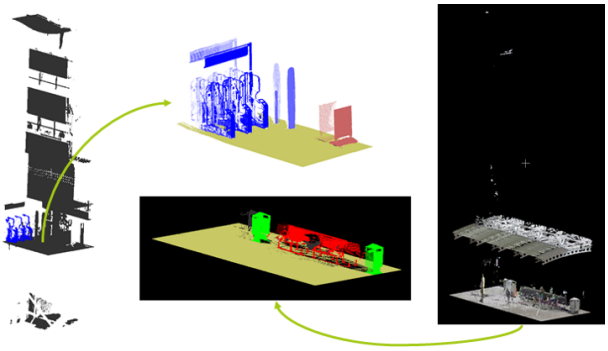

Figure 2. Data partition and manual labelling.

clutter as another class in our classification scheme and end up having the number of points not proportional between classes. Based on our initial implementation through different partition sets, we found that the training did not perform well on the unbalanced data. Meanwhile, a local scene labelled correctly was able to produce plausible results. Thus, the large partitions are not suitable to train because each scene contains many undesirable objects classified as clutter. This paper uses the small data subsets generated from larger partitions with desired objects. The comparison between large and small partitions is illustrated in Figure 2. The final data subsets are shown in Figure 3. We use consistently three scenes to train the networks and the other two for testing.

3.2 Method

This paper uses three different neural network architectures to evaluate on our dataset, namely PointNet++ (Qi et al., 2017a), PointCNN (Li et al., 2018), and DGCNN (Wang et al., 2019). Specifically, we used the PyTorch implementation for PointNet++ (Wijmans, 2018) and DGCNN (Tao, 2020), and the ArcGIS API for PointCNN (Esri, 2021). These networks have been used for semantic segmentation tasks on the private indoor point clouds (Dai et al., 2017, Armeni et al., 2016).

As shown in Figure 4, we perform training of the three networks using the same dataset. All common hyperparameters of the networks are structured as with the S3DIS dataset (Armeni et al., 2016). We adapt the data preparation process to fit the data we have. We use the default setting, except for the block size in PointCNN, which is changed from 1.5 m to 1 m as we consider

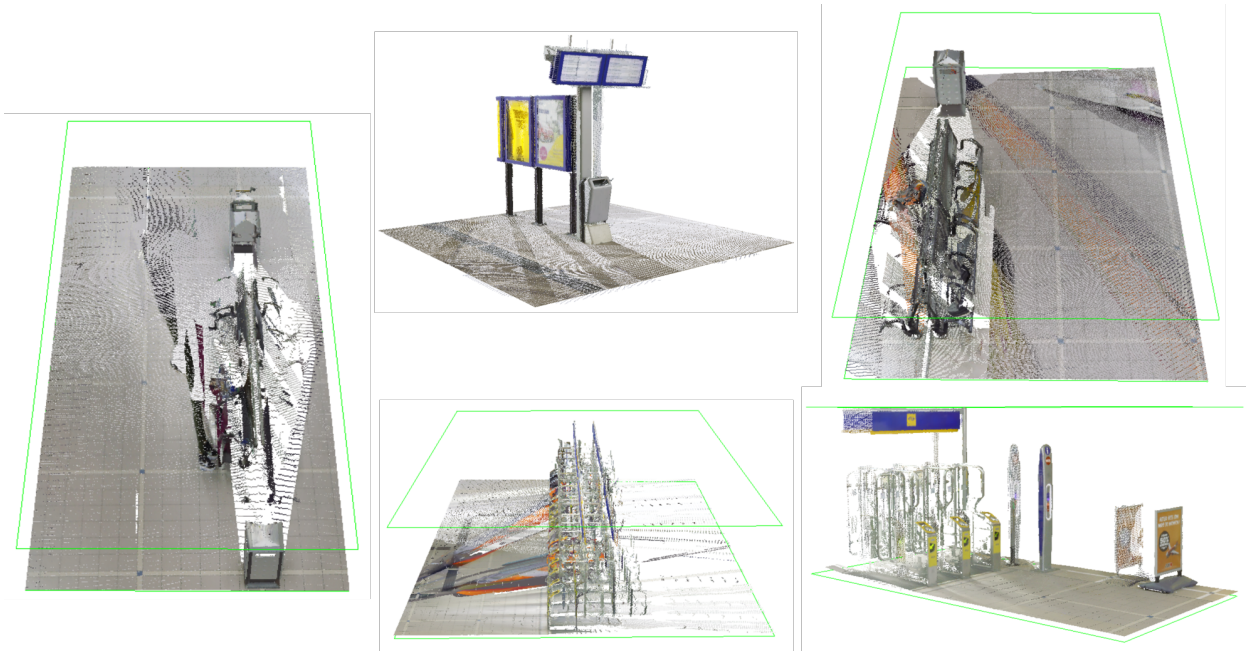

Figure 3. Final data partition to be labelled. Three scenes on the left are used for training, and the rest is for testing.

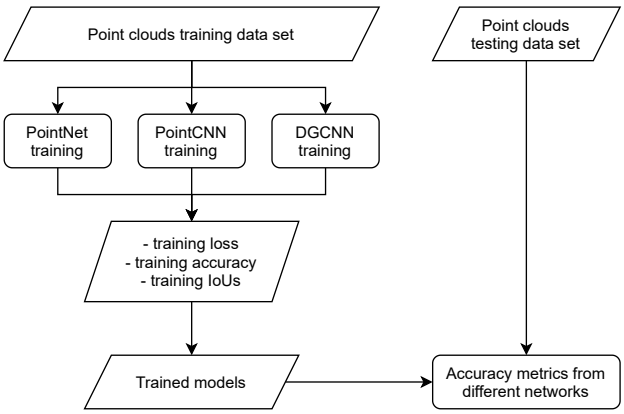

Figure 4. Flowchart of the experiments conducted in this study.

the objects in our scenes to have different dimensions from that of S3DIS.

The training is analyzed by monitoring the loss and accuracy. We stop the training when there is no significant improvement of these metrics. Then, we evaluate each trained model of the networks in the testing stage. Here we use standard approaches to measure the quality of segmentation results by comparing the predicted and ground truth values w.r.t. the overall accuracy and Intersection over Union (IoU).

The overall accuracy describes the ratio between the numbers of points equal to truth values with the total number of points. It is given by

$$OA = \frac{(TP + TN)}{(TP + TN + FP + FN)} \quad (1)$$

where OA is the overall accuracy, TP is the total number of true positive points (e.g. if labelled couch predicted as a couch), TN is the total number of true negatives (e.g. if labelled non-couch predicted as a non-couch), FP is the total number of false positives (e.g. when labelled non-couch predicted as a couch), and FN is the total number of false negatives (when labelled couch predicted as a non-couch). In the overall accuracy, the numerator's sum is equal with the total number of predicted values that are classified correctly for each class, while the sum of the denominator is equal to the total number of ground truth points.

IoU expresses the ratio of the overlapping area and the union area between the predicted and the ground truth:

$$IoU = \frac{Intersection}{Union} \quad (2)$$

4. RESULTS AND DISCUSSIONS

Table 1 presents the evaluation results of each network. All overall accuracy can reach more than 90%. However, the class accuracy varies between 50% and 80%, and a similar observation also appears for the mean IoU. The reason is that some classes have significant low accuracy than others, as indicated in Table 2.

Table 1. Results on different networks.

Networks	Overall accuracy (%)	Mean class accuracy (%)	Mean IoU
PointNet++	95.6	79.9	66.5
PointCNN	92.9	49.9	47.3
DGCNN	94.7	79.9	70.9

Table 2. Class IoUs of each network.

Classes	Class IoU (%)		
	PointNet++	PointCNN	DGCNN
Entrance gate	95.1	71.7	91.5
Couch	57.9	7.5e-5	77.1
Garbage can	75.1	57.7	72.8
Floor	98.3	98.4	99.2

All networks appear to have accurate predictions to the class floor. However, the results suggest that some networks have dif-

ficulty predicting points belonging to the couch or garbage can. In this case, PointCNN almost cannot detect the couch at all. From 106,306 points, only 35 points are predicted as the couch, and 8 of them are classified correctly; most couch points are detected as the garbage can. We suspect that the block size in our PointCNN setting may cause errors in the prediction since 1.0 m block size is relatively small considering the size of a couch. However, PointNet++ and DGCNN, which use the same block size as PointCNN, pinpoint a significantly higher accuracy.

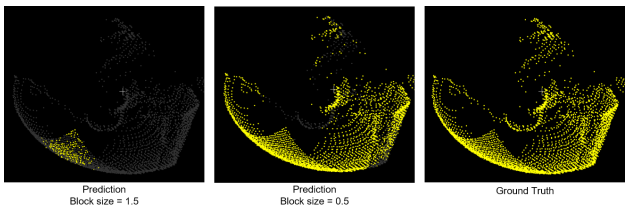

Figure 5. Results of PointCNN on the surveillance cameras using two different block sizes.

To further determine the effect of block size in PointCNN, we clip out one particular type of objects, which are the surveillance cameras inside the station, and evaluate on them using different block sizes. With five cameras as input, the first test result for a block size of 1.5 m has only achieved an accuracy of 30%. The surveillance cameras are considered small, having a dimension of approximately 30 cm x 20 cm. After we reduce the block size to 0.5 m, the accuracy increased significantly to 97%. Figure 5 illustrates the prediction results. This experiment shows a high sensitivity of the block size for PointCNN.

Figure 6 presents the prediction results of each network. We observe that some objects, including clutter, have a similar shape. For example, the board above the entrance gate is similar to the advertisement board on the floor labelled as clutter in the ground truth data. We argue that this affects how the networks learn from the training data and may cause misclassification in the prediction.

Another observation is that the scanner has difficulty capturing the objects completely. For example, the couch's points have holes because it was partly occupied, and we removed the objects above it in the data cleaning process. Moreover, some objects may be obscured by others, so the LiDAR scanner cannot measure them fully, e.g. missing garbage can facades.

5. CONCLUSIONS AND FURTHER WORK

In this paper, a comparison of the semantic segmentation neural networks is addressed for a public indoor point cloud captured in a railway station. Our study scene is different from the existing indoor datasets in terms of the layout, shape and size of the objects, and the presence of moving objects. The results obtained by PointNet++, PointCNN and DGCNN are compared, and some factors that may influence the semantic segmentation performance are analyzed. First, the objects in the station were not completely recorded by the LiDAR scanner, given the difficulty of measuring the public space. Second, noise still exists in the data even after the manual cleaning process. Finally, a similar shape of the different objects, including the unclassified points, may cause misclassification. A caveat to this study is that only a minimal amount of data subset and classes are used. Despite the limitations in the data acquisition, this data represents a point cloud in the real-world indoor public space where several restrictions such as time, budget and administrative effort have to be taken into consideration.

The implementation of our data partitioning to create small scenes as input neglects an enormous number of points from the raw data. With an overall accuracy of more than 90%, though, the pre-trained model may still not be suitable to be used in a larger scene. Moreover, the quality of the data and the proportion of the number of points per class may affect the segmentation performance. The adaptability is also heavily dependent on the training location, for which the railway station's overall characteristics make a trained network for one location less suitable for prediction on another. Further study with more classes and attributes is required to analyze semantic segmentation with public indoor point cloud data comprehensively.

ACKNOWLEDGEMENTS

The authors would like to thank Remco Bunder from Nederlandse Spoorwegen for his support during this research and for using their data of the Amersfoort railway station. On request, the dataset is available for reproducible research. This research also benefits from the Indonesian Endowment Fund for Education (LPDP), Republic of Indonesia. The authors also thank Xunta de Galicia given through human resources grant (ED481B-2019-061).

REFERENCES

Ardabili, S., Mosavi, A., Dehghani, M., Várkonyi-Kóczy, A. R., 2019. Deep learning and machine learning in hydrological processes climate change and earth systems: A systematic review. *International Conference on Global Research and Education*, Springer, 52–62.

Armeni, I., Sener, O., Zamir, A. R., Jiang, H., Brilakis, I., Fischer, M., Savarese, S., 2016. 3D semantic parsing of large-scale indoor spaces. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1534–1543.

Battaglia, P. W., Hamrick, J. B., Bapst, V., Sanchez-Gonzalez, A., Zambaldi, V., Malinowski, M., Tacchetti, A., Raposo, D., Santoro, A., Faulkner, R. et al., 2018. Relational inductive biases, deep learning, and graph networks. *arXiv preprint arXiv:1806.01261*.

Dai, A., Chang, A. X., Savva, M., Halber, M., Funkhouser, T., Nießner, M., 2017. ScanNet: Richly-annotated 3D reconstructions of indoor scenes. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 5828–5839.

Esri, 2021. ArcGIS API for Python: Point Cloud Segmentation using PointCNN.

Fang, Y., Chen, J., Cho, Y., Zhang, P., 2016. A point cloud-vision hybrid approach for 3D location tracking of mobile construction assets. *33rd International Symposium on Automation and Robotics in Construction (ISARC 2016). Proceedings of the International Symposiumon Automation and Robotics in Construction*, 33, 1–7.

Flikweert, P., Peters, R., Díaz-Vilariño, L., Voûte, R., Staats, B., 2019. Automatic extraction of a navigation graph intended for IndoorGML from an indoor point cloud. *ISPRS Annals of Photogrammetry, Remote Sensing & Spatial Information Sciences*, 4.

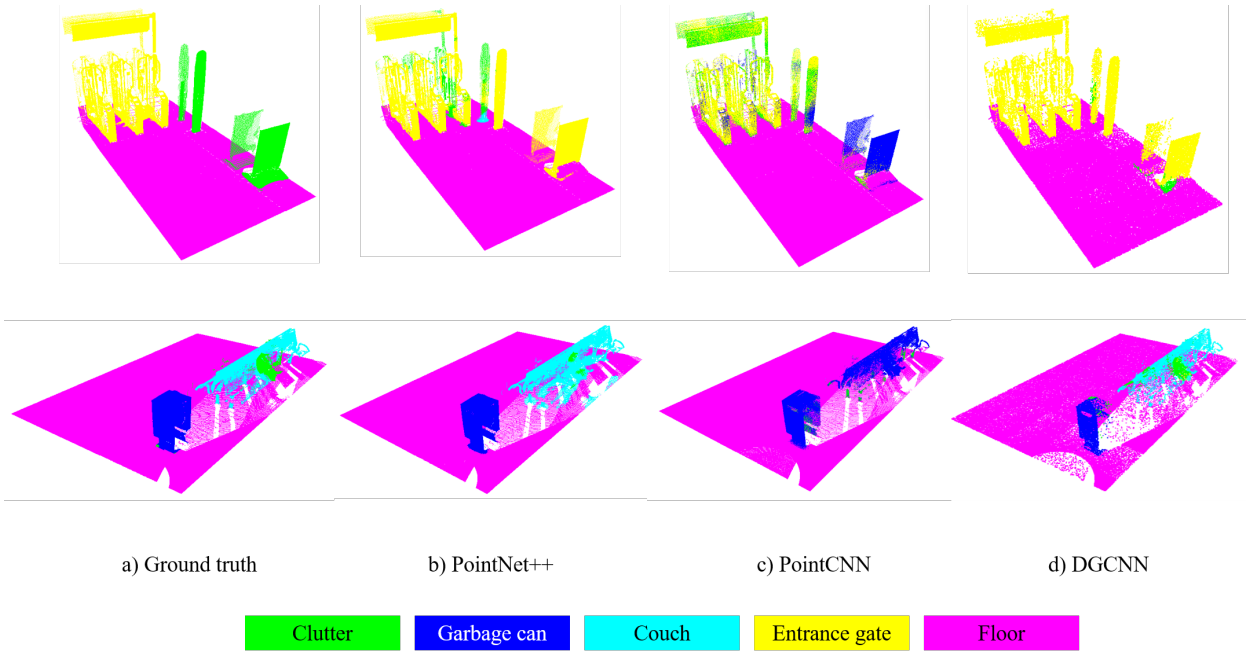

Figure 6. Visualization of each network on the test data.

Khoshelham, K., Vilariño, L. D., Peter, M., Kang, Z., Acharya, D., 2017. The ISPRS benchmark on indoor modelling. *Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci*, 42, 367–372.

Landrieu, L., Simonovsky, M., 2018. Large-scale point cloud semantic segmentation with superpoint graphs. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4558–4567.

Li, Y., Bu, R., Sun, M., Wu, W., Di, X., Chen, B., 2018. PointCNN: Convolution on x-transformed points. *Advances in neural information processing systems*, 31, 820–830.

Ma, L., Liu, Y., Zhang, X., Ye, Y., Yin, G., Johnson, B. A., 2019. Deep learning in remote sensing applications: A meta-analysis and review. *ISPRS journal of photogrammetry and remote sensing*, 152, 166–177.

Maturana, D., Scherer, S., 2015. Voxnet: A 3D convolutional neural network for real-time object recognition. *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, IEEE, 922–928.

Meijers, M., Zlatanova, S., Pfeifer, N., 2005. 3D geo-information indoors: Structuring for evacuation. *Proceedings of next generation 3D city models, Bonn, Germany*, 6, 11–16.

Ochmann, S., Vock, R., Wessel, R., Klein, R., 2016. Automatic reconstruction of parametric building models from indoor point clouds. *Computers & Graphics*, 54, 94–103.

Pang, Y., Zhang, C., Zhou, L., Lin, B., Lv, G., 2018. Extracting indoor space information in complex building environments. *ISPRS International Journal of Geo-Information*, 7(8), 321.

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

Qi, C. R., Su, H., Nießner, M., Dai, A., Yan, M., Guibas, L. J., 2016. Volumetric and multi-view CNNs for object classification on 3D data. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 5648–5656.

Qi, C. R., Yi, L., Su, H., Guibas, L. J., 2017b. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *arXiv preprint arXiv:1706.02413*.

Quintana, B., Prieto, S., Adán, A., Vázquez, A. S., 2016. Semantic scan planning for indoor structural elements of buildings. *Advanced Engineering Informatics*, 30(4), 643–659.

Tao, A., 2020. DGCNN Pytorch. `https://github.com/AnTao97/dgcnn.pytorch`.

Thomas, H., Qi, C. R., Deschaud, J.-E., Marcotegui, B., Goulette, F., Guibas, L. J., 2019. Kpconv: Flexible and deformable convolution for point clouds. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 6411–6420.

Urech, P. R., Dissegna, M. A., Girot, C., Grêt-Regamey, A., 2020. Point cloud modeling as a bridge between landscape design and planning. *Landscape and Urban Planning*, 203, 103903.

Wang, J., Zhang, S., Teizer, J., 2015. Geotechnical and safety protective equipment planning using range point cloud data and rule checking in building information modeling. *Automation in Construction*, 49, 250–261.

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.

Wijmans, E., 2018. Pointnet++ Pytorch. `https://github.com/erikwijmans/Pointnet2_PyTorch`.

Wu, Z., Song, S., Khosla, A., Yu, F., Zhang, L., Tang, X., Xiao, J., 2015. 3D shapenets: A deep representation for volumetric

shapes. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1912–1920.

Zaheer, M., Kottur, S., Ravanbakhsh, S., Poczos, B., Salakhutdinov, R., Smola, A., 2017. Deep sets. *arXiv preprint arXiv:1703.06114*.

Zhou, J., Cui, G., Hu, S., Zhang, Z., Yang, C., Liu, Z., Wang, L., Li, C., Sun, M., 2020. Graph neural networks: A review of methods and applications. *AI Open*, 1, 57–81.

Zhu, X. X., Tuia, D., Mou, L., Xia, G.-S., Zhang, L., Xu, F., Fraundorfer, F., 2017. Deep learning in remote sensing: A comprehensive review and list of resources. *IEEE Geoscience and Remote Sensing Magazine*, 5(4), 8–36.