

Comparison of Deep-Learning Classification Approaches for Indoor Point Clouds

VLADETA STOJANOVIC¹, MATTHIAS TRAPP¹, RICO RICHTER¹ & JÜRGEN DÖLLNER¹

Abstract: In this paper we present an approach for classifying indoor point clouds - particularly areas of clutter, noise, and missing segments that are often featured in point clouds captured using commodity mobile devices with photogrammetric or depth-sensing capabilities. We describe and evaluate two deep-learning approaches for classification of point clouds: using a 2D or 3D Convolutional Neural Network (CNN), for multiview and object-based classification tasks. We present a case study for classification of furniture items using two different CNN architectures (Inception V3 and PointNet++). We report on the classification accuracy and the practicality of using both approaches, and evaluate an experimental approach for generation 3D floorplans from bounding boxes of segmented point clusters. The experimental results of the case study show that the use of PointNet++ CNN offers superior classification accuracy in comparison to the multiview classification, though multiview classification offers a better alternative in terms of computational requirements and lightweight software component integration.

1 Introduction

The built environment around us is constantly changing, and reflecting this change digitally is a key challenge in research and development for Smart Cities and Real Estate applications. Point clouds allow us to capture and digitally represent the built environment. While point clouds can capture the physical surface of a real world object and its location, they lack associated semantics that are often required for further analysis and decision making. With recent advancements in machine learning (particularly deep learning), new methods and approaches have become available for generating semantics for point cloud data. Convolutional Neural Networks (CNN) allow for a given classification model to predict the semantics of point clouds. Approaches for training CNN-based classification models can either make use of point clouds, geometric approximations of point regions (e.g., voxels), 2D images of specific 3D objects and/or real-life photographs of their counterparts. The main focus of this research is a case study that examines and compares the practical aspects of training and using 2D and 3D CNNs for classification of indoor point clouds.

1.1 Problem Statement

Access to training data for indoor point clouds required for training 3D CNNs is usually difficult to acquire, as there is a paucity of freely available datasets. 3D CNNs that make use of unstructured point data as the primary data source for unsupervised learning have versatile applications for semantic classification of the built environment. However, their use is also often restricted by

¹ Hasso Plattner Institute, Faculty of Digital Engineering, University of Potsdam, Prof.-Dr.-Helmert-Straße 2-3, D-14482 Potsdam, Email: [vladeta.stojanovic, matthias.trapp, rico.richter, juergen.doellner]@hpi.de

computational hardware requirements that are usually far greater when compared to those of using 2D CNNs for training and classification. In most cases, the use of GPU-accelerated computation is required, and point cloud data used for training can be very large in terms of memory and storage requirements (often Gigabytes in size). An alternative to 3D CNNs is the use of 2D CNNs trained on image data. In such cases, a multiview classification approach can be used to classify images of spatially-partitioned point clusters, and stream the classification results of the associated point cluster from the model classification result outputs.

1.2 Research Contributions

For this research we have compared two approaches for detecting common office furniture objects: The first using multiview classification with a retrained version of the Inception V3 2D CNN (SZEGEDY et al. 2016), retrained using real-life photographs of common office furniture, and the second approach using PointNet++ 3D CNN (QI et al. 2017), trained on the publicly available Stanford Large-Scale 3D Indoor Spaces Dataset (S3DIS) (ARMENI et al. 2016), using a semantic segmentation approach for detecting furniture objects. We also focused on the practicality of using such approaches on commodity computer hardware, in order to assess the feasibility of such an approach without access to high-end workstation of cluster computation resources. Finally, we present and discuss a practical approach for generating bounding-box and triangulated 3D geometry from segmented point clouds representing the floorplan of a given indoor area, which can be combined with the classified furniture point clusters to create a semantically-enriched indoor representation model.

2 Related Work

Semantic segmentation enables the assignment of points to surface categories (e.g, wall, chair). Furthermore, semantically enriched point clouds can be used as base-data for Building Information Modeling (BIM) and Digital Twin (DT) representations, particularly within the realm of Real Estate 4.0 (STOJANOVIC et al. 2018a). BABACAN et al. (2017) present an approach of semantically enriching indoor point clouds using a combined approach of classification with a 3D CNN and further segmentation of planar point clusters using the RANSAC algorithm. Other recent approaches for semantic segmentation of indoor scenes make combined use of both 2D and 3D CNNs for clustering and semantic labelling of unstructured indoor point clouds (WANG et al. 2019; CHIANG et al. 2019).

The use of 3D CNNs for classification tasks has proven to be a viable option for semantic enrichment of unstructured point clouds. The PointNet++ CNN architecture is able to learn specific features of indoor point clouds in a supervised manner, and apply this model to semantic segmentation of various point cloud models - including point clouds representing the built environment (MALINVERNI et al. 2019), and LIDAR representations of outdoor and urban scenes (WINIWARTER et al. 2019; WOLF et al. 2019). PointNet++ is able to classify unstructured point clouds regardless of their scale, density or orientation, and is invariant to permutations. It uses local distance between neighbouring points projected in Euclidean space to enable direct feature

learning of point cloud data, and recursively sub-samples point regions for evaluating the local features of a point neighbourhood. The training of the network is based on mapping an input point set to a vector representation that is first fed to a multilayer perceptron, then to a max pooling layer. The global feature vector used for generating the output score is obtained from the subsequent local feature vectors of subsampled regions.

However, the use of CNNs such as PointNet++ impose a considerable amount of overhead in terms of hardware resources and computation times. An alternative to using 3D CNNs for classification of point cloud is using 2D CNNs instead, which classify images depicting multiple views of point clusters and associating the classification results of such views back with corresponding point clusters (SU et al. 2015). Subsequent evaluation of this approach, known as *multiview classification*, has shown that it can in certain cases provide just as accurate results as using a 3D CNN, but without the computational overheads and data bandwidth requirements needed for parsing and processing large point cloud datasets (WANG et al. 2017). Another approach is a combination of both 2D and 3D CNNs for classification of 3D data, and variants of such approaches have been described by IOANNIDOU et al. (2017).

Finally, one of the other main benefits of using point clouds is that planar regions representing walls, floors and ceilings can be segmented and reconstructed as BIM data at a given Level-of-Detail (LOD). Since manual generation of BIM data tends to be a laborious and expensive process (VOLK et al. 2014), the use of various segmentation algorithms can automate the process of “scan to BIM” reconstruction (OCHMANN et al. 2019).

3 Approach

We implemented and evaluated two different point cloud classification methods, one using image-based classification with the multiview approach with a 2D CNN (Inception V3), and the other using point cloud-based classification (object-based) using a 3D CNN (PointNet++). Both CNN retraining and classification methods were implemented using Python 3.6, with Tensorflow (for Inception V3) and PyTorch (for PointNet++). Segmentation and generation of bounding boxes of point clusters were implemented as custom command line tools in C++ using the PCL framework (RUSU and COUSINS 2011). Finally, the visualization outputs of the semantically-enriched point clouds were implemented using the WebGL-based Three.js framework in JavaScript and HTML 5, in order to enable accessible and flexible web-based 3D visualization (CABELLO 2010). Both classification approaches were tested using the S3DIS dataset and on a custom point cloud dataset. The training and testing of both CNNs and classification approaches were performed on a commodity laptop with an Intel i5 1.8 GHz CPU, 8 GB RAM, and NVidia GeForce MX150 GPU with 2 GB video memory.

3.1 Custom Dataset Capture and Pre-Processing

The custom dataset used for testing of the semantic segmentation was captured using a Google Tango specification compatible mobile phone, and utilizing Time of Flight (ToF) and depth-

sensing. The capture of the main hallway, conference room, kitchen and three office rooms was completed during day time under natural lighting conditions. Windows and highly-reflective surfaces were not fully captured, and were thus excluded from the scan. The point cloud segments were then manually aligned using the official floorplan as guidance. The initially captured point cloud was further manually edited where noisy and partially scanned clusters without significance were captured. The point cloud originally contained 2 506 858 points, but was sub-sampled to 501 372 points to decrease processing time. Finally, the office rooms, hallway, kitchen and communal area were segmented to form 7 different point cloud scenes for evaluation (Fig. 1).

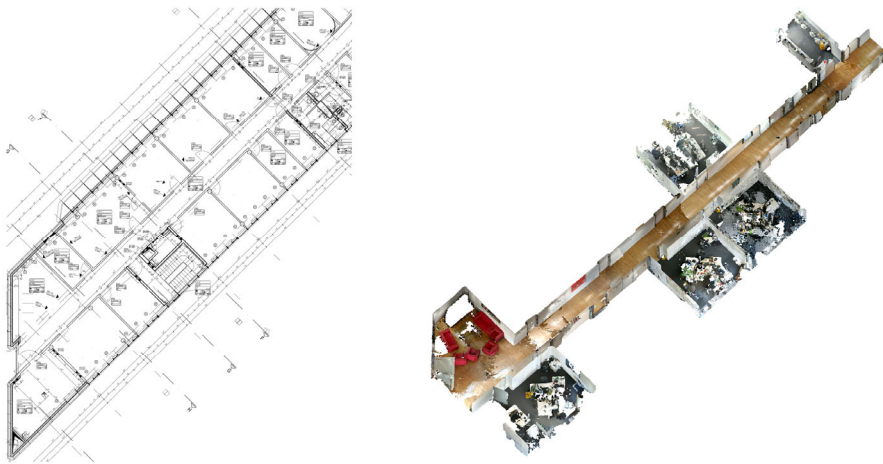


Fig. 1: The floorplan (left) and the captured point cloud of selected areas of the office (right) used to generate the custom dataset for the case study.

3.2 Multiview Classification Implementation Details

The multiview classification is based on the approach described in STOJANOVIC et al. (2019b). This approach uses a pretrained 2D CNN (Inception V3) to classify images obtained from entropy-selected viewpoints of 3D point clusters. The last bottleneck layer of the CNN was pretrained with 4000 epochs, with a learning rate of 0.01 and using 9759 different RGB images of chairs, tables and sofas (with a 70/30 split for training and validation images). Two CNN versions were used, one for detecting only images chairs and tables, and the other for detecting images of sofas as well. The predicted classification accuracy of the CNN version used to classify only chairs and tables (the most common office furniture), was 92.9% while version of the CNN used to classify scenes with both chairs, sofas and tables was 94.5% (with sofas being a less common office furniture item). Prior to classification, the 3D point clouds are clustered, so that point clusters representing cluttered and smaller items can be separated from furniture objects that we want to classify. We used the k -means clustering approach in order to spatially partition the furniture objects represented by point clusters prior to multiview image generation and classification. We then use randomly sampled 3D point normal vectors as an entropy descriptor to calculate the direction and position of the virtual camera placed around furniture object clusters. The viewpoint entropy method then synthesizes 2D images of a given cluster from multiple views with positions and

directions that have highest visual entropy prior to classification. On average, our multiview approach generates approximately 10 images per point cluster, which are then classified by the retrained CNN and their probability-based results are assigned back to each of the corresponding point clusters.

3.3 Object-Based Classification Implementation Details

A PyTorch implementation of the PointNet++ 3D CNN architecture was used in order to assess the feasibility of using 3D CNNs and point clouds for training and classification of indoor office environments. We trained two different versions of the PointNet++ model using the S3DIS dataset (one with and one without additional RGB information of the point data). Both the RGB and non-RGB versions of the PointNet++ CNN model were trained using 112 different scenes from the SI3D dataset, featuring typical indoor offices (with a 81/31 split for training/testing data). The CNN was trained for 200 epochs, with a learning rate of 0.001. We sampled 8192 points for each training cycle. The validation accuracy for the RGB version of the CNN was 64.4%, while for the non-RGB version it was 67.2% (this includes the validation accuracy of the whole scenes, not just furniture objects).

We did not use precomputed point normals as a training feature, since by default the S3DIS set does not contain any normals. A small batch size of 1 was chosen due to computational hardware restraints. The S3DIS already featured labelled data, which was used later on to generate ground truth models to test the accuracy for semantic segmentation of the specific scenes (for the multiview classification approach as well).

3.4 As-is BIM Generation Approach

An object-oriented bounding box (OOBB) representation of point clusters can be extracted and used for generating a 3D floorplan model. These planar clusters are the result of the *Region Growing* (RG) segmentation process, which is suited towards segmenting planar region point clusters that most commonly represent walls, floors and ceilings in indoor point clouds (BASSIER et al. 2017). The 3D floorplan generation involves computing the bounding box for each of the segmented planar point clusters, and exporting them as a text file along with the associated semantic label. We implemented this as a custom command-line tool using the PCL framework. The resulting clusters are exported and used for further editing, analysis, representation and decision making tasks, such as generation of Industry Foundation Classes (IFC) data.

4 Case Study

We presented experimental results for the multiview and object-based classification approaches for common office furniture (Fig. 2), as well as basic reconstruction of segmented core structural features (e.g., walls, floors, and ceilings). For comparing the classification accuracy of the multiview and object-based classification approaches, we empirically evaluated and compared the

semantically segmented point clouds of 14 offices from Area 1 of the S3DIS dataset. We evaluated both RGB and non-RGB based semantic segmentation using PointNet++. The 14 office areas from the S3DIS set used for testing were chosen due to featuring a large amount of visual clutter. They contain three of the furniture object categories used to train the multiview CNN to classify the data (tables, chairs and sofas).



Fig. 2: Examples of point clouds representing common office furniture and indoor locations

Furthermore, we evaluated the classification capabilities of the multiview-based approach for semantic segmentation on our custom point cloud dataset. The custom point cloud dataset features extracted office furniture items from the main office, kitchen and common areas. The accuracy of the two classification approaches for semantic segmentation was obtained by detecting and finding the average of intersection points between the ground truth and predicted point sets - where an intersecting point is defined as having the same location and color as the corresponding point in the ground truth set. Finally, we present experimental results for 3D mesh generation of the floorplan of our custom dataset using OOB-based reconstruction of point clusters segmented using the RG algorithm.

4.1 Classification Results

We present classification results using both the multiview and object-based classification methods, tested using the SI3D dataset and our custom dataset. The results from the multiview-based semantic segmentation show that the multiview-based approach has an accuracy of 52.22% for the custom dataset (Tab. 1), and an average accuracy of 39.56% for the SI3D dataset (Tab. 2). The obtained results from the object-based semantic segmentation of the SI3D dataset show that it has an average classification accuracy of 83.51% for the non-RGB version of the CNN, and an average classification accuracy of 81.14% for the RGB version of the CNN (Tab. 3).

4.2 Experimental 3D Floorplan Generation Results

In addition, we also investigated the use of segmented point clusters of walls, floors and ceilings for the generation of as-is BIM and geometry data. The point clusters were segmented using RG

segmentation, and reconstructed using the OOBBs of the point clusters. Since the OOBBs are generally not correctly aligned, further manual editing was needed in order to create the final 3D floorplan representation (Fig.3). An alternative method for floorplan generation is also possible using horizontally segmented regions of the complete indoor point clouds. Using such an approach, concave shapes obtained from the boundary evaluation of the point cloud in the 2D projected plane is used to generate vectorized contours that can be extruded into 3D polygonal shapes (Stojanovic et al. 2019c).

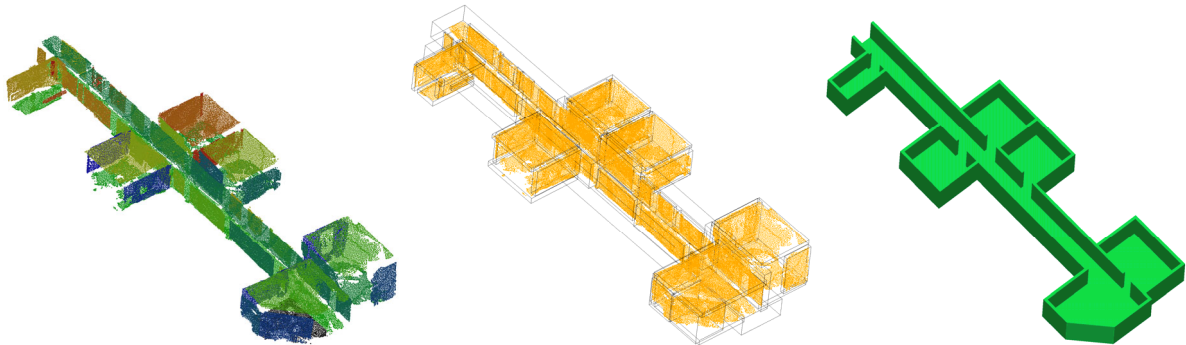
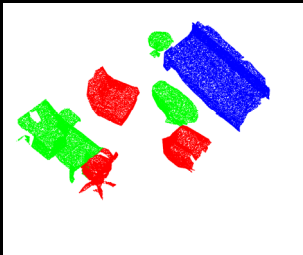
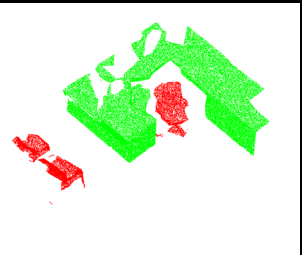


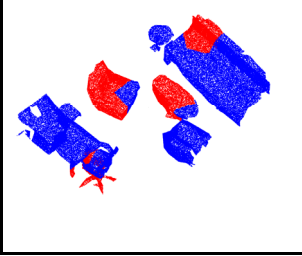
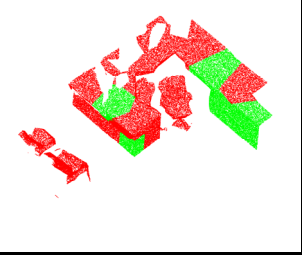
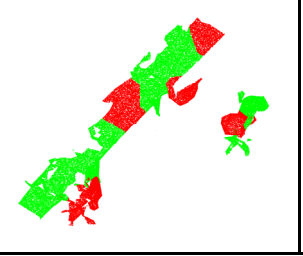



Fig. 3: Example RG-based segmentation of the main hallway and office areas of the custom dataset (left). The resulting OOBBs (middle) can then be exported and used to generate an as-is BIM representation at LOD300 level (right), using various specialist BIM software or software libraries.

Tab. 1: The ground truth (top) and predicted (bottom) results for the custom dataset classified using the multiview-based approach. Red point clusters represent chair objects, blue represent sofas and green represent tables.

Common Area	Kitchen	Office 1	Office 2

Tab. 2: The ground truth (top) and predicted (bottom) results for the SI3D classified the using the multiview-based approach

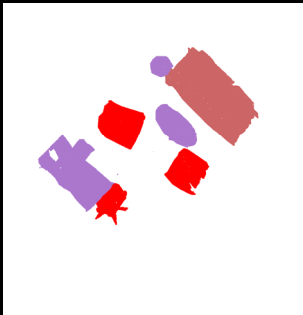

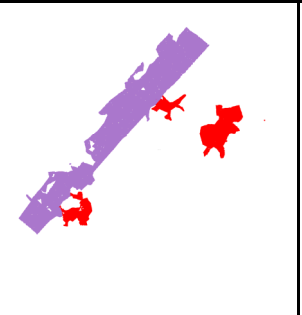
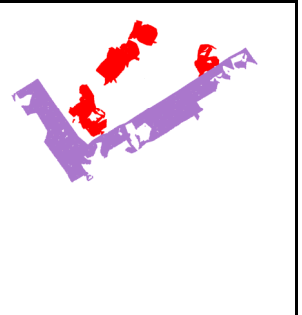
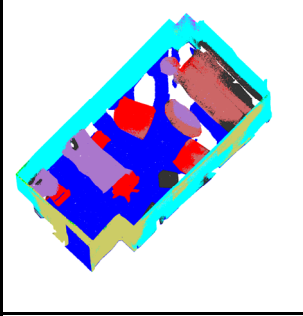
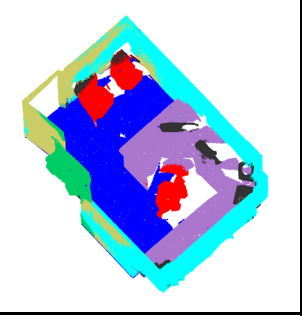
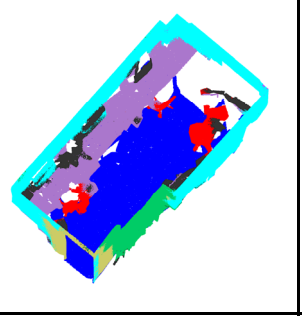
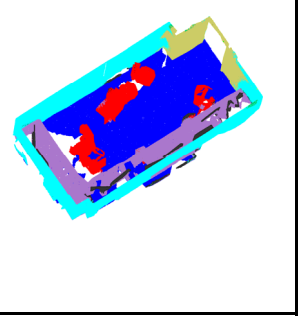
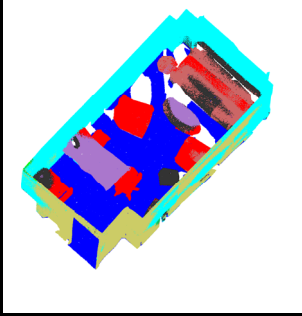
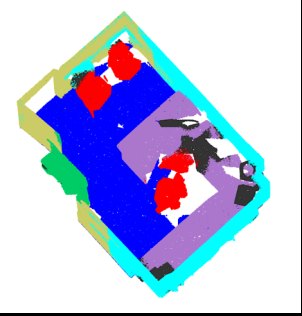
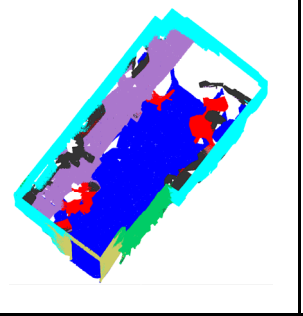
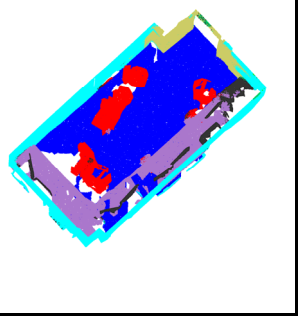
Area 1 Office 2	Area 1 Office 6	Area 1 Office 18	Area 1 Office 24
			
			

5 Discussion

For applications where classification accuracy is more critical, the use of PointNet++ is a better option if high-end computing resources are available. The retraining of the Inception V3 CNN with images of real-life furniture, used for the multiview-based approach, takes only a fraction of the time compared to training the PointNet++ 3D CNN using the S3DIS dataset. The retraining time for Inception V3 is approximately 30 minutes, while the training time for the PointNet++, which had to be trained from scratch, was approximately 15 hours. The semantic segmentation using PointNet++ on each of the offices from the SI3D takes on average approximately 90 minutes, while the multiview-based approach on segmented office clusters from the same dataset takes on average approximately 6 minutes. The main challenge when using PointNet++ for semantic segmentation is that it relies on point cloud data for training, and this is impractical for integration with commodity computing hardware compared to the multiview-based classification approach (as most implementations of PointNet++ are designed to leverage the computing power of parallel multi-GPU architectures). We have also observed that the PointNet++ CNN versions making use of RGB point color features had slightly worse evaluated classification accuracy. While the multiview-based approach offers worse classification accuracy in comparison to the object-based approach, it is generally more adaptable for use on commodity hardware, and retraining the CNN (using only the last bottleneck layer), is more convenient with access to potentially millions of images on the internet depicting common office furniture in different categories (in comparison to having to find and use point clouds of such objects, which is often difficult). One drawback of the multiview-based approach is that the viewpoint entropy method used for generating the multiview images relies on correctly calculated point normals that conform to the surface curvature (in terms of viewing the object from a given distance from its center with 3D perspective). Finally, the

reconstruction of 3D floor plans from RG-segmented point clusters provides favourable experimental results for further use for generation of as-is BIMs, and can be expanded to include reconstruction of higher levels of detail and information.

Tab. 3: The ground truth (top) and predicted non-RGB CNN version (middle), and RGB CNN version (bottom) results for the SI3D classified the using the multiview-based approach. Red point clusters represent chair objects, copper represents sofas and purple represents tables. For these results the semantically segmented walls, floors and ceilings are also included, but were not evaluated as part of the classification accuracy result

Area 1 Office 2	Area 1 Office 6	Area 1 Office 18	Area 1 Office 24
			
			
			

6 Conclusions and Outlook

Our research investigated the practical use of CNNs for semantic enrichment of indoor point clouds. We compared multiview versus object-based classification for indoor point clouds, with a particular focus on semantic segmentation. We also experimented with a basic method for reconstructing 3D floor plans using bounding boxes of point clusters representing walls, floors and ceilings. We have found that the use of the use of object-based classification with PointNet++

offers superior classification accuracy in comparison to the multiview-based classification approach. However, there is a trade-off between practicality and performance versus classification accuracy when deciding which approach to use for deep-learning-based semantic segmentation. Future work will investigate the possible combined use of a 2D CNN with a 3D CNN for classification of indoor point clouds. We predict that a balance in terms of computational performance and classification accuracy can be achieved using a combined 2D and 3D CNN approach, along with the use of semantic reasoning.

7 Acknowledgments

This work has been partially funded by the Research School on Service-Oriented Systems Engineering of the Hasso Plattner Institute, Faculty of Digital Engineering, University of Potsdam, Germany.

8 References

- ARMENI, I., SENER, O., ZAMIR, A. R., JIANG, H., BRILAKIS, I., FISCHER, M. & SAVARESE, S., 2016: 3D Semantic Parsing of Large-Scale Indoor Spaces. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 1534-1543.
- BABACAN, K., CHEN, L. & SOHN, G., 2017: Semantic Segmentation of Indoor Point Clouds Using Convolutional Neural Network. ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences, 101-108.
- BASSIER, M., BONDUÉL, M. VAN GENECHTEN, B., & VERGAUWEN, M., 2017: Segmentation of Large Unstructured Point Clouds Using Octree-Based Region Growing and Conditional Random Fields. The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences **42**(2W8), 25-30.
- CABELLO, R., 2010: Three.js. URL: <https://github.com/mrdoob/three.js>.
- CHIANG, H. Y., LIN, Y. L., LIU, Y. C. & HSU, W. H., 2019: A Unified Point-based Framework for 3D Segmentation. IEEE International Conference on 3D Vision (3DV), 155-163.
- IOANNIDOU, A., CHATZILARI, E., NIKOLOPOULOS, S. & KOMPATSIARIS, I., 2017: Deep Learning Advances in Computer Vision with 3D Data: A Survey. ACM Computing Surveys (CSUR) **50**(2), 20.
- MALINVERNI, E. S., PIERDICCA, R., PAOLANTI, M., MARTINI, M., MORBIDONI, C., MATRONE, F. & LINGUA, A., 2019: Deep Learning for Semantic Segmentation of 3D Point Cloud. ISPRS-International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences **4215**, 735-742.
- OCHMANN, S., VOCK, R. & KLEIN, R., 2019: Automatic Reconstruction of Fully Volumetric 3D Building Models from Oriented Point Clouds. ISPRS journal of photogrammetry and remote sensing, **151**, 251-262.
- QI, C. R., YI, L., SU, H. & GUIBAS, L. J., 2017: Pointnet++: Deep Hierarchical Feature Learning on Point Sets in a Metric Space. In Advances in neural information processing systems, 5099-5108.

- RUSU, R. B. & COUSINS, S., 2011: 3D is here: Point cloud library (pcl). IEEE international conference on robotics and automation, 1-4.
- STOJANOVIC, V., TRAPP, M., RICHTER, R., HAGEDORN, B. & DÖLLNER, J. 2018a: Towards The Generation of Digital Twins for Facility Management Based on 3D point Clouds. In Proceedings of the 34th Annual ARCOM Conference.
- STOJANOVIC, V., TRAPP, M., DÖLLNER, J. & RICHTER, R. 2019b: Classification of Indoor Point Clouds Using Multiviews. In The 24th International Conference on 3D Web Technology, 1-9.
- STOJANOVIC, V., TRAPP, M., RICHTER, R. & DÖLLNER, J. 2019c: Generation of Approximate 2D and 3D Floor Plans from 3D Point Clouds. In Proceedings of the 14th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications 1, 177-184.
- SU, H., MAJI, S., KALOGERAKIS, E. & LEARNED-MILLER, E., 2015: Multi-View Convolutional Neural Networks for 3D Shape Recognition. In Proceedings of the IEEE international conference on computer vision, 945-953.
- SZEGEDY, C., VANHOUCHE, V., IOFFE, S., SHLENS, J. & WOJNA, Z., 2016: Rethinking the Inception Architecture for Computer Vision. In Proceedings of the IEEE conference on computer vision and pattern recognition, 2818-2826.
- VOLK, R., STENGEL, J. & SCHULTMANN, F., 2014: Building Information Modeling (BIM) for Existing Buildings—Literature Review and Future Needs. *Automation in construction* **38**, 109-127.
- WANG, C., PELILLO, M. & SIDDIQI, K., 2019: Dominant Set Clustering and Pooling for Multi-View 3D Object Recognition. arXiv preprint arXiv:1906.01592.
- WANG, W., YU, R., HUANG, Q. & NEUMANN, U., 2018: SGPN: Similarity Group Proposal Network for 3D Point Cloud Instance Segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2569-2578.
- WINIWARTER, L. & MANDLBURGER, G., 2019: Classification of 3D Point Clouds using Deep Neural Networks. *Publikationen der Deutschen Gesellschaft für Photogrammetrie, Fernerkundung und Geoinformation e.V.*, Band **28**, 663-674.
- WOLF, J., RICHTER, R. & DÖLLNER, J., 2019: Techniques for Automated Classification and Segregation of Mobile Mapping 3D Point Clouds. In Proceedings of the 14th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications 1, 201-208.