Attentive Enhanced Convolutional Neural Network for Point Cloud Analysis

Bifang Ma, Yifei Chen

Abstract—**Although methods based on deep learning have achieved remarkable success in the field of point cloud analysis, the recognition accuracy of point cloud analysis is still far from practical applications. In this paper, we propose a novel end-to-end attentive enhanced convolutional neural network for point cloud analysis, named AECNN, which can analyze point clouds with high accuracy. The key component in our method is the attentive enhanced convolution (AEConv) module. It designs an attention mechanism to enhance the features of each group through the interaction of information between groups, so that the feature expression is more sufficient and richer. Based on AEConv, we further design an attentive enhanced convolutional neural network (AECNN) for 3D model classification and segmentation. AECNN continuously abstracts the entire point cloud by stacking downsampling and AEConv to obtain feature descriptors that can represent the entire point cloud, and finally uses a classification head and a segmentation head to complete the task of point cloud classification and segmentation, respectively. Experiments are conducted on the point cloud recognition and large-scale point cloud outdoor segmentation datasets, i.e., ModelNet40 and vKITTI, respectively. Extensive experiments show that our AECNN can achieve high performance for classification and segmentation tasks, which exceeds previous point cloud recognition and segmentation algorithms.**

Index Terms—**point cloud analysis, attentive enhanced, classification, segmentation, convolutional neural network**

I. Introduction

POINT cloud analysis is a key technology and basic means in the field of computer vision, and has a wide range of practical applications. Deep learning-based methods have made significant progress in the field of 2D image processing. However, unlike 2D images, the sparsity and disorder of point clouds make point cloud analysis remain challenging in computer vision. In addition, the accuracy of existing 3D point cloud classification and segmentation methods needs to be further improved.

To improve the recognition accuracy by using the surface detail information of point cloud, the previous multi view-based methods use 2D image segmentation algorithms to process irregular point cloud data. Rendering a group of point clouds into 2D images is conducive to feature learning using classical 2D CNN, and the pixel-level semantic labels are back-projected to the point clouds to obtain the classification and segmentation results. However, the multi view-based methods will inevitably lose some distinctive geometric information, and the perspective selection of projection needs extensive expert knowledge. Voxel-based methods regularize the point cloud into voxel structures, which can preserve the geometric information of objects to a large extent, but cannot subdivide the geometric information of the object boundary. In addition, low resolution tends to lose boundary details easily. Although point cloud-based methods are easy to obtain local geometric information, only part of the geometric information is discriminative for the overall structure of the object, and the absolute position information of points and the relative position information between point pairs lack the ability to describe the high-level global geometric structure of the object. The operation of CNN takes a lot of time to construct local point cloud data, resulting in increased time cost. Recently, deep learning [1] techniques have made progress in the machine vision field for many tasks [2], [3], [4], [5], [6], [7]. In the realm of machine vision, deep learning techniques have recently made progress in a number of tasks. The method of directly processing point cloud data can utilize the original geometric information of the point cloud, avoiding the complex operations to completely obtain all the information of the point cloud. However, the original point cloud has the characteristics of the irregular, sparse and disordered structure, and it is difficult for traditional convolutional neural networks to capture the useful features.

To this end, we postulate an attentive enhanced convolutional neural network to deal with point cloud classification and segmentation tasks, which can improve the performance of point cloud recognition and segmentation while obtaining stronger feature representation capabilities. Our method learns a feature enhancement through information interaction between groups to extract a fuller and richer point cloud representation. The contributions are three-fold:

- We propose a novel attentive enhanced convolution module for point cloud feature extraction. It adaptively enhances the features of each group by information interacting between groups, effectively improving the feature representation capability of convolution operations.
- We propose an attentive enhanced convolutional neural network for 3D model analysis. It extracts discriminative feature descriptors by stacking downsampling and attentive enhanced convolution to improve the accuracy.
- We obtain high accuracy for both point cloud classification and segmentation benchmarks.

II. Related Work

A. Volumetric-based and Projection-based Methods

Literature [8] is early attempts to apply deep learning methods to 3D model recognition and segmentation tasks. First, the raw unstructured point cloud is transformed into voxel data, and then features of the voxel data are extracted

Manuscript received September 01, 2022; revised February 16, 2023.

This paper was supported by the Guiding Project Fund of Science and Technology Department of Fujian Province (No.2020H0027) and Natural Science Foundation of Fujian Province (No.2022J01972).

Bifang Ma, the corresponding author, is an associate professor of School of Electronic and Mechanical Engineering, Fujian Polytechnic Normal University, Fuqing, 350300, P.R.China. (e-mail: mabf555@126.com).

Yifei Chen is the deputy director of the education management Information Center of Fujian Province, Fuzhou, 350003, P.R.China. (e-mail: 13356537@qq.com).

using a 3D convolutional neural network. In [8], the features of the voxel model are extracted by a stochastic gradient descent algorithm and an additional regularization term by adding a voxel model. In literature [9], a multi-view convolutional neural network is designed for the first time. Features are extracted from 2D views of 3D models, and then global feature descriptors are obtained through feature aggregation through a view pooling layer to predict category information. However, these methods may lose useful information during data format conversion and increase time cost and memory overhead.

B. Point Cloud-based Learning Methods

Compared with the multi-view-based method and the volumetric-based method, the point cloud-based method is the most concise, and the deep learning technology can be designed on the 3D point cloud, avoiding the time-consuming and laborious operations such as multi-directional photography and voxelization, which has become the mainstream research method. In [10], PointNet is proposed, which first adopt the multi-layer perceptron (MLP) to express independent points, and then obtains global feature descriptors through the aggregation of max-pooling layers to complete the tasks of 3D model recognition and segmentation. However, PointNet only focuses on each independent point without considering capturing local geometric features. In [11], the local features of the point cloud are extracted in different scales, but the distance measurement between point pairs is not considered. The ability to capture local geometric features is lacking. To sum up, the feature representation capability of the existing point cloud recognition and segmentation methods needs to be further improved, and it is difficult to be compatible with practical applications.

III. Attentive Enhanced Convolutional Neural Network

We propose the attentive enhanced convolution and attentive enhanced convolutional neural network to boost the segmentation accuracy. We first design a attention enhanced convolution better to extract features of point cloud, then build an attention enhanced convolutional neural network by repeatedly using down sampling and attention enhanced convolution to extract feature descriptors that can represent the whole point cloud, and finally use a classification head and a segmentation head to achieve point cloud processing.

A. Group Convolution

The previous three-dimensional model recognition and segmentation methods mainly use multi-layer perceptron and graph convolution as the basic point cloud feature representation modules, and achieve certain phased results. However, the convolution kernel of the multi-layer perceptron has a shallow channel width, and the information that can be encoded is limited, and the feature expression is insufficient. Considering that the group convolution [12] has a wider channel and can encode more information, we introduce the group convolution to extract discriminative information by increasing the convolution kernel width. Specifically, we divide the input point cloud and convolution kernel into g groups, so that the features and convolution parameters are decreased to 1 / g. Compared with the standard multi-layer perceptron and graph convolution, the parameters of the packet can be reduced to the original 1 / g, which effectively reduces the network parameters and improves the calculation efficiency. At the same time, multiple groups increase the channel width so that the network can encode more useful features. Moreover, more discriminative information can be encoded through multiple channels, greatly enhancing the feature representation ability. Since different groups only contain part of the point cloud features, in order to capture complete features, previous methods concatenate features of different groups together to obtain more abundant features, and shuffle the channels to enhance the information interaction between groups, thereby capturing the discriminative features of all groups. In this way, the group convolution can increase the channel width, encode more useful features of the point cloud, and obtain richer feature representation capabilities.

B. Attentive Enhanced Convolution

To further enhance the information interaction between groups and use the information between groups to enhance the useful channel features, we design the attention enhanced convolution to obtain stronger feature representation ability. The network structure is shown in Figure 1. Specifically, we first concatenate the features of different groups together to obtain a connection map, the process is formulated as

$$f^g = f_1^g \oplus f_2^g \oplus f_2^g \oplus, ..., f_g^g, \tag{1}$$

where f is the features of the point cloud, g represents the group operation. And then the connection feature map is input into a convolution (conv) layer to learn discriminative features, and the learned features are input into the activation function sigmoid for nonlinear mapping, and then a feature selector is obtained as an attentive mask ($attM$) for feature enhancement.

$$attM = sigmoid(conv(f^g)) \tag{2}$$

Next, we use the learned attention mask as a feature enhancement module to adaptively enhance the features of each group by an element-wise multiplication operation, and the process is formulated as

$$\widetilde{f_1^g} = attM \odot f_1^g \tag{3}$$

$$\widetilde{f_2^g} = attM \odot f_2^g \tag{4}$$

where $\odot$ denotes an element-wise multiplication operation. Finally, we concatenate the enhanced features of each group together to obtain more discriminative and richer features $\widetilde{f^g}$.

$$\widetilde{f^g} = \widetilde{f_1^g} \oplus \widetilde{f_2^g} \oplus \widetilde{f_2^g} \oplus, ..., \widetilde{f_g^g} \tag{5}$$

In this way, the proposed attentive enhanced convolution can improve the feature representation of point clouds by learning between-group information, extracting discriminative enhanced features that are beneficial for identifying elusive categories.

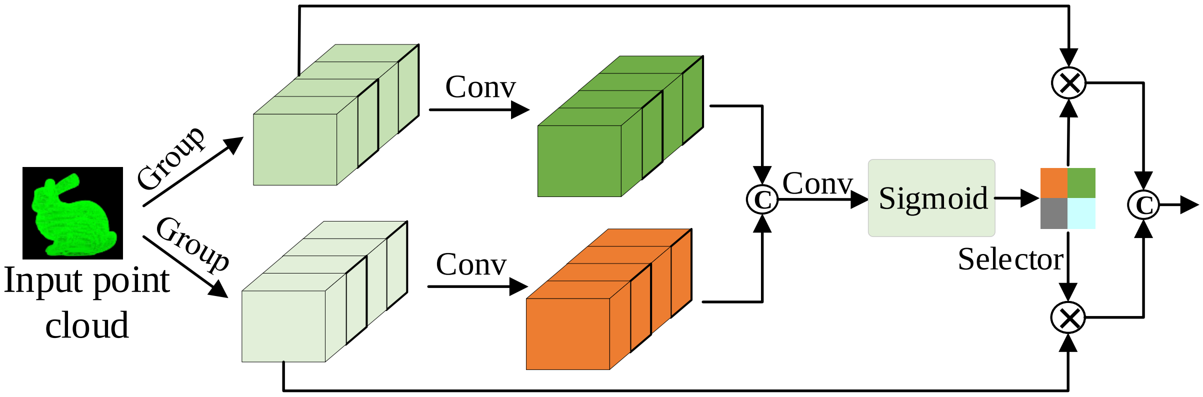

Fig. 1. The structure of attentive channel convolution.

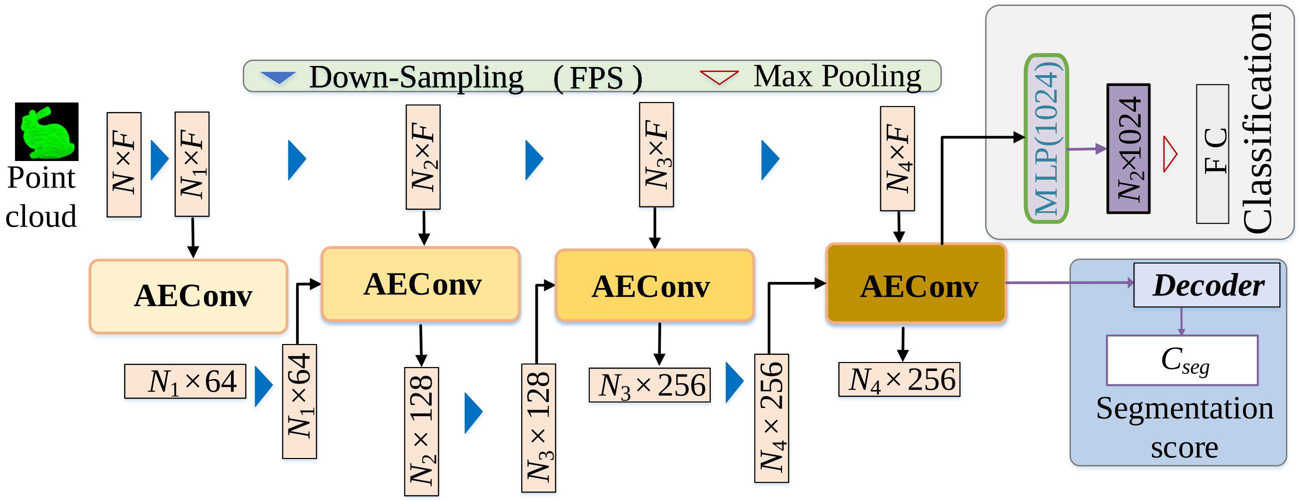

Fig. 2. The framework of our AECNN.

C. Attentive Enhanced Convolutional Neural Network

Based on AEConv as the basic feature extraction operation, we can further complete AECNN for 3D model classification and large-scale point cloud segmentation tasks. The network structure is shown in Figure 2. To significantly reduce computational cost and memory overhead, we employ downsampling operations to progressively decrease the point cloud density. Specifically, we use three downsampling operations to enrich the attentive semantic information, and finally abstract the point cloud into a feature descriptor that can represent the point cloud. Since our attentive enhanced convolution has powerful feature representation capabilities, the sampling points have rich semantic information. By continuously stacking downsampling and attentive enhanced convolution, the original point cloud is getting smaller and smaller but the semantic information is richer and richer, and finally the point cloud is abstracted into a feature descriptors that can represent entire point cloud. Point cloud analysis task are completed by performing the heads of classification and segmentation on this descriptor. For the classification task, we use max pooling to aggregate useful features while discarding useless redundant features. Finally, we adopt three fully connected layers to obtain classification scores to complete the point cloud classification task. For segmentation, the structure of encoder layers follows the classification network, but it has a deeper number of layers so that the features have more abstract semantic information. We downsample four times in total to reduce computation for the segmentation network. Furthermore, we introduce skip connections for feature reuse and feature propagation.

IV. Experiments

A. Datasets

For the point cloud classification task, the standard dataset ModelNet40 [13] is selected for experiments. ModelNet40 has 12,311 point cloud of 40 categories. We use 9,843 models for training and 2,468 models for testing. For segmentation, experiments are carried out on a outdoor scene segmentation dataset vKITTI [14]. vKITTI is an outdoor actual autonomous driving scenes dataset, divided into 6 different urban scenes, in which all points are labeled as cars, and trees in the autonomous driving scene of 13 semantic categories.

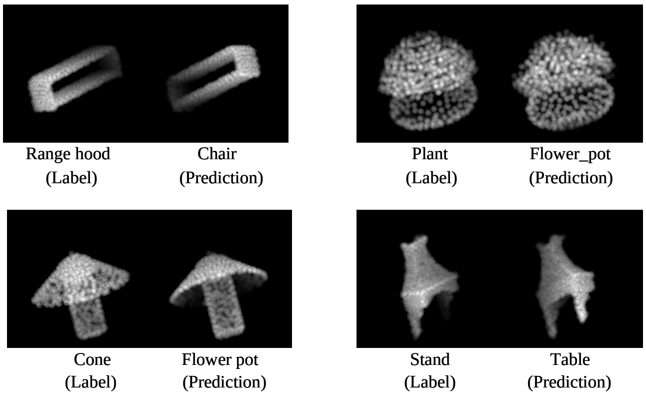

Fig. 3. Misclassification results on ModelNet40.

TABLE I: Results on MODELNET40 (%).

Algorithms	mA	OA
VoxNet [8]	83.0	85.9
PointNet [10]	86.0	89.2
PointNet++ [11]	-	90.7
AECNN (Ours)	91.8	93.2

TABLE II: Quantitative results on VKITTI (%).

Algorithms	OA	mIoU
PointNet [10]	79.7	34.4
AECNN(Ours)	86.1	54.0

B. Implementation Details

Experimental environment: the operating system is Linux, the hardware environment is a RTX 2080Ti (8GB GPU), and the deep learning framework is Tensorflow. The experiments adopt adam's step size for updating stochastic gradient descent (SGD) and initialize network parameters using the xavier optimizer. The network is optimized with momentum-based SGD with a momentum factor of 0.9 and an initial learning rate of 0.001. An activation function Relu is used in the model.

C. Evaluation Metrics

For classification, we adopt overall accuracy (OA) and mean accuracy (mA) to evaluate model classification performance. For the point cloud segmentation task, we use OA and mean IOU (mIoU) to evaluate model segmentation performance.

D. Point Cloud Classification

To verify that our algorithm has a strong feature representation ability for classification, we select the ModelNet40 [13] dataset, and results are represented in Table I. It can be seen that our method achieves a ideal accuracy of 93.2%. Furthermore, compared to the voxel-based method VoxNet [8], our AECNN improves mA and OA by 8.8% and 7.3%, respectively. Additionally, Figure 3 visualizes the misclassification results on the ModelNet40 dataset to qualitatively analyze the classification performance. It can be seen from the cases that there are many similar models in the point cloud classification dataset. Because our algorithm lacks the ability to identify similar point cloud models that are easy to be confused, the similar models are classified incorrectly.

E. Point Cloud Segmentation

To verify the superiority of our algorithm in this paper in processing 3D point cloud segmentation task, the recognition accuracy is compared with other advanced algorithms on the vKITTI [14] dataset. As shown in Table III, the proposed algorithm achieves good segmentation performance with an mIoU of 85.4%. Compared with the current advanced algorithm, our algorithm has certain advantages in segmentation accuracy. The experiment fully verifies that our method has the strong feature recognition ability. Figure 4 visualizes the prediction results of our AECNN. Our AECNN can accurately segment the entire autonomous driving scene, especially has a strong discriminating ability for fine-grained boundaries. The reason is that the proposed AEConv can attentively extract discriminative feature-enhanced information through the information interaction among different groups.

TABLE III: Ablation study on Modelnet40.

Model	Ablation	ModelNet40(OA)
A	Baseline [10]	89.2%
B	Group convolution	91.3%
C	AEConv	92.4%
D	AECNN(Fully Network)	93.2%

F. Ablation Study

We conduct ablation studies on Modelnet40. The results are represent in table III. Here, "AECNN" means attention enhanced convolutional neural network, and "AEConv" means attention enhanced convolution. It can be seen that compared with baseline PointNet [10], the recognition accuracy is improved by 2.1% when we use group convolution. When we adopt AEConv instead of group convolution, the classification accuracy is further improved by 1.1%. The reason is that our AEConv is able to focus on important channel features. In addition, after constructing AECNN by stacking AEConv and downsampling, the recognition accuracy is improved by 0.8%. The reason is that our AECNN has the stronger feature representation ability for point cloud analysis. The above experiments prove the effectiveness of attentive enhanced convolution and attentive enhanced convolutional neural network, and fully prove the effectiveness of the proposed method.

V. Conclusion

In this paper, we propose an attentive enhanced convolutional neural network for point cloud recognition and segmentation tasks. Firstly, we design an attentive enhanced convolution to obtain more discriminative point cloud features by using the information between groups. Besides, an attentive enhanced convolutional neural network is constructed by stacking downsampling and attentive enhanced convolution multiple times to reduce the density of the point cloud while continuously enhancing the semantic information of the features to obtain more discriminative feature descriptors for completing the classification and segmentation

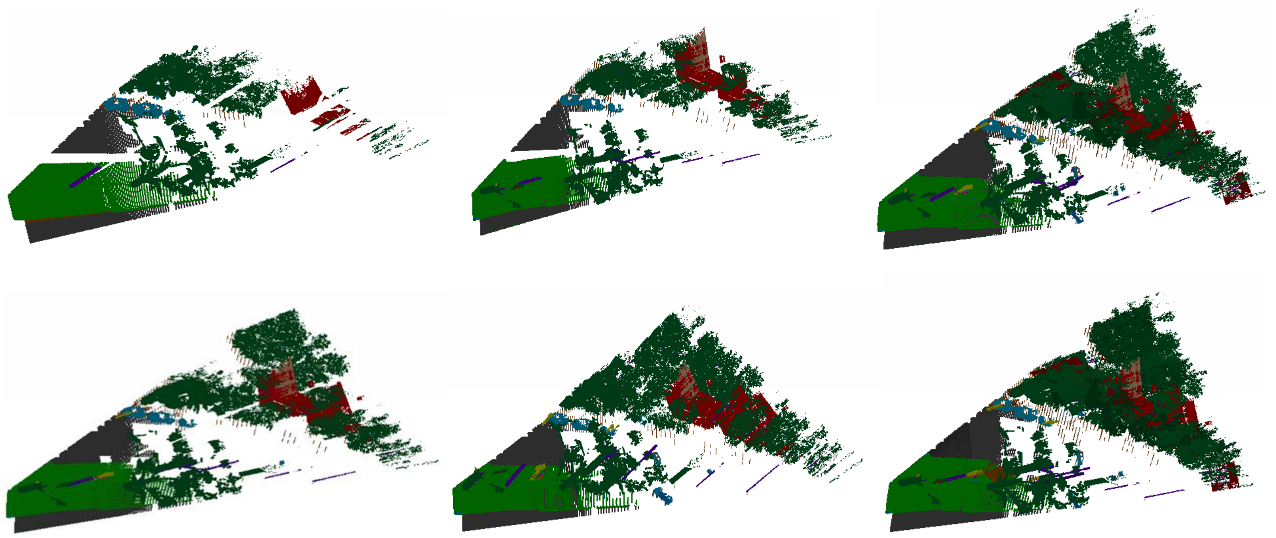

Fig. 4. Qualitative results of CADCNN on outdoor vKITTI.

of point clouds. Experiments prove that our method has better performance in recognition and segmentation due to its strong feature representation for 3D point cloud classification and segmentation tasks.

References

[1] Krizhevsky Alex, Ilya Sutskever and Geoffrey E. Hinton, "Imagenet Classification with Deep Convolutional Neural Networks," *Advances in Neural Information Processing Systems*, vol. 60, no. 6, pp. 84–90, 2017.

[2] Felice, Liccardo and Salvatore, Strano and Mario, TerzoLiccardo, "Real-time Nonlinear Optimal Control of A hydraulic actuator," *Engineering Letters*, vol. 21, no. 4, pp. 241–246, 2013.

[3] Yang, Li and Yang, Yonghong and Li, Yaqin and Zhang, Tianwei, "Almost Periodic Solution for a Lotka-Volterra Recurrent Neural Networks with Harvesting Terms on Time Scales," *Engineering Letters*, vol. 24, no. 4, pp. 455–460, 2016.

[4] Rivero, Jesus E and Valdovinos, Rosa M and Herrera, Edgar and Montes-Venegas, Hector A and Alejo, Roberto, "Thermal Neutron Classification in the Hohlraum Using Artificial Neural Networks," *Engineering Letters*, vol. 23, no. 2, pp. 87–91, 2015.

[5] Fereidouni, Alireza and Masoum, Mohammad Sherkat, "Study on Adaptive Harmonic Extraction Approaches in Active Power Filter Acpplications," *Engineering Letters*, vol. 22, no. 4, pp. 209–214, 2014.

[6] Okamoto, Shingo and Ito, Akihiko, "Effect of Nitrogen Atoms and Grain Boundaries on Shear Properties of Graphene by Molecular Dynamics Simulations," *Engineering Letters*, vol. 22, no. 3, pp. 142–148, 2014.

[7] Wan, Genshun and Song, Xuehua and Bettati, Riccardo, "An Improved EZW Algorithm and Its Application in Intelligent Transportation Systems," *Engineering Letters*, vol. 22, no. 2, pp. 63–69, 2014.

[8] Maturana Daniel and Sebastian Scherer, "VoxNet: A 3d Convolutional Neural Network for Real-time Object Recognition," in *IEEE/RSJ International Conference on Intelligent Robots and Systems 2015*, pp. 922–928.

[9] Hang Su, Subhransu Maji, Evangelos Kalogerakis and Erik Learned-Miller, "Multi-view Convolutional Neural Networks for 3d Shape Recognition," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 945–953.

[10] Charles R. Qi, Hao Su, Kaichun Mo and Leonidas J. Guibas, "Pointnet: Deep Learning on Point Sets for 3d Classification and Segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2017*, pp. 652–660.

[11] Charles Ruizhongtai Qi, Li Yi, Hao Su and Leonidas J. Guibas, "Pointnet++: Deep Hierarchical Feature Learning on Point Sets in a Metric Space," *arXiv preprint arXiv:1706.02413*, 2017.

[12] Zhang, Xiangyu and Zhou, Xinyu and Lin, Mengxiao and Sun, Jian, "Shufflenet: An Extremely Efficient Convolutional Neural Network for Mobile Devices," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2018*, pp. 6848–6856.

[13] Zhirong Wu, Shuran Song, Aditya Khosla, Fisher Yu, Linguang Zhang, Xiaoou Tang and Jianxiong Xiao, "3d Shapenets: A Deep Representation for Volumetric Shapes," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2015*, pp. 1912–1920.

[14] Francis Engelmann, Theodora Kontogianni, Alexander Hermans and Bastian Leibe, "Exploring Spatial Context for 3d Semantic Segmentation of Point Clouds," in *Proceedings of the IEEE International Conference on Computer Vision Workshops 2017*, pp. 716–724.