Semantic Segmentation of Large-scale 3D Point clouds

Chinmay Khamesera
*c*khamesra@wpi.edu

Aditya Dilip Patil
*a*patil@wpi.edu

Durga prakash Karuppannan
*d*karuppannan@wpi.edu

Youness Bani
*y*bani@wpi.edu

***Abstract*—This project focuses on the semantic segmentation of large-scale 3D point clouds for autonomous driving, using LiDAR sensors data. Semantic segmentation is crucial for object identification and classification on the road. Deep learning has made significant strides in real-time semantic segmentation using camera data, but this suffers greatly from erratic camera data under different weather and lighting conditions. LiDAR sensors, on the other hand, are nearly unbreakable in any lighting situation, including day and night, glare and shadows. This paper compares two deep learning architectures, PointNet and SqueezeSegV3 on KITTI dataset , for segmenting LiDAR point clouds with respect to speed and large dataset computation.**

***Index Terms*—Point-Cloud Segmentation, PointNet, Squueze-SegV3, KITTI-Semantic**

I. Introduction

For autonomous driving systems to comprehend their surroundings, classify, and locate objects on road including cars, traffic lights, pedestrians and other barriers, robust perception is crucial. Semantic segmentation is the term used to describe this process of object identification and classification. For semantic segmentation, the perception module uses either camera data or LiDAR 3D point cloud data.

Deep learning has recently made significant strides toward effectively performing semantic segmentation tasks in real time with camera data. However, these methods suffer greatly from the erratic camera data that occurs under different illumination and weather circumstances. LiDAR scanners are nearly unbreakable in any lighting situation, including day and night, glare and shadows present or absent. LiDAR sensors are frequently employed in a variety of fields, most notably autonomous driving[1,2]. Most of the solutions for level 4 and level 5 autonomous vehicles rely on LiDAR to produce a point-cloud representation of the surroundings.LiDAR point clouds can be used in a variety of ways to understand the environment, including point-cloud segmentation, multi-modal fusion, and 2D/3D object detection[3,4]. In this paper, point-cloud segmentation is the main topic. The goal of this exercise is to label each point in a point-cloud according to the type of entity it belongs to. Point-cloud segmentation can be used for autonomous driving to classify things like pedestrians and cars, and more.

The processing of 3D point clouds for segmentation started with PointNet architecture[5]. Using shared multi-layer perceptrons, it learns per-point characteristics (MLPs). Although computationally effective, it fails to capture wider context information of each point. Many specialized neural modules have subsequently and quickly been designed to learn richer local structures. The following categories can be used to classify these modules as graph message forwarding, nearby feature pooling, attention-based aggregation and kernel-based convolution. Two primary subcategories of recent point-cloud segmentation research, focusing on small-scale or large-scale point-clouds, are presented. The majority of current approaches are based on PointNet and are designed to solve small-scale issues, such as object parsing and interpreting indoor scenes. Although having competitive performance in many 3D tasks, PointNet-based approaches have a limited processing speed, especially for large-scale point clouds. PointNet-based techniques are utterly incapable of efficiently processing point clouds at such scales, let alone in real time like autonomous driving. As a result, the spherical projection approach is used in a lot of contemporary research. These techniques convert a 3D LiDAR point cloud into a 2D LiDAR image before using 2D ConvNets to segment the point cloud, as opposed to processing 3D points directly[6].

This research focuses in particular on the following things:

1) With the KITTI dataset, we experimentally performed semantic segmentation using PointNet as a baseline model based on multi-view and volumetric representations.

2)The PointNet has constrained size of receptive fields which prevents it from effectively capturing complex structures for a large-scale point cloud, which typically comprises of hundreds of objects. We implemented SqueezeSegV3 for the superior spatially adaptive convolution technique to segment LiDAR point clouds.

3) We examine and compared the PointNet and SqueezeSegV3 architectural approaches, determining that random sampling is the most effective element for effective learning on huge point clouds. SqueezeSegV3 shows measurable memory and computational improvements over the Point-

Net on the SemanticKITTI dataset.

II. Literature Review

The semantic segmentation of large-scale 3D point clouds is an active area of research with numerous recent works proposing various deep learning architectures and methods [7,8,9]. In this literature review, we will address some of the important publications in this filed [14, 15]. PointNet, which was put forth in [5], is one of the groundbreaking achievements in this discipline. This study suggests PointNet, a deep learning architecture for point cloud analysis. PointNet takes raw point clouds as input and applies a series of shared multi-layer perceptrons (MLPs) to extract local features from each point. These local features are then aggregated and transformed into a global feature vector using max pooling and another MLP. Finally, the global feature vector is used for classification or segmentation. The authors evaluated PointNet on several benchmark datasets and achieved state-of-the-art results on 3D object classification and segmentation tasks, demonstrating the effectiveness of the architecture. PointNet's ability to handle unordered and irregular point clouds without any preprocessing or hand-engineered features makes it a powerful tool for 3D point cloud analysis.

Another notable work is SqueezeSeg [8]. This paper proposes a real-time road-object segmentation system called SqueezeSeg. The system takes as input 3D LiDAR point clouds and outputs a pixel-wise segmentation map of the road and surrounding objects. SqueezeSeg uses a deep convolutional neural network (CNN) to extract features from the point cloud, followed by a recurrent conditional random field (CRF) to refine the segmentation output. The network architecture is designed to be computationally efficient, using a lightweight SqueezeNet-like architecture that reduces the number of parameters while maintaining high accuracy. The authors evaluated SqueezeSeg on several publicly available datasets and achieved state-of-the-art results in real-time road-object segmentation. SqueezeSeg's ability to operate in real-time makes it a promising technology for applications such as autonomous driving and robotics.

SqueezeSegV2 is an extension of the original SqueezeSeg architecture [9] that introduces a novel spatiotemporal convolutional layer that takes advantage of the temporal continuity of LiDAR data to improve segmentation accuracy. The spatiotemporal convolutional layer combines both spatial and temporal information of the point cloud data to capture both local and global features, and is designed to be lightweight and efficient.

SqueezeSegV3 is a further extension of the SqueezeSeg architecture which introduces a new type of convolutional layer called a "Dilated Squeeze Module" that expands the receptive field of the network while reducing computation, and a new "Deformable RoI Pooling" operation that improves feature extraction for small objects. It also introduces a "Sparse Point-wise Convolution" operation that is designed to handle sparse point clouds with missing data.

RandLA-Net, proposed by Hu et al. (2020) [10], is another notable work in this domain. In order to efficiently and accurately segment large-scale point clouds into semantically distinct subsets, this paper suggests a neural network design dubbed RandLA-Net. RandLA-Net uses a novel hierarchical processing approach to handle large-scale point clouds. The point cloud is first partitioned into several smaller overlapping regions, which are then processed by a series of shared sub-networks. Each sub-network uses an edge convolution operation that considers both local and global features to extract discriminative features from the points in the region. The features from all sub-networks are then fused and passed through a global network to produce the final semantic segmentation. The authors also proposed a randomization strategy to improve the generalization ability of the network and reduce overfitting. The authors evaluated RandLA-Net on several large-scale point cloud datasets and achieved state-of-the-art performance in terms of accuracy and efficiency. RandLA-Net's hierarchical processing approach and edge convolution operation make it well-suited for processing large-scale point clouds, which is a critical requirement for many real-world applications such as autonomous driving and urban planning.

Other notable works in this domain include PointCNN [11]. PointCNN extends the PointNet architecture by introducing a novel type of convolution operation that is designed to be permutation invariant, meaning that it can handle input point clouds that are rotated or translated without affecting the output. The PointCNN convolution operation operates on a local geometric structure around each point in the input point cloud, rather than on the points themselves. This geometric structure is defined by a set of neighboring points, and the PointCNN convolution operation applies a set of learnable weights to these neighboring points to compute a new feature vector for the central point. The permutation invariant nature of the PointCNN convolution operation makes it particularly well-suited for processing point clouds that may have arbitrary rotations or translations.

Spherical Polygonal Graph [12] is a neural network architecture proposed for semantic segmentation of large-scale 3D point clouds. SPG uses a graph-based representation of the point cloud, where each point is associated with a local 3D polygonal region, and neighboring polygons are connected to form a graph. SPG applies a series of graph convolutions to the graph to extract features from the local polygonal regions, and uses a feature propagation operation to aggregate information from neighboring polygons. The resulting features are then passed through a series of fully-connected layers to produce the final semantic segmentation. The authors demonstrated that SPG achieves state-of-the-art performance on several benchmark datasets for semantic segmentation of large-scale point clouds, and showed that it is particularly effective at handling point

clouds with varying densities and non-uniform distributions.

KPConv (Kernel Point Convolution) [13] is a neural network architecture proposed for 3D point cloud analysis tasks such as segmentation, classification, and detection. It operates directly on the raw point cloud data, without requiring any intermediate voxelization or projection steps. KPConv uses a novel kernel-based convolution operation that involves representing each point in the point cloud as a weighted sum of nearby kernel points. The kernel points are learned during training, and are used to compute a weighted sum of local features from the input point cloud. KPConv also includes a multi-scale feature extraction component that applies the kernel-based convolution operation at different scales to capture both local and global features in the point cloud. The resulting features are then passed through a series of fully-connected layers to produce the final classification or segmentation output.

These works demonstrate the rapid progress and exciting potential of deep learning for semantic segmentation of large-scale 3D point clouds, which has numerous applications in fields such as autonomous driving, robotics, and urban planning [16,17,18]. In summary, point cloud segmentation is a critical technique in autonomous vehicles that allows them to understand and interact with the environment effectively. It is essential for accurate perception, navigation, and decision-making, making it a crucial component in the development of autonomous vehicles [19].

III. Problem Description

A. Dataset

SemanticKITTI dataset is used for our project. The Semantic KITTI dataset is a large-scale benchmark for autonomous driving research. It is based on the KITTI Vision Benchmark Suite and consists of 21 sequences of point cloud data with over 43000 densely annotated scans.

B. Network Architecture

1) PointNet: PointNet is an innovative architecture for extracting valuable features from an unordered set of 3D points. Its versatile capabilities include performing 3D shape classification, shape part segmentation, and scene semantic parsing tasks. Compared to prior methods, PointNet achieves superior performance and efficiency.

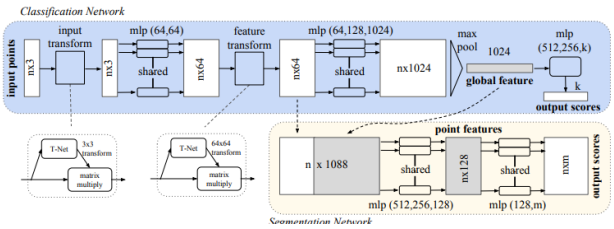

Fig. 1. PointNet Architecture.

Figure 1. shows a diagram of the complete PointNet architecture, where it can be observed that the classification network and the segmentation network share many structures.

PointNet comprises three crucial modules: Firstly, a max pooling layer acts as a symmetrical function to gather information from all points in the input point cloud. Secondly, a structure that combines local and global information. Finally, two joint alignment networks, referred to as "T-Net"s, which align the input points and point features.

2) SqueezeSegV3-Net: The SqueezeSegNet projects a LiDAR point cloud to create an image, which is then analyzed using spatially adaptive convolutions (SAC). The resulting predictions of this network are capable of labeling individual 3D points within the original point cloud. Using standard convolutions to process LIDAR images presents a challenge because these filters only detect and analyze local features that are limited to specific areas in the image. This causes the network's capacity to be underutilized, which results in poorer segmentation performance. To address this issue, Spatially-Adaptive Convolution (SAC), where different filters are applied to various locations based on the input image, is used in this project.

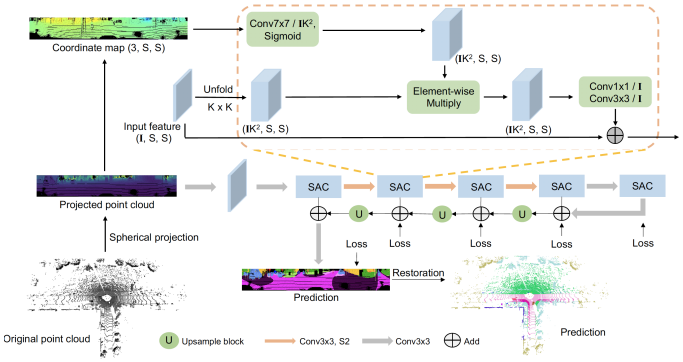

Fig. 2. SqueezeSeg Prediction.

In order to ensure accurate comparison, the backbone architecture of SqueezeSegV3 is based on RangeNet, which consists of five stages as shown in Figure 2. Each stage includes several blocks that begin with downsampling and end with upsampling to recover resolution. Every block in RangeNet comprises two convolutions. The last two downsampling stages are eliminated, and the channel sizes of the last two stages are reduced to maintain the same FLOPs. As a result of removing the last two downsampling operations, only three upsample blocks are employed using transposed convolution and convolution methods. Multi-layer cross-entropy loss is used to train the proposed network. The intermediate supervisions used in the model are more effective in guiding the formation of features with semantic meaning compared to the single-stage cross-entropy and also serve the purpose of reducing the vanishing gradient problem during training.

IV. Results

We assessed the performance of PointNet and SqueezeSegV3 models on the Semantic-Kitti dataset.Figure 3 shows the training and validation loss of SqueezeSegV3 model. We employed 5 semantic classes to seperate point cloud: Car, Bicycle, Motorcycle, Trucks and Person as shown in Figure 4. The Mean IoU was evaluated to obtain the performance metric of both the models. Figure 5 shows the visualisation of point cloud segmentation performed on PointNet architecture. Figure 6 and 7 represents the visualization of point cloud segmentation performed on SqueezeSeg architecture.

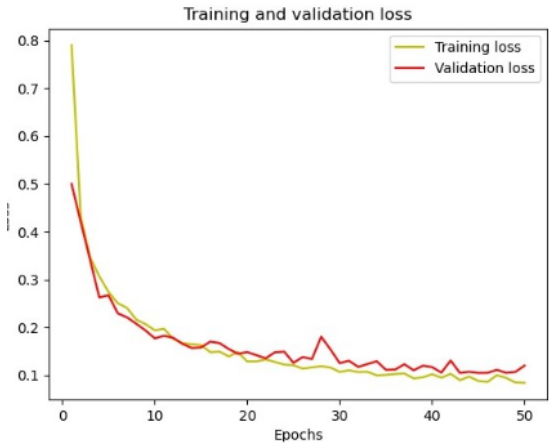

Fig. 3. Training and Validation loss of SqueezesegV3

Method	mIoU	Car	Bicycle	Motorcycle	Truck	Person
PointNet	13.5	50.2	5.4	0.2	1.6	0.5
SegSquueezeV3	28.2	78.6	15.0	7.3	2.3	10.4

Fig. 4. Comparision table of PointNet and SqueezeSeg

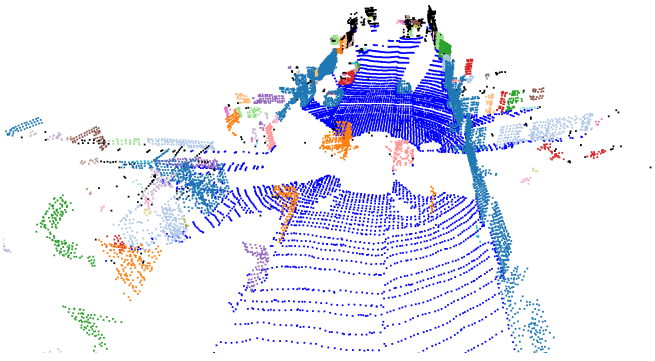

Fig. 5. PointNet point-cloud segmentation

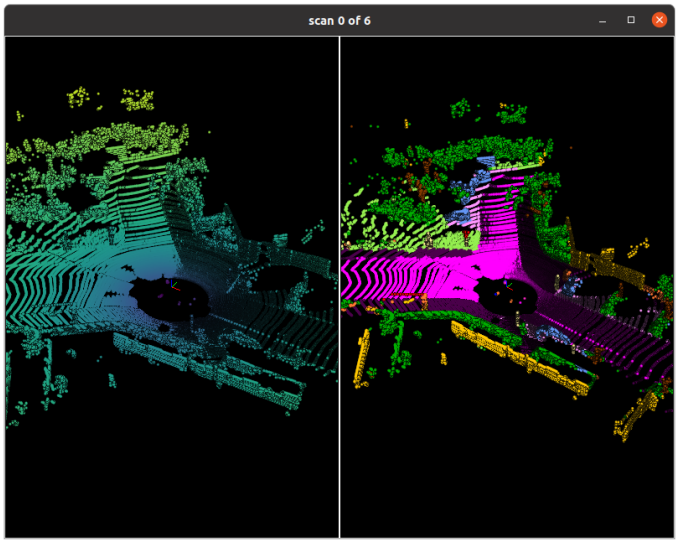

Fig. 6. SqueezeSeg point-cloud segmentation

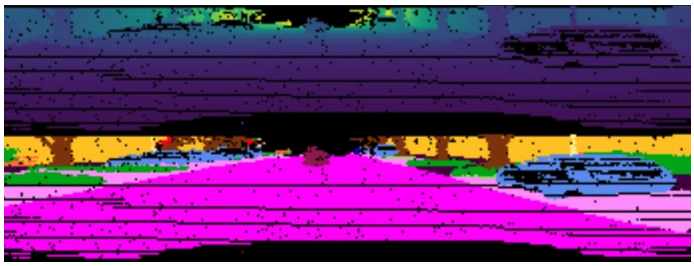

Fig. 7. SqueezeSeg Results.

V. Conclusion and Future Work

In this study, we have demonstrated the excellent accuracy of 3D semantic segmentation on lidar point clouds using deep learning techniques like PointNet and SqueezeSegV3. We have demonstrated that our method is capable of reliably labeling each point in a point cloud with its matching semantic class by measuring the performance of our models using metrics like MIOU and accuracy. SqueezeSegV3 outperforms PointNet in terms of performance and accuracy. This work is crucial for the development of more precise and effective autonomous driving systems that can confidently navigate challenging environments.

Future employment opportunities in this field are numerous. In order to increase the precision and effectiveness of our models, we could first investigate the usage of more sophisticated deep learning architectures and methods, such as graph convolutional networks or attention mechanisms. Second, to enhance the performance of our models on datasets from various contexts or geographies, we may look into the usage of transfer learning or domain adaptation techniques. The use of lidar data in combination with other sensor modalities, such cameras or radar, may also be investigated in order to create autonomous driving systems that are more durable and dependable.

REFERENCES

[1] Bo Li, Tianlei Zhang, and Tian Xia. Vehicle detection from 3D lidar using fully convolutional network. In RSS, 2016.

[2] Iro Armeni, Sasha Sax, Amir R Zamir, and Silvio Savarese. Joint 2D-3D-semantic data for indoor scene understanding. In CVPR, 2017.

[3] Jens Behley, Martin Garbade, Andres Milioto, Jan Quenzel, Sven Behnke, Cyrill Stachniss, and Juergen Gall. SemanticKITTI: A dataset for semantic scene understanding of lidar sequences. In ICCV, 2019.

[4] Alexandre Boulch, Bertrand Le Saux, and Nicolas Audebert. Unstructured point cloud semantic labeling using deep segmentation networks. In 3DOR, 2017.

[5] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. PointNet: Deep learning on point sets for 3D classification and segmentation. In CVPR, 2017.

[6] Chenfeng Xu, Bichen Wu, Zining Wang, Wei Zha, Peter Vajda, Kurt Keutzer, and Masayoshi Tomizuka1. SqueezeSegV3: Spatially-Adaptive Convolution for Efficient Point-Cloud Segmentation.

[7] Aksoy, Eren Erdal, Baci, Saimir, and Cavdar, Selcuk. Salsanet: Fast road and vehicle segmentation in lidar point clouds for autonomous driving. In 2020 IEEE intelligent vehicles symposium (IV), pp. 926–932. IEEE, 2020.

[8] Wu, Bichen,Wan, Alvin, Yue, Xiangyu, and Keutzer, Kurt. Squeezeseg: Convolutional neural nets with recurrent crf for real-time road-object segmentation from 3d lidar point cloud. In 2018 IEEE International Conference on Robotics and Automation (ICRA), pp. 1887–1893. IEEE, 2018.

[9] Wu, Bichen, Zhou, Xuanyu, Zhao, Sicheng, Yue, Xiangyu, and Keutzer, Kurt. Squeezesegv2: Improved model structure and unsupervised domain adaptation for roadobject segmentation from a lidar point cloud. In 2019 International Conference on Robotics and Automation (ICRA), pp. 4376–4382. IEEE, 2019.

[10] Hu, Qingyong, Bo Yang, Linhai Xie, Stefano Rosa, Yulan Guo, Zhihua Wang, Niki Trigoni, and Andrew Markham. "Randla-net: Efficient semantic segmentation of large-scale point clouds." In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 11108-11117. 2020.

[11] Li, Yangyan, Rui Bu, Mingchao Sun, Wei Wu, Xinhan Di, and Baoquan Chen. "Pointcnn: Convolution on x-transformed points." Advances in neural information processing systems 31 (2018).

[12] Landrieu, Loic, and Martin Simonovsky. "Large-scale point cloud semantic segmentation with superpoint graphs." In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 4558-4567. 2018.

[13] Thomas, Hugues, Charles R. Qi, Jean-Emmanuel Deschaud, Beatriz Marcotegui, François Goulette, and Leonidas J. Guibas. "Kpconv: Flexible and deformable convolution for point clouds." In Proceedings of the IEEE/CVF international conference on computer vision, pp. 6411-6420. 2019.

[14] Cortinhal, Tiago, Tzelepis, George, and Aksoy, Eren Erdal. Salsanext: Fast, uncertainty-aware semantic segmentation of lidar point clouds for autonomous driving. arXiv preprint arXiv:2003.03653, 2020.

[15] Milioto, Andres, Vizzo, Ignacio, Behley, Jens, and Stachniss, Cyrill. Rangenet++: Fast and accurate lidar semantic segmentation. In 2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 4213–4220. IEEE, 2019.

[16] Qi, Charles Ruizhongtai, Yi, Li, Su, Hao, and Guibas, Leonidas J. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. Advances in neural information processing systems, 30, 2017b.

[17] Rosu, Radu Alexandru, Sch¨utt, Peer, Quenzel, Jan, and Behnke, Sven. Latticenet: Fast point cloud segmentation using permutohedral lattices. arXiv preprint arXiv:1912.05905, 2019.

[18] Shelhamer, Evan, Long, Jonathan, and Darrell, Trevor. Fully convolutional networks for semantic segmentation. IEEE transactions on pattern analysis and machine intelligence, 39(4):640–651, 2016.

[19] Simon, Martin, Amende, Karl, Kraus, Andrea, Honer, Jens, Samann, Timo, Kaulbersch, Hauke, Milz, Stefan, and Michael Gross, Horst. Complexer-yolo: Real-time 3d object detection and tracking on semantic point clouds. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, pp. 0– 0, 2019.