Hochschule
Bonn-Rhein-Sieg
University of Applied Sciences

Bonn-Aachen
International Center for
Information Technology

Out-of-Distribution Detection in 3D Semantic Segmentation

Master Thesis

September 02, 2022

Lokesh Veeramacheneni

Supervisors

Prof. Dr Paul G Ploeger, Prof. Dr Matias Valdenegro Toro, Prof. Dr Sebastian Houben

1. Introduction

2. Experimental Setup & Methodology

3. Experiments & Results

4. Conclusion

Out-of-Distribution detection

- Tesla autonomous driving system detects the moon as yellow traffic light
- These faulty predictions might result in autonomous agent behaving unpredictably
- An ideal trustworthy visual recognition system
  - Produce accurate predictions on known examples
  - Detect and reject unknown examples

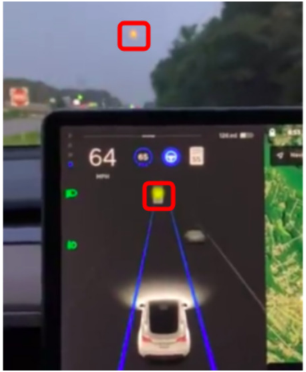

Figure 1: Misdetection of Moon as yellow signal light in Tesla driving platform. Image taken from [6].

Out-of-Distribution detection

- Deep Neural Networks (DNNs) are trained based on closed world assumption
- closed world assumption - test data is assumed to be drawn from same distribution as training data which is called In-Distribution (ID)
- When deployed in real world (open world scenario) the test samples can be Out-of-Distribution (OOD) i.e. the test samples can be,
  - from different class
  - from different domain
- Moon is calssified as an OOD object in the above example

Importance of OOD detection

- Figure 2 depicts the pipeline of modules in Apollo driving platform.
- Prediction and motion planning module are dependent on perception module.
- A misdetection of an OOD sample will propogate the error to motion planning and affects the total vehicle control and this might lead to unfortunate consequence

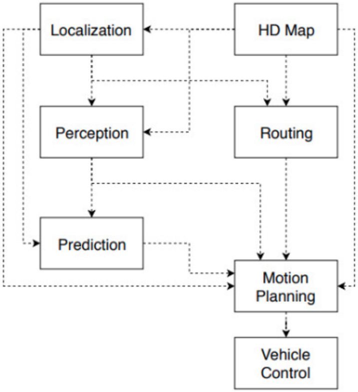

Figure 2: Module pipeline for Apollo autonomous driving platform. Image taken from [1].

3D Light Detection And Ranging (LiDAR)

- Uses pulsed lasers to find the range to the objects
- Unlike images, LiDAR is insusceptible to illumination and provide rich 3D information.
- Figure 3 depicts the sample point cloud with LiDAR is placed in round white circle found at the center of point cloud
- Typically, features of each point in point cloud include
  - spatial features (XYZ)
  - Colour (RGB)
  - Intensity

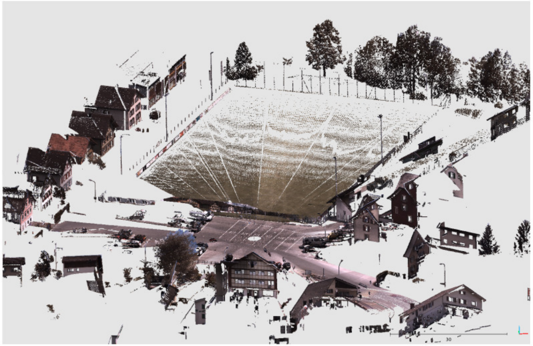

Figure 3: Sample LiDAR point cloud collected in a outdoor scene. Image taken from [3].

3D Semantic Segmentation

- An important task in computer vision because of its use in scene understanding
- Further helps in navigation and planning of robots
- Objective - Assign each point in the point cloud a specific class

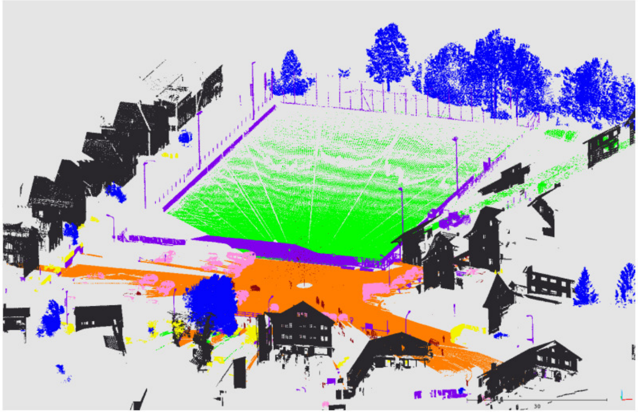

Figure 4: Segmented output of sample point cloud. Image taken from [3].

Thesis objective

- OOD detection in the 3D semantic segmentation setting
- Create a benchmark datasets for OOD detection among existing 3D LiDAR datasets. We define OOD data based on two categories
  - if the point is from different class than training data
  - if the point has inferior quality
- We also study whether uncertainty estimation is a practical approach for OOD detection in 3D domain

Setup

- 3D Semantic Segmentation model
- Uncertainty methods
- OOD score methods
- Datasets

RandLA-Net

- Lightweight, efficient computation, meomry usage and inputs 3D point cloud directly
- Random point sampling and local feature aggregation module are most important modules
- Local feature aggregation module is subdivided into local spatial encoding, attentive pooling and dilated residual block
- Encoder-Decoder style architecture as depicted in Figure6

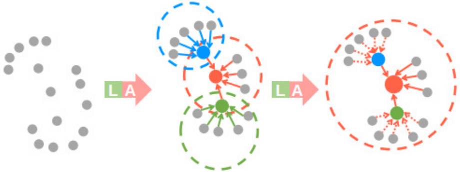

Figure 5: Image depicting the working of Dilated residual block with each circle representing the receptive field of the block for feature extraction. LA represents the combination of Local Spatial Encoding and Attentive Pooling modules combined. Image taken from [4].

RandLA-Net

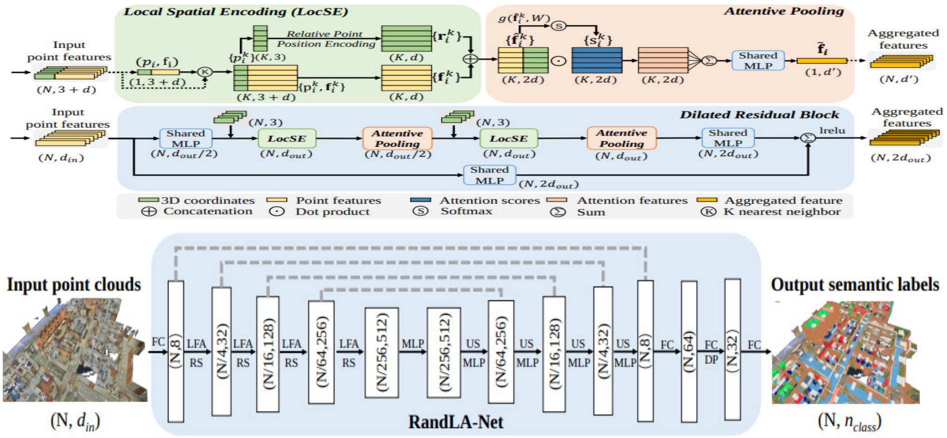

Figure 6: Illustration of (a) local feature aggregation module in RandLa-Net and (b) architecture of RandLA-Net. Both the images are taken from [4].

Setup

- 3D Semantic Segmentation model - RandLA-Net
- Uncertainty methods
- OOD score methods
- Datasets

Deep Ensembles

- Ensemble learning technique - train N randomly initialized models with same data
- Reulting N predictions are then averged
- Perormance boosting along with uncertainty value for a prediction
- Requires more computation power

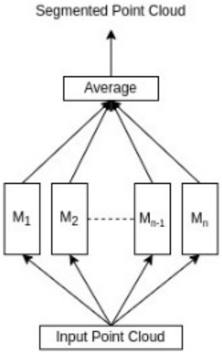

Figure 7: Illustration of test dataflow in Deep Ensembles, where input point cloud is fed into multiple randomly initialized models.

Flipout

- Introduced as a method to decorrelate gradients in a mini batch of examples
- Add independent weight perturbations sampled from a distribution
- The output of Flipout versioned neuron is
- Train single instance of Flipout versioned network and then perform multiple forward passes for same input

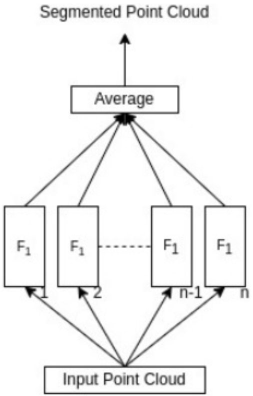

Figure 8: Illustration of test dataflow in Flipout. Here F1 represents the Flipout trained model and we compute n forward passes of the same point cloud on F1 .

Setup

- 3D Semantic Segmentation model - RandLA-Net
- Uncertainty methods - Deep Ensembles & Flipout
- OOD score methods
- Datasets

OOD Score calculation

- We use the following two methods to generate the OOD scores.
- Maximum Softmax Probability
  - $max(y_n), y_n = [P(C_1), P(C_2), ..., P(C_n)]$
- Entropy
  - $-\sum_i P(x_i) log(P(x_i))$ with i iterates across all the classes for point x

Setup

- 3D Semantic Segmentation model - RandLA-Net
- Uncertainty methods - Deep Ensembles & Flipout
- OOD score methods - Maximum Softmax Probability & Entropy
- Datasets

3D LiDAR datasets

acquisition mode	dataset	frames	total points (in million)	classes	scene type
static	Oakland[60]	17	1.6	44	outdoor
	Paris-lille-3D[71]	3	143	50	outdoor
	Paris-rue-Madame[74]	2	20	17	outdoor
	S3DIS[5]	5	215	12	indoor
	ScanObjectNN[85]	-	-	15	indoor
	Semantic3D[31]	30	4009	8	outdoor
	TerraMobilita/IQmulus[88]	10	12	15	outdoor
	TUM City Campus[26]	631	41	8	outdoor
	DALES[90]	40 (tiles)	492	8	outdoor
sequential	A2D2[27]	41277	1238	38	outdoor
	AIO Drive[96]	100	-	23	outdoor
	KITTI-360[100]	100K	18000	19	outdoor
	nuScenes-lidarseg[12]	40000	1400	32	outdoor
	PandaSet[99]	16000	1844	37	outdoor
	SemanticKITTI[7]	43552	4549	28	outdoor
	SemanticPOSS[62]	2988	216	14	outdoor
	Sydney Urban[19]	631	-	26	outdoor
	Toronto-3D[79]	4	78.3	8	outdoor
synthetic	SynthCity[30]	75000	367.9	9	outdoor

Table 1: 3D LiDAR datasets classified based on the acquisition type. Table updated from [2].

Semantic3D

- Huge 3D point cloud benchmark classification static dataset with 4 million points
- Scenes are taken in european streets around church, stations and fields
- Point features include XYZ, RGB and Intensity values.
- It has 8 classes with distribution of points represented in Figure9
- cite states that the scanning artefacts, hardscapes and cars are the most challenging classes

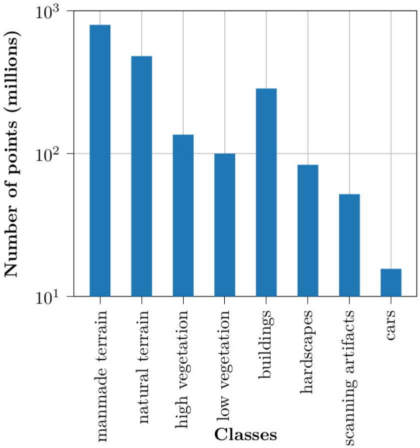

Figure 9: Graph depicting the number of points per class (in millions) in the Semantic3D dataset.

Semantic3D

Figure 10: Illustration of the Semantic3D point clouds of various outdoor scenes.

S3DIS

- Indoor dataset with scans from various buildings
- Dataset include scans of personal offices, restrooms, open spaces, lobbies and hallways
- It has 12 classes, further subdivided into two types
  - structural elements
  - everyday items
- One of the most evaluated datasets for indoor semantic segmentation

Figure 11: Illustration of the S3DIS point clouds of various indoor scenes.

OOD Benchmark datasets

ID dataset	OOD dataset	OOD detection difficulty	Summary
Semantic 3D	S3DIS	Easy	1. No class overlap 2. Less structural similarity 3. Different domain(outdoor-vs-indoor)
	Semantic3D without color	Hard	1. Same structural properties 2. Difference in RGB values 3. Same domain as ID dataset 4. Same classes

Table 2: Table representing the ID dataset and corresponding OOD datasets, difficulty in OOD detection and the summary of reasons to chose this OOD dataset.

Setup

- 3D Semantic Segmentation model - RandLA-Net
- Uncertainty methods - Deep Ensembles & Flipout
- OOD score methods - Maximum Softmax Probability & Entropy
- Datasets - Semantic3D, S3DIS & Semantic3D w/o colour

Experiments

- Semantic3D (ID) vs S3DIS (OOD)
  - Deep Ensembles
  - Flipout
  - Area Under Receiver Operating Characteristic (AUROC) score comparison
- Semantic3D vs Semantic3D w/o colour

Semantic3D vs S3DIS - Deep Ensembles

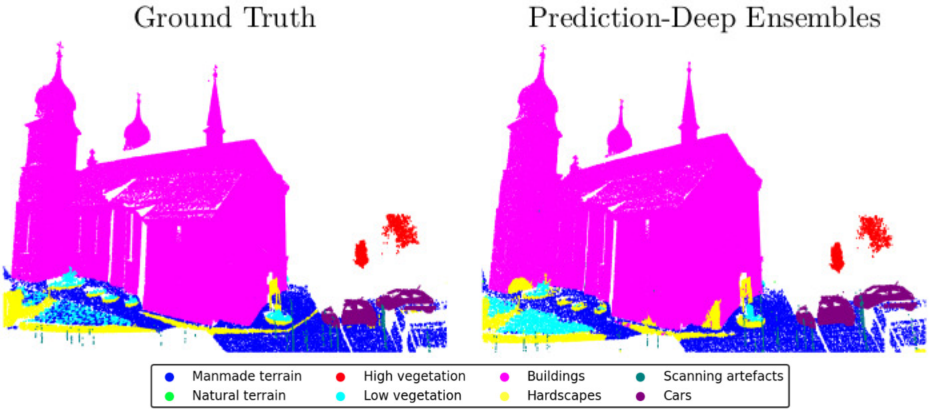

Figure 12: Image representing the predictions (last column) from Deep Ensemble with an ensemble size of 15 on Semantic3D dataset. The first column depict the ground truth.

Semantic3D vs S3DIS - Deep Ensembles

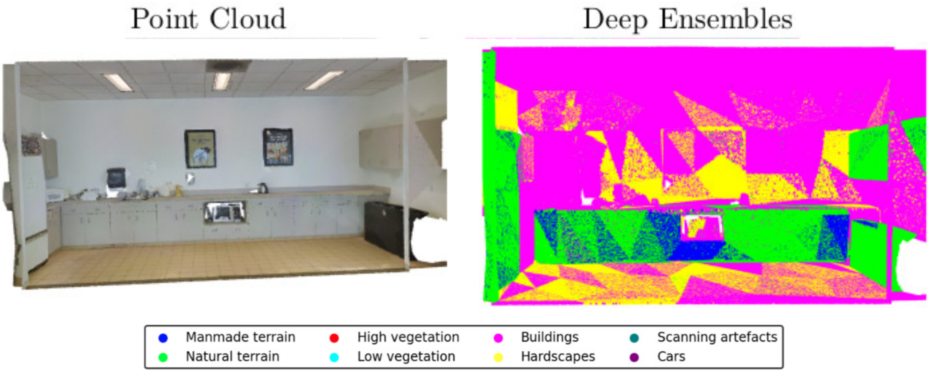

Figure 13: Predictions of RandLA-Net on S3DIS (OOD) dataset. First column representing the point cloud, second column presenting the predictions of Deep Ensembles (15 Ensemble size)

Semantic3D vs S3DIS - Deep Ensembles

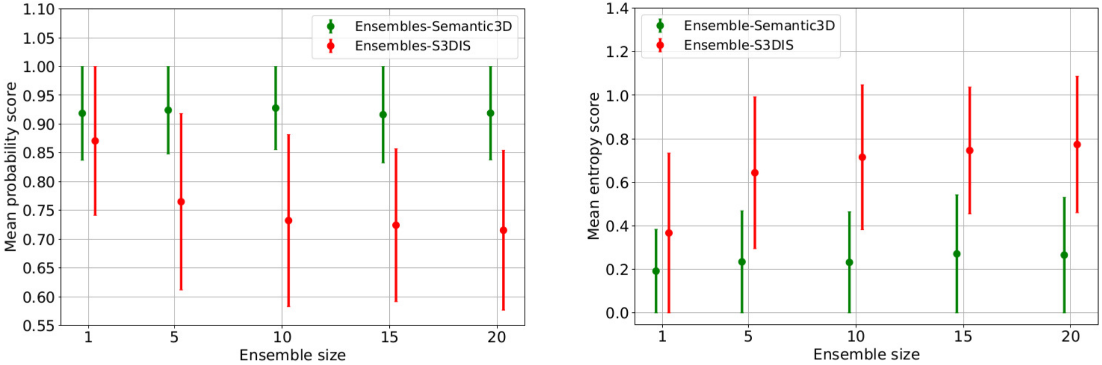

Figure 14: Graphs representing the mean probability value and entropy as a dot for Semantic3D (ID) in green and S3DIS (OOD) in red when using Deep Ensembles. The variance is represented via the error bars.

Semantic3D vs S3DIS - Flipout

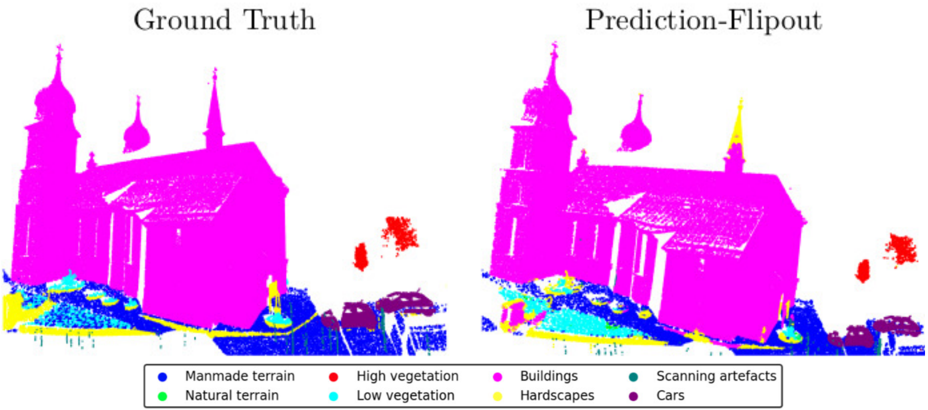

Figure 15: Image representing the predictions (last column) from Flipout with 15 number of passes on Semantic3D dataset. The first column depict the ground truth.

Semantic3D vs S3DIS - Flipout

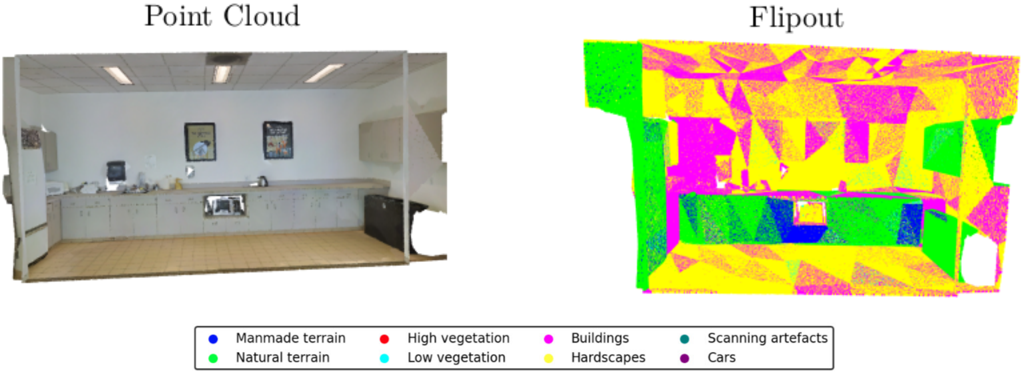

Figure 16: Predictions of RandLA-Net on S3DIS (OOD) dataset. First column representing the point cloud, second column presenting the predictions from Flipout (15 number of passes)

Semantic3D vs S3DIS - Flipout

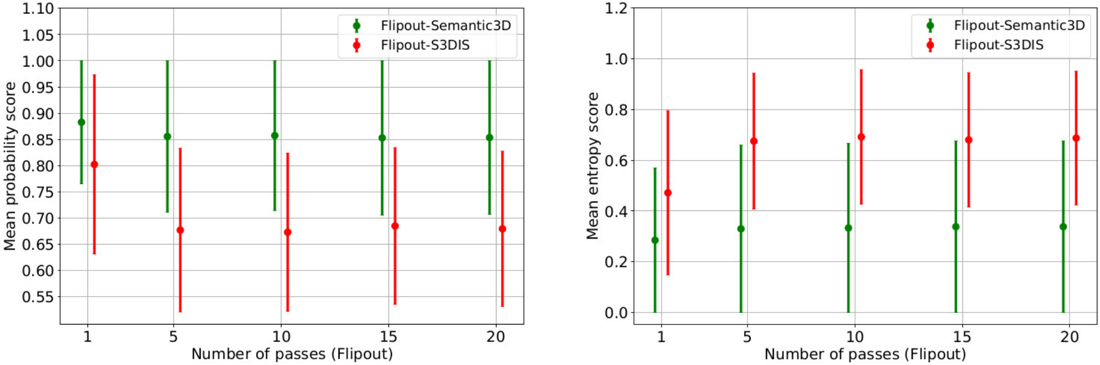

Figure 17: Graphs representing the mean probability value and entropy as a dot for Semantic3D (ID) in green and S3DIS (OOD) in red when using Flipout. The variance is represented via the error bars.

Semantic3D vs S3DIS - AUROC Scores

Ensemble size/ #passes	Method	AUROC	
		MSP	Entropy
1	Dropout	0.53311	0.53041
	Flipout	**0.69988**	**0.69368**
	Deep Ensembles	0.62020	0.62529
5	Dropout	0.58439	0.57821
	Flipout	0.77885	0.76934
	Deep Ensembles	**0.84013**	**0.83665**
10	Dropout	0.60168	0.59925
	Flipout	0.78728	0.78327
	Deep Ensembles	**0.87929**	**0.87541**
15	Dropout	0.59773	0.59557
	Flipout	0.7667	0.76741
	Deep Ensembles	**0.88486**	**0.88246**
20	Dropout	0.59766	0.59661
	Flipout	0.77331	0.77237
	Deep Ensembles	**0.89338**	**0.89052**

Table 3: AUROC scores calculated for all the points in the test sets of Semantic3D and S3DIS for Dropout, Flipout, and Deep Ensembles generated using MSP and entropy values for various ensemble sizes and forward passes.

Semantic3D vs S3DIS - AUROC Scores

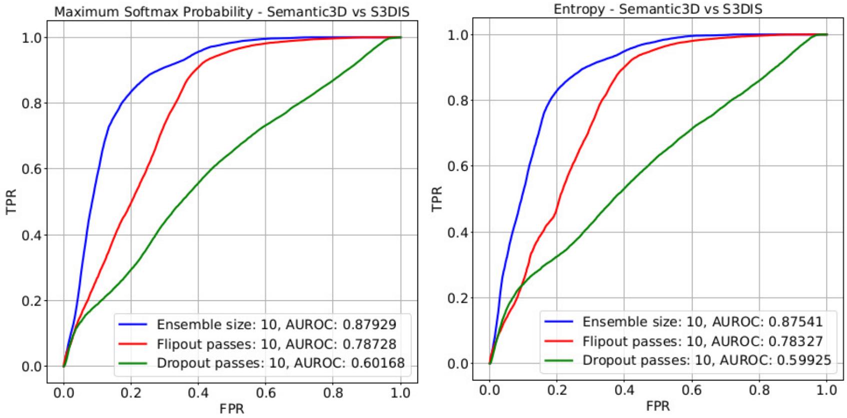

Figure 18: ROC curves of Semantic3D-vs-S3DIS for 10 Ensembles, 10 forward passes for Flipout and Dropout. (a) representing the ROC curve from MSP and (b) representing the ROC curve from the entropy score.

Experiments

- Semantic3D (ID) vs S3DIS (OOD)
- Semantic3D vs Semantic3D w/o colour
  - Deep Ensembles
  - Flipout
  - Area Under Receiver Operating Characteristic (AUROC) score comparison

Semantic3D colour vs w/o colour - Deep Ensembles

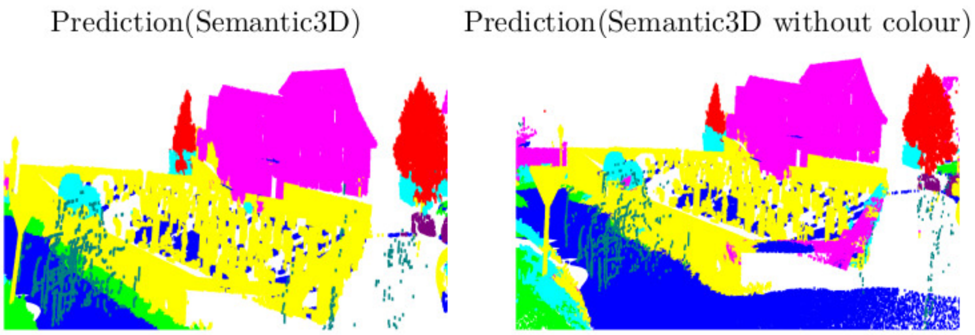

Figure 19: Output predictions of the RandLA-Net over the Semantic3D dataset and Semantic3D without colour dataset using Deep Ensembles (Ensemble size of 10)

Semantic3D colour vs w/o colour - Deep Ensembles

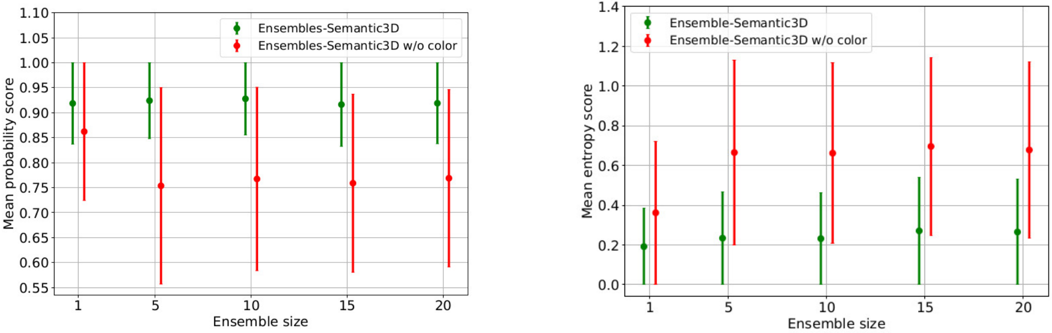

Figure 20: Graphs representing the mean probability value and entropy as a dot for Semantic3D (ID) in green and Semantic3D w/o colour (OOD) in red when using Deep Ensembles. The variance is represented via the error bars.

Semantic3D colour vs w/o colour - Flipout

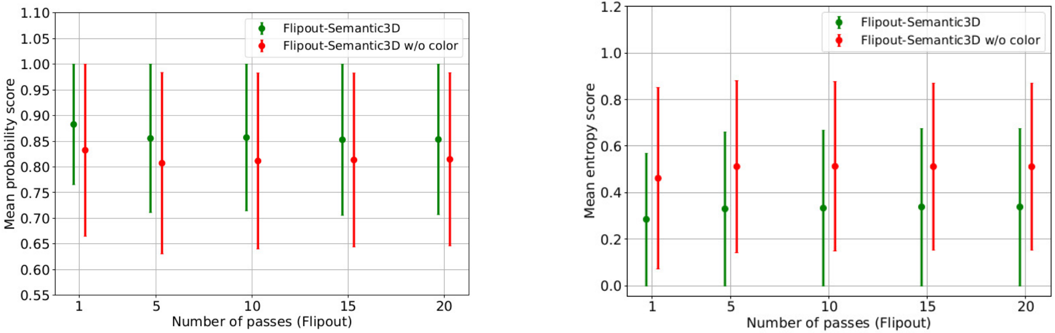

Figure 21: Graphs representing the mean probability value and entropy as a dot for Semantic3D (ID) in green and Semantic3D w/o colour (OOD) in red when using Flipout. The variance is represented via the error bars.

Semantic3D colour vs w/o colour - AUROC Scores

Ensemble size/ #passes	Method	AUROC	
		MSP	Entropy
1	Dropout	0.66349	0.65908
	Flipout	0.64221	0.66157
	Deep Ensembles	**0.67855**	**0.67866**
5	Dropout	0.69448	0.68507
	Flipout	0.63743	0.66536
	Deep Ensembles	**0.76769**	**0.77120**
10	Dropout	0.68568	0.68004
	Flipout	0.63712	0.66535
	Deep Ensembles	**0.77837**	**0.78142**
15	Dropout	0.68975	0.68347
	Flipout	0.63022	0.65976
	Deep Ensembles	**0.77302**	**0.77881**
20	Dropout	0.68447	0.68199
	Flipout	0.63017	0.65934
	Deep Ensembles	**0.77031**	**0.77584**

Table 4: AUROC scores in case of Semantic3D-vs-Semantic3D without colour for Dropout, Flipout, and Deep Ensembles generated using MSP and entropy values for various ensemble sizes and forward passes.

Semantic3D colour vs w/o colour - AUROC Scores

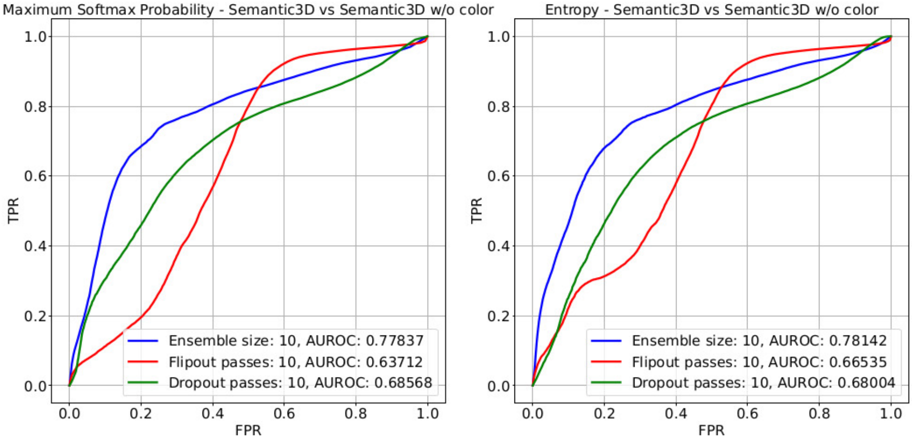

Figure 22: ROC curves of Semantic3D-vs-Semantic3D without colour for 10 ensembles, 10 forward passes for Flipout and Dropout. (a) representing the ROC curve from MSP and (b) representing the ROC curve from the entropy score.

Conclusion

- We propose two benchmark datasets
  - Semantic3D-vs-S3DIS (Outdoor-vs-Indoor) - Easy OOD identification
  - Semantic3D-vs-Semantic3D without colour - Hard OOD identification
- The second case is hard becuase of same point geometry between ID and OOD datasets
- Both Maximum Softmax Probability and Entropy are able to identify OOD points
- Deep Ensembles outperform Flipout and Dropout in both the benchmark datasets

Lessons Learned

Learning's during the duration of the thesis are

1. Training and evaluation of 3D DNNs are time consuming and resource intensive.
2. Finding the proper prior for Flipout layers is hard and currently we use brute force to find the best fitting prior.
3. LiDAR datasets have large memory requirements especially for the preprocessing and metric computation.
4. Getting 100% OOD detection performance is not possible with the post-hoc methods used as some points in the ID dataset also have low probability scores.

Future Work

This thesis can be extended in the following ways.

1. This thesis is limited to only point based models, this can be extended to graph and projection based models.
2. The datasets involved are only static datasets and this thesis study can be further extended to other type of datasets such as synthetic and sequential datasets.
3. Since this thesis utilzes post-hoc threshold methods for OOD detection. Other methods such as Mahalanobis distance based OOD detection [5] or MetaSeg [7] can be added as an extension to this thesis.

References (1/3)

- Haoyang Fan, Fan Zhu, Changchun Liu, Liangliang Zhang, Li Zhuang, Dong Li, Weicheng Zhu, Jiangtao Hu, Hongye Li, and Qi Kong.
  Baidu apollo em motion planner.
  arXiv preprint arXiv:1807.08048, 2018.
- Biao Gao, Yancheng Pan, Chengkun Li, Sibo Geng, and Huijing Zhao.
  Are we hungry for 3d lidar data for semantic segmentation? a survey of datasets and methods.
  IEEE Transactions on Intelligent Transportation Systems, pages 1–19, 2021.
- Timo Hackel, Nikolay Savinov, Lubor Ladicky, Jan D Wegner, Konrad Schindler, and Marc Pollefeys.
  Semantic3d. net: A new large-scale point cloud classification benchmark.
  arXiv preprint arXiv:1704.03847, 2017.

References (2/3)

- Qingyong Hu, 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 (CVPR)*, June 2020.
- Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin.
  A simple unified framework for detecting out-of-distribution samples and adversarial attacks.
  Advances in neural information processing systems, 31, 2018.

References (3/3)

Tim Levin.
Tesla's full self-driving tech keeps getting fooled by the moon, billboards, and burger king signs, 2021.
[Online; accessed December 24, 2021].

Philipp Oberdiek, Matthias Rottmann, and Gernot A. Fink.
Detection and retrieval of out-of-distribution objects in semantic segmentation.
In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR Workshops*, pages 1331–1340. Computer Vision Foundation / IEEE, 2020.

What is OOD Detection?

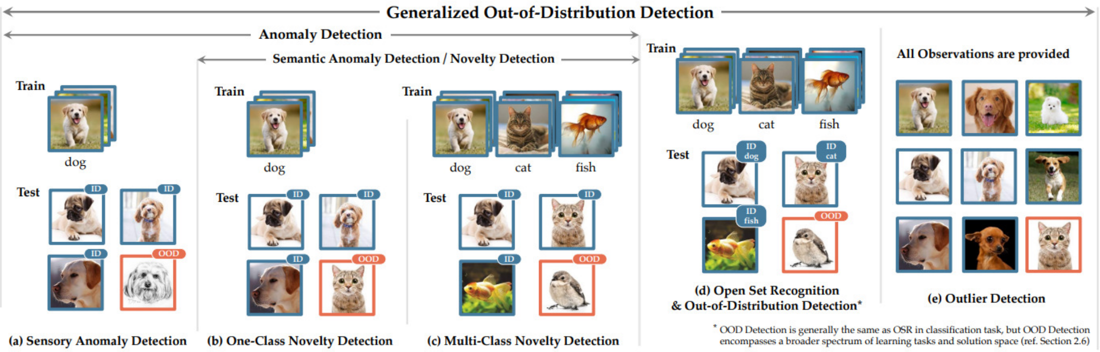

Figure 23: Generalized Out-of-Distribution Detection: A Survey

Semantic3D-Deep Ensembles

		IoU per-class								
Ensemble size	**meanIoU**	**C1**	**C2**	**C3**	**C4**	**C5**	**C6**	**C7**	**C8**	**Accuracy**
1	68.19	94.55	81.19	84.67	29.43	81.37	18.85	64.74	90.74	88.78
5	69.51	94.73	81.92	84.42	28.05	**86.41**	28.50	61.03	91.03	90.04
10	69.97	95.25	83.73	86.63	30.36	84.13	18.60	**66.01**	92.61	89.94
15	70.32	95.27	83.54	**88.22**	**32.19**	84.82	26.17	61.67	90.75	**90.57**
20	**70.80**	**95.55**	**84.11**	86.65	29.60	85.41	**29.58**	62.47	**93.06**	90.56

Table 5.1: Illustration of performance of RandLA-Net on Semantic3D over ensemble size. meanIOU, IOU per-class and overall accuracy are represented here. C1 to C8 are the classes of Semantic3D which are Manmade terrain, Natural terrain, High vegetation, Low vegetation, Buildings, Hardscapes, Scanning artefacts, and Cars.

Semantic3D-Flipout

		IoU per-class								
#Passes	**MeanIoU**	**C1**	**C2**	**C3**	**C4**	**C5**	**C6**	**C7**	**C8**	**Accuracy**
1	69.95	94.24	80.09	86.16	22.48	88.70	39.41	57.42	91.12	90.71
5	69.83	94.38	80.21	84.10	23.32	87.80	39.68	57.75	91.43	90.43
10	69.84	94.38	80.16	83.90	23.46	87.73	39.75	57.83	91.47	90.40
15	69.86	94.38	80.17	83.80	23.48	87.73	39.82	57.96	91.57	90.40
20	69.87	94.38	80.18	83.80	23.57	87.72	39.84	57.92	91.57	90.40

Table 5.2: Illustration of performance of Flipout-versioned RandLA-Net on Semantic3D dataset. meanIOU, IOU per-class and overall accuracy are represented here. C1 to C8 are the classes of Semantic3D which are Manmade terrain, Natural terrain, High vegetation, Low vegetation, Buildings, Hardscapes, Scanning artefacts, and Cars.