



Hochschule
Bonn-Rhein-Sieg
University of Applied Sciences



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 result in unpredictable behaviour
- An ideal trustworthy visual recognition system
 - Produce accurate predictions on known samples
 - Detect and reject unknown samples



Figure 1: Misdetetection of Moon as yellow signal light in Tesla driving platform. Image taken from [7].

Out-of-Distribution detection

- Deep Neural Networks (DNNs) are trained based on closed world assumption
- Closed world assumption - test data is assumed to be from same distribution as **training data** which is called **In-Distribution (ID)** data
- When deployed in real-world (**open world scenario**), the test samples can be
 - from different class
 - from different domain
- These test samples are called **Out-of-Distribution** samples

Importance of OOD detection

- Prediction and motion planning modules depend on perception module
- OOD sample misdetection propagates error to motion planning
- This affects the total vehicle control and lead to catastrophic consequences

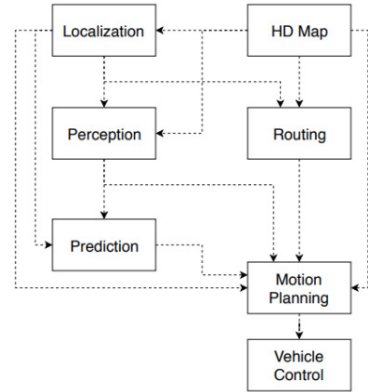


Figure 2: Module pipeline for Apollo autonomous driving platform. Image taken from [2].

3D Light Detection And Ranging (LiDAR)

- Uses **pulsed lasers** to find the **range** to the objects
- Unlike images, LiDAR is insusceptible to illumination
- It provide rich 3D information
- Typically, features of each point includes
 - Spatial features (XYZ)
 - Colour (RGB)
 - Intensity
 - Surface normals



Figure 3: Sample LiDAR point cloud collected in a outdoor scene. Image taken from [4].

3D Semantic Segmentation

- An important task in computer vision because of scene understanding
- Helps in navigation and planning of robots
- Objective - Assign **each point** in the point cloud to a **specific class**

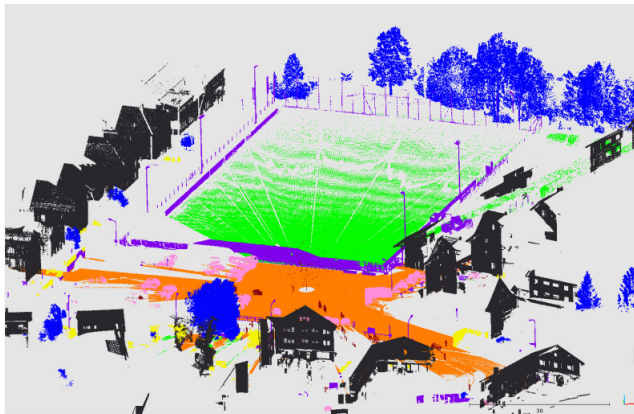


Figure 4: Segmented output of sample point cloud. Image taken from [4].

Thesis objective

- OOD detection in the 3D semantic segmentation.
- Propose datasets for OOD detection benchmarking. We define OOD data based on two categories
 - if the point is from different class to training data
 - if the point has inferior quality
- Study whether uncertainty estimation a practical approach for OOD detection in 3D Semantic Segmentation.

1. Introduction

2. Experimental Setup & Methodology

3. Experiments & Results

4. Conclusion



Setup

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

RandLA-Net

- Lightweight, efficient computation, memory 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

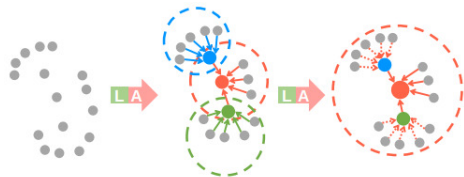


Figure 5: Image depicting the working of Dilated residual block with each circle representing the receptive field of the block for feature extraction. LA represents the combination of Local Spatial Encoding and Attentive Pooling modules combined. Image taken from [5].

RandLA-Net

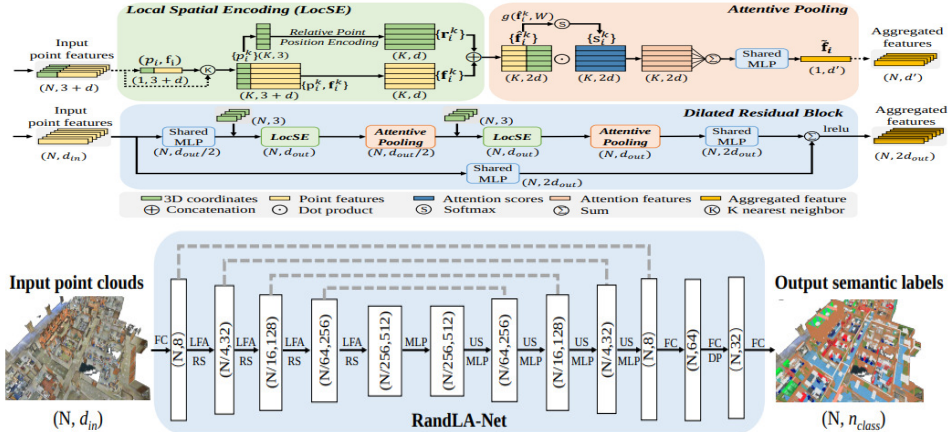


Figure 6: Illustration of local feature aggregation module in RandLA-Net in top image and architecture of RandLA-Net in bottom image. Both the images are taken from [5].

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
- Resulting N predictions are then averaged
- **Performance boosting** along with uncertainty value for a prediction
- Requires **more computation power**

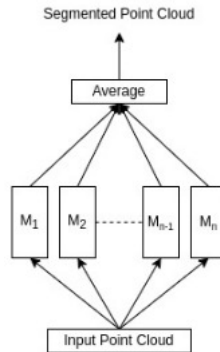


Figure 7: Illustration of test dataflow in Deep Ensembles, where input point cloud is fed into multiple randomly initialized models M_1 to M_n .

Bayesian Neural Networks - Flipout

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

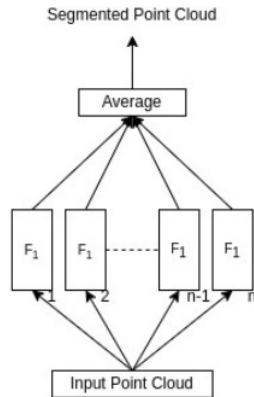


Figure 8: Illustration of test dataflow in Flipout. Here F_1 represents the Flipout trained model and we compute n forward passes of the same point cloud on F_1 .

Bayesian Neural Networks - Flipout

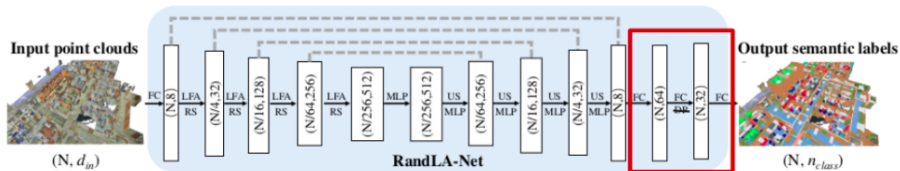


Figure 9: Flipout-versioned RandLA-Net where the last three FC layers as depicted in red box are made Flipout-compatible. Image taken from [5].

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
 - $\max(y_n), y_n = [P(C_1), P(C_2), \dots, P(C_n)]$
- 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 |
| | SynthCity [30] | 75000 | 367.9 | 9 | outdoor |
| synthetic | SynthCity [30] | 75000 | 367.9 | 9 | outdoor |

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

Semantic3D

- 3D point cloud benchmark dataset with 4 million points
- Scenes are captured around european streets including church, stations and fields
- Point features include XYZ, RGB and Intensity values.
- It has 8 classes with distribution of points represented in adjacent figure
- [4] states that the scanning artefacts, hardscapes and cars are the most challenging classes

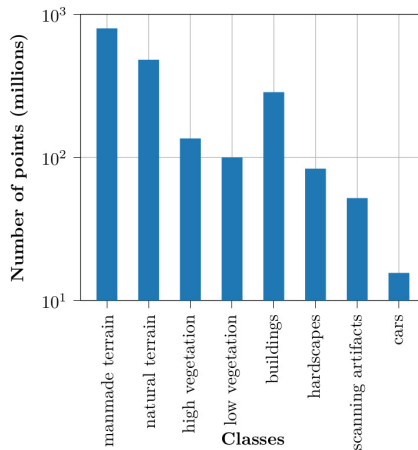


Figure 10: Graph depicting the number of points per class (in millions) in the Semantic3D dataset.

Semantic3D



Figure 11: Illustration of the Semantic3D point clouds of various outdoor scenes. Dataset from [4].

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



Figure 12: Illustration of the S3DIS point clouds of various indoor scenes. Dataset from [1].

OOD Benchmark datasets

| ID dataset | OOD dataset | OOD detection difficulty | Summary |
|-------------|---------------------------|--------------------------|--|
| Semantic 3D | S3DIS | Easy | <ul style="list-style-type: none">• No class overlap• Less structural similarity• Different domain(outdoor-vs-indoor) |
| | Semantic3D without colour | Hard | <ul style="list-style-type: none">• Same structural properties• Difference in RGB values• Same domain as ID dataset• Same classes |

Table 2: Table representing the ID dataset and corresponding OOD datasets, difficulty in OOD detection and the 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-vs-S3DIS & Semantic3D-vs-Semantic3D without colour

1. Introduction

2. Experimental Setup & Methodology

3. Experiments & Results

4. Conclusion



Experiments

- Semantic3D-vs-S3DIS
 - Deep Ensembles
 - Flipout
 - Area Under Receiver Operating Characteristic (AUROC) score comparison
- Semantic3D-vs-Semantic3D without colour

Semantic3D-vs-S3DIS - Deep Ensembles

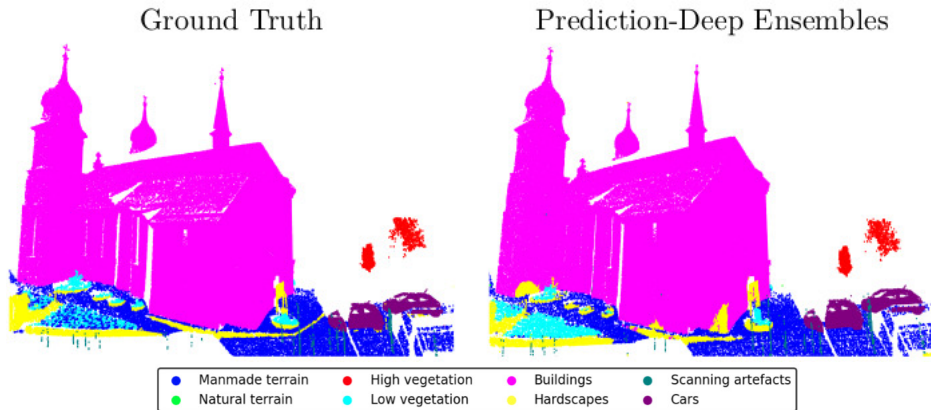


Figure 13: Image representing the predictions (last column) from Deep Ensemble with an ensemble size of 15 on Semantic3D dataset. The first column depict the ground truth.

Semantic3D-vs-S3DIS - Deep Ensembles

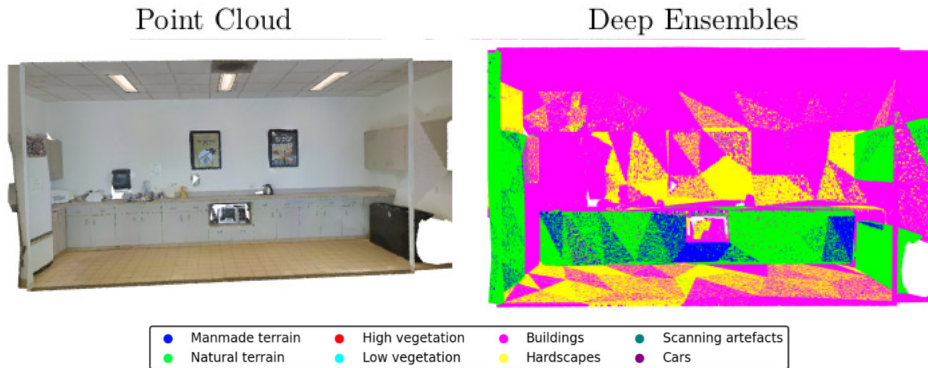


Figure 14: Predictions of RandLA-Net on S3DIS (OOD) dataset. First column representing the point cloud, second column presenting the predictions of Deep Ensembles (15 Ensemble size).

Semantic3D-vs-S3DIS - Deep Ensembles

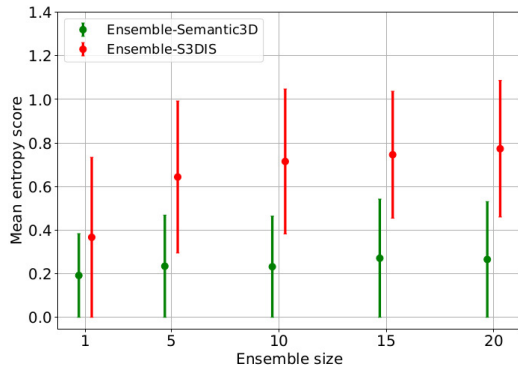
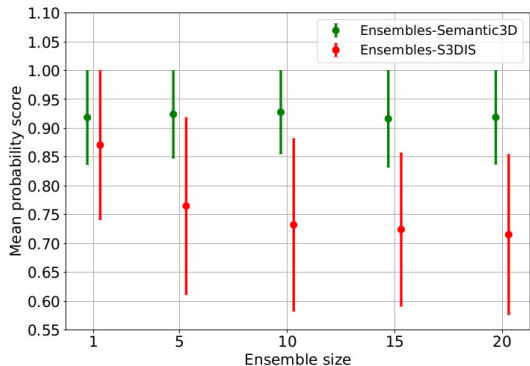


Figure 15: Graphs representing the mean probability value and mean entropy as a dot for Semantic3D (ID) in green and S3DIS (OOD) in red when using Deep Ensembles. The variance is represented via the error bars.

Semantic3D-vs-S3DIS - Flipout

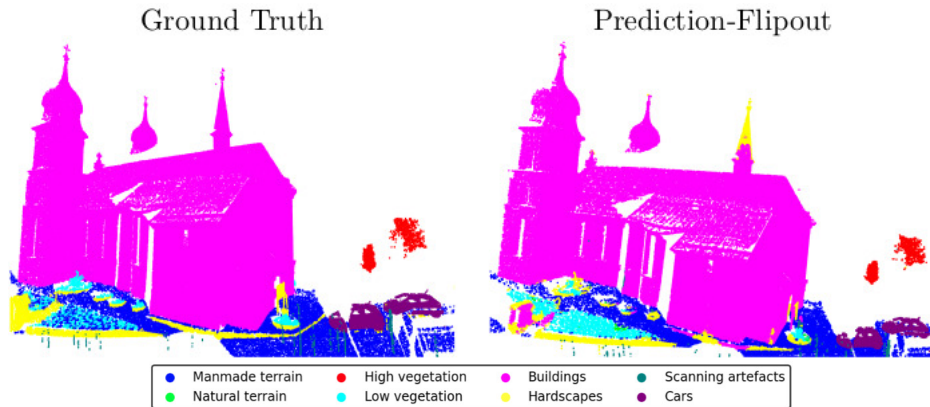
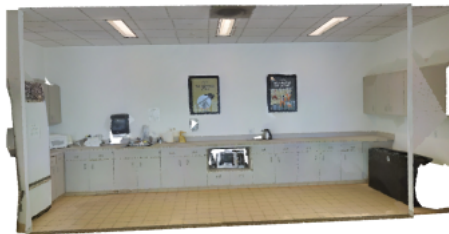


Figure 16: Image representing the predictions (last column) from Flipout with 15 number of passes on Semantic3D dataset. The first column depict the ground truth.

Semantic3D-vs-S3DIS - Flipout

Point Cloud



Flipout



| | | | |
|-------------------|-------------------|--------------|----------------------|
| ● Manmade terrain | ● High vegetation | ● Buildings | ● Scanning artefacts |
| ● Natural terrain | ● Low vegetation | ● Hardscapes | ● Cars |

Figure 17: 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

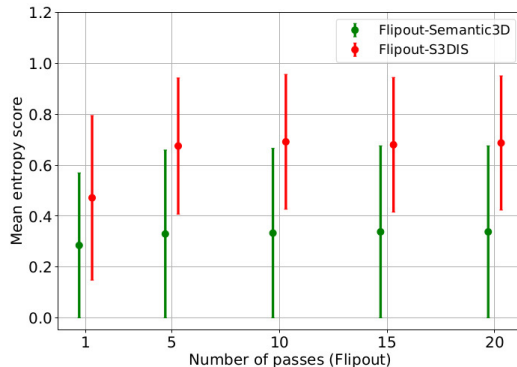
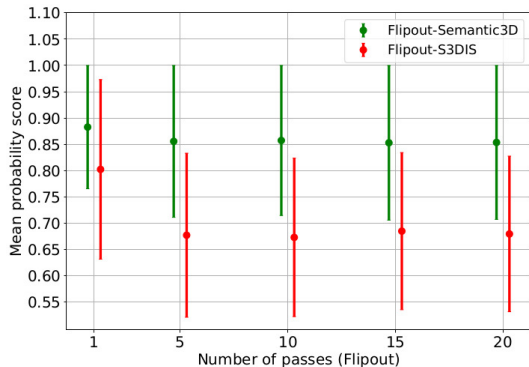


Figure 18: Graphs representing the mean probability value and mean entropy as a dot for Semantic3D (ID) in green and S3DIS (OOD) in red when using Flipout. The variance is represented via the error bars.

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

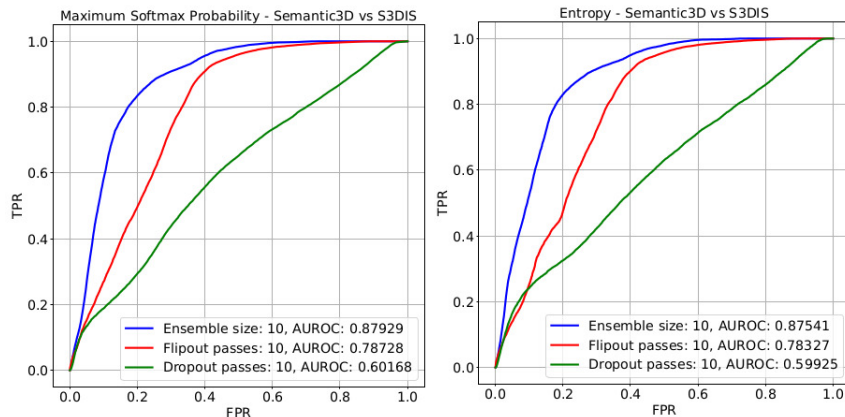


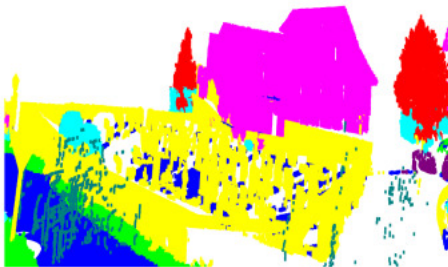
Figure 19: ROC curves of Semantic3D-vs-S3DIS for 10 Ensembles, 10 forward passes for Flipout and Dropout using Maximum Softmax Probability and Entropy respectively.

Experiments

- Semantic3D-vs-S3DIS
- Semantic3D-vs-Semantic3D without colour
 - Deep Ensembles
 - Flipout
 - Area Under Receiver Operating Characteristic (AUROC) score comparison

Semantic3D colour-vs-without colour - Deep Ensembles

Prediction(Semantic3D)



Prediction(Semantic3D without colour)

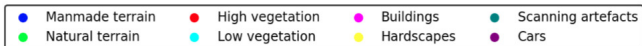
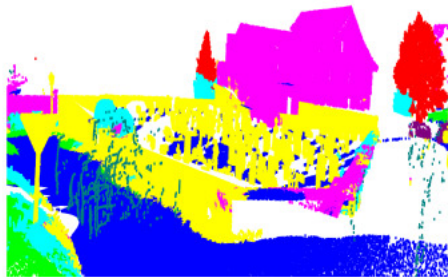


Figure 20: Output predictions of the RandLA-Net over the Semantic3D dataset and Semantic3D without colour dataset using Deep Ensembles (Ensemble size of 10).

Semantic3D colour-vs-without colour - Deep Ensembles

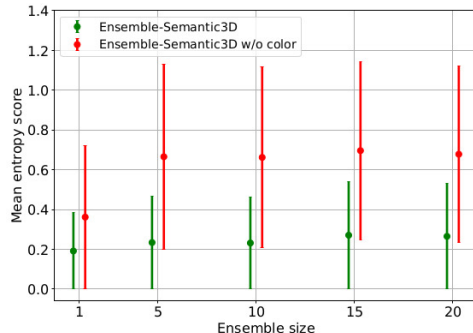
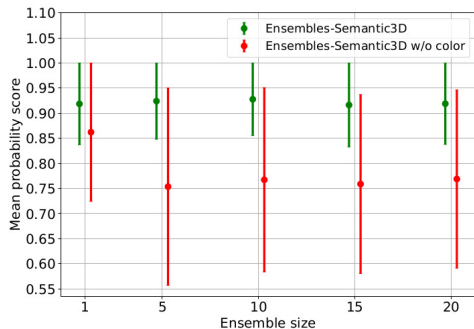


Figure 21: Graphs representing the mean probability value and mean entropy as a dot for Semantic3D (ID) in green and Semantic3D without colour (OOD) in red when using Deep Ensembles. The variance is represented via the error bars.

Semantic3D colour-vs-without colour - Flipout

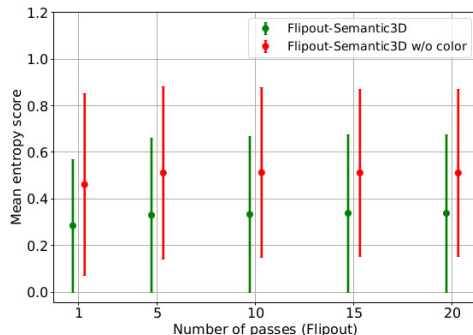
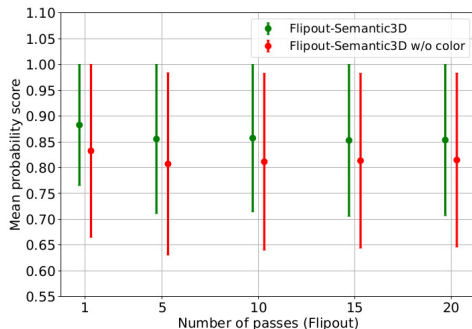


Figure 22: Graphs representing the mean probability value and mean entropy as a dot for Semantic3D (ID) in green and Semantic3D without colour (OOD) in red when using Flipout. The variance is represented via the error bars.

Semantic3D colour-vs-without 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-without colour - AUROC Scores

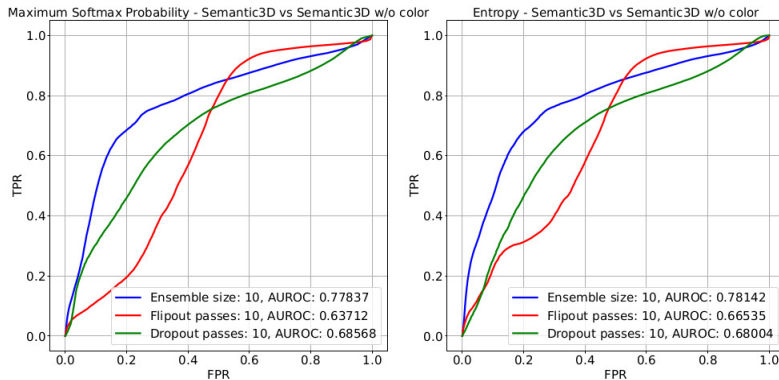


Figure 23: ROC curves of Semantic3D-vs-Semantic3D without colour for 10 ensembles, 10 forward passes for Flipout and Dropout using Maximum Softmax Probability and Entropy scores respectively.

1. Introduction

2. Experimental Setup & Methodology

3. Experiments & Results

4. Conclusion



Conclusion

- We propose two datasets for OOD benchmarking
 - Semantic3D-vs-S3DIS (Outdoor-vs-Indoor) - Easy OOD identification
 - Semantic3D-vs-Semantic3D without colour - Hard OOD identification
- The second case is hard because 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

- Training and evaluation of 3D DNNs are time-consuming and resource-intensive
- Finding the proper prior for Flipout layers is hard
- LiDAR datasets have large memory requirements especially for pre-processing and metric computation
- Getting 100% OOD detection performance is not possible with the post-hoc methods

Future Work

- This thesis is limited to only point-based models, this can be extended to graph and projection-based models
- 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
- Similar post-hoc methods like Mahalanobis distance-based OOD detection [6] or MetaSeg [8] can be added as an extension to this thesis

References (1/4)

 Iro Armeni, Ozan Sener, Amir R. Zamir, Helen Jiang, Ioannis Brilakis, Martin Fischer, and Silvio Savarese.

3d semantic parsing of large-scale indoor spaces.



In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2016.

 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.

References (2/4)

-  Biao Gao, Yancheng Pan, Chengkun Li, Sibogeng, 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 (3/4)

 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 (4/4)

-  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.
-  Jingkang Yang, Kaiyang Zhou, Yixuan Li, and Ziwei Liu.
Generalized out-of-distribution detection: A survey.
arXiv preprint arXiv:2110.11334, 2021.

Semantic3D-Deep Ensembles

| Ensemble size | meanIoU | IoU per-class | | | | | | | | Accuracy |
|---------------|--------------|---------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|
| | | C1 | C2 | C3 | C4 | C5 | C6 | C7 | C8 | |
| 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

| #Passes | MeanIoU | IoU per-class | | | | | | | | Accuracy |
|---------|---------|---------------|-------|-------|-------|-------|-------|-------|-------|----------|
| | | C1 | C2 | C3 | C4 | C5 | C6 | C7 | C8 | |
| 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.

OOD vs Anomaly

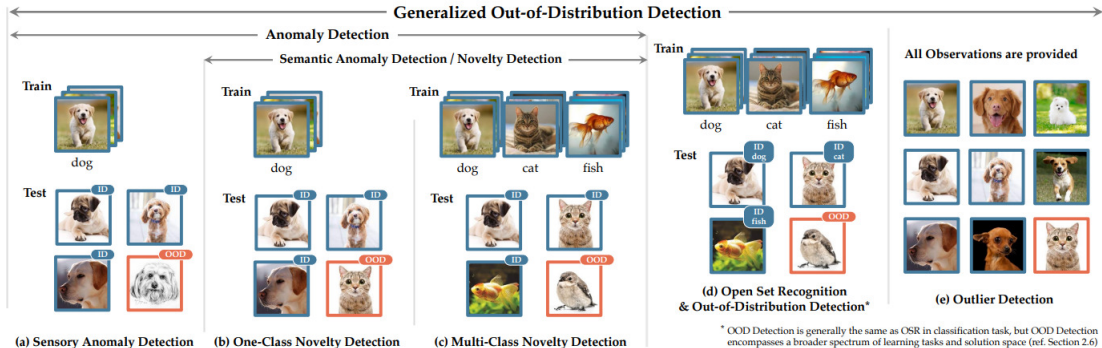


Figure 24: Image representing OOD vs Anomaly vs Novelty detection. Image taken from [9].

OOD vs Anomaly



(a)



(b)



(c)



(d)



(e)



(f)

Figure 2.5: Illustration of Distributional Shift, Anomaly and Out-of-Distribution examples using various kind of ships. (a) represents the sail ship during 18th century. (b) depicts the current training data. (c), and (d) represents the anomalous ship data and (e), and (f) represents the OOD data. Images are taken from [75], [36], [58], [76], [24], and [84] respectively in the order they appear.