
A Benchmark for Out of Distribution Detection in Point Cloud 3D Semantic Segmentation

Lokesh Veeramacheni¹, Matias Valdenegro-Toro²

¹ Bonn-Rhein-Sieg University of Applied Sciences, 53757 Sankt Augustin, Germany.

² Department of AI, University of Groningen, 9747 AG Groningen, The Netherlands.

lokesh.Veeramacheni@smail.inf.h-brs.de, m.a.valdenegro.toro@rug.nl

Abstract

Safety-critical applications like autonomous driving use Deep Neural Networks (DNNs) for object detection and segmentation. The DNNs fail to predict when they observe an Out-of-Distribution (OOD) input leading to catastrophic consequences. Existing OOD detection methods were extensively studied for image inputs but have not been explored much for LiDAR inputs. So in this study, we proposed two datasets for benchmarking OOD detection in 3D semantic segmentation. We used Maximum Softmax Probability and Entropy scores generated using Deep Ensembles and Flipout versions of RandLA-Net as OOD scores. We observed that Deep Ensembles outperform Flipout model in OOD detection with greater AUROC scores for both datasets.

1 Introduction

Deep Neural Networks (DNNs) have made tasks such as object detection and classification easy to solve, allowing deployment in various real-world scenarios such as autonomous driving, and robotic surgery. These networks are usually trained and evaluated on similar datasets. This is called the closed world assumption, where the train and test sets have the same classes and similar distributions for inputs and labels. But this assumption often does not hold in practice, requiring out of distribution detection, in order for safe use of perception models in the real world. If a model is uncertain, it can abstain to provide an answer and could trigger additional processing or release control to a human.

Semantic segmentation is a popular way to understand a scene. Since autonomous vehicles and other kinds of robots operate in 3D environment, point clouds are preferred to perform perception, leading into 3D semantic segmentation, where each point in the cloud receives a class label.

This paper introduces the problem of out of distribution detection in 3D semantic segmentation tasks. We propose a benchmark using two well known LiDAR datasets [7] (Semantic3D and S3DIS) in order to benchmark out of distribution detection capabilities. We initially benchmark uncertainty-based methods, namely Dropout, Deep Ensembles, and Flipout, using RandLA-Net [11] as model for 3D semantic segmentation. Our benchmark contains two dataset combinations, namely Semantic3D vs S3DIS (Benchmark A), and Semantic3D vs Semantic3D without color information (Benchmark B).

Our initial results indicate that OOD detection in point cloud segmentation is possible, with the best result given by Deep Ensembles (AUROC 0.893) with maximum probability on benchmark A, and 0.773 on Benchmark B.

The contributions of this paper are: a definition of a benchmark for out of distribution detection in semantic segmentation of point clouds, initial benchmark results for uncertainty-based methods (Dropout, Ensembles, Flipout), using both entropy and maximum probability.

2 Related Work

2.1 3D Semantic Segmentation

Existing 3D semantic segmentation can be broadly grouped into three different kinds. The first type includes point-based models where the model directly feeds on a 3D point cloud. Pointnet [20], Pointnet++ [21], SPLATNet [22] and RandLA-Net [11] are few example networks of this type. The second type includes projection-based models where the data is projected onto a 2D range image. SqueezeSegV3 [28], RangeNet++ [19], KPRNet [13] and 3DMiniNet [1] are few example models for this type. Projection-based models also include projecting onto bird eye view projection of point cloud. SalsaNext [4], PolarNet [29] and Cylinder3D [30] are few models are of this subtype.

The final type of model is graph neural networks. Dynamic graph CNN [26] and GACNet [25] are example models using graph neural networks. In this study, we use RandLA-Net [11] which is point based model with fewer parameters and also state-of-the-art performance in point-based methods without expensive operations like kernelization. Moreover, RandLA-Net doesn't require preprocessing like range image computation or farthest point sampling and employs simple random point sampling.

2.2 Out-of-Distribution detection

A dataset is referred to as the OOD dataset if the whole dataset consists of only OOD objects and the training dataset is called In-Distribution (ID) dataset. Multiple approaches exist to generate the scores for OOD detection. These approaches employ a threshold-based method and have no idea of how OOD data is during training. [10] provides a baseline method using Maximum Softmax Probability (MSP) scores for OOD detection. An improved method called ODIN for the baseline method is proposed in [17]. ODIN utilizes calibrated softmax scores along with input noise perturbations making training adversarial. [16] proposed the use of Mahalanobis distance as OOD score instead of softmax, where Mahalanobis distance is calculated between each activation map and multivariate Gaussian distribution. [12] uses the combination of Mahalanobis distance and ODIN for OOD score generation. [23] proposed a method for OOD score generation called ReAct, where the proposed ReAct activation is applied before the softmax layer to suppress the higher activations to a constant.

Based on the fact that the OOD data has higher uncertainty scores when compared to ID data, [14] estimates epistemic uncertainty using Deep Ensembles to classify ID and OOD. Similarly [24] uses epistemic uncertainty calculated from radial basis function to detect OOD data. [5], [15] utilizes Bayesian neural networks for uncertainty estimation and then for OOD detection.

In this paper, we use the MSP score proposed in [10] with uncertainty estimates from Deep Ensembles [14] and Flipout [27] for OOD score generation. Additionally we use Entropy to integrate information from the whole predictive distribution. Both MSP and Entropy metrics are computed from the output predictive distribution, which is the average of ensemble/forward pass outputs of each network. The selection of UQ methods is motivated as these methods are scalable [8] and have been tested to work well for point cloud semantic segmentation [3].

3 Datasets and Benchmark

All the experiments discussed in this paper use Semantic3D proposed in [9] as the In-Distribution dataset. Semantic3D is chosen as the ID dataset because it is one of the dense datasets along with the RGB color. We hypothesize that having RGB colors will help in the improved performance of OOD detection.

The first OOD dataset we used is S3DIS proposed in [2]. We especially chose the S3DIS as the OOD dataset because the dataset consists of indoor objects whereas the ID dataset consists of outdoor objects. This domain difference in scenes makes S3DIS an ideal OOD dataset. We expect the Semantic3D trained RandLA-Net model to detect the S3DIS dataset as OOD with ease and high confidence. This expectation is due to the difference in point geometry between these datasets. We call this combination Benchmark A.

The second OOD dataset we used is Semantic3D without color. Having this dataset as an OOD dataset ensures the same point geometries between ID and OOD datasets but the difference in other

Ensemble size/ #passes	Method	AUROC Benchmark A		AUROC Benchmark B	
		MSP	Entropy	MSP	Entropy
1	Dropout	0.53311	0.53041	0.66349	0.65908
	Flipout	0.69988	0.69368	0.64221	0.66157
	Deep Ensembles	0.62020	0.62529	0.67855	0.67866
5	Dropout	0.58439	0.57821	0.69448	0.68507
	Flipout	0.77885	0.76934	0.63743	0.66536
	Deep Ensembles	0.84013	0.83665	0.76769	0.77120
10	Dropout	0.60168	0.59925	0.68568	0.68004
	Flipout	0.78728	0.78327	0.63712	0.66535
	Deep Ensembles	0.87929	0.87541	0.77837	0.78142
15	Dropout	0.59773	0.59557	0.68975	0.68347
	Flipout	0.7667	0.76741	0.63022	0.65976
	Deep Ensembles	0.88486	0.88246	0.77302	0.77881
20	Dropout	0.59766	0.59661	0.68447	0.68199
	Flipout	0.77331	0.77237	0.63017	0.65934
	Deep Ensembles	0.89338	0.89052	0.77031	0.77584

Table 1: AUROC scores calculated for all the points in the test sets of both datasets. MSP and Entropy values are represented for Deep Ensembles, Flipout and Dropout methods with increase in ensemble size or number of passes with step size of 5. Benchmark A represents Semantic3D vs S3DIS and Benchmark B is Semantic3D vs Semantic3D without color.

point properties like color. This simulates sensor failure. Because of the above reason, we expect RandLA-Net to struggle in detecting this OOD data. We call this combination Benchmark B.

4 Experiments and Results

The experimental setup of RandLA-Net on Deep Ensembles and Flipout is reported in Supplementary Section A. Training results (mean IoU, per-class Iou, Overall Accuracy) are also provided in Supplementary Section C. Here we discuss the results of OOD detection on two proposed datasets using the AUROC score as the evaluation metric.

Benchmark A - Semantic3D vs S3DIS

In this section, we evaluate the OOD detection performance on the Semantic3D-vs-S3DIS dataset with Semantic3D being ID and S3DIS being the OOD dataset. We use the AUROC scores generated using Maximum Softmax Probability and Entropy from Deep Ensembles, Flipout and Dropout techniques. From AUROC (Dataset1) column in Table 1, we observe that the Deep Ensembles outperform Flipout and Dropout models in OOD detection, with Flipout only outperforming the other two methods in the case of a single ensemble/forward pass, which indicates an advantage of the Gaussian weight distribution learned using Flipout. We also observe that after the Ensemble size or number of passes 10, the performance improvements are little to none.

Figure 1a and 1b depict the ID points represented in green and OOD points represented in red for both Semantic3D (ID) and S3DIS (OOD) datasets respectively. Here we observe that a few points such as the edges of the church in Figure 1a are classified with a low probability score and these points are also classified as OOD points in the ID dataset. Similarly, a few points in S3DIS dataset in Figure 1b are also classified as a part of ID dataset (green points). These green points are because of the walls being classified as a building which is partly true. Overall, we observe that the S3DIS dataset has higher OOD points (more reddish) whereas most of the ID dataset is greenish in color.

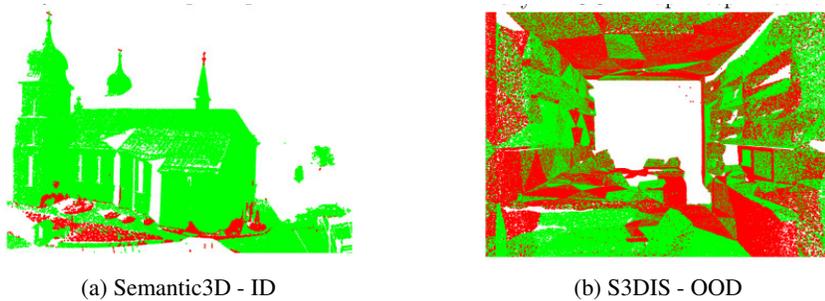


Figure 1: Images depicting the ID points in green and OOD points in red for Semantic3D (ID) dataset in (a) and S3DIS (OOD) dataset in (b). ID-OOD classification is made using Maximum Softmax Probability values generated from Deep Ensembles with ensemble size of 10.

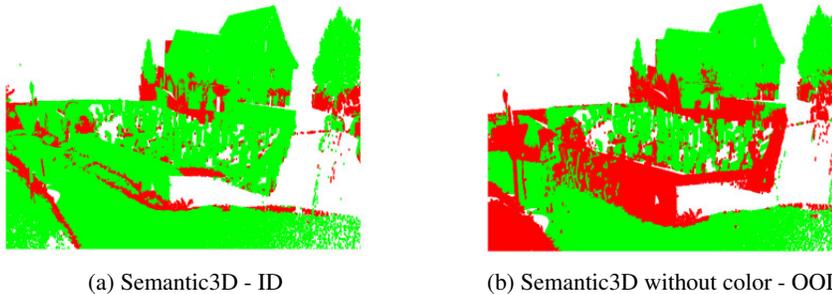


Figure 2: Images depicting the ID points in green and OOD points in red for Semantic3D (ID) dataset in (a) and Semantic3D without color (OOD) dataset in (b). ID-OOD classification is made using Maximum Softmax Probability values generated from Deep Ensembles with ensemble size of 10.

Benchmark B - Semantic3D vs Semantic3D without color

Similarly, the second set of AUROC columns in Table 1 represents the AUROC scores generated using MSP and Entropy for the second OOD benchmark. Deep Ensembles outperform Flipout and Dropout in this dataset also. In both the dataset performances, we observe that the AUROC scores for the Entropy and MSP are similar. Performance of OOD detection maxes out at ensemble size or the number of passes of 10 and no improvements are observed further increase in size. Figures 2a and 2b depict the ID and OOD points (green and red color respectively) for Semantic3D and Semantic3D without color respectively. Most of the misclassified points in the ID dataset are classified as OOD points as shown in Figure 2a. In Semantic3D without color (OOD dataset) we majorly observe that walls and low vegetation are prone to misclassify as buildings and manmade terrain because no color information is available to differentiate between them. So most of the points belonging to walls and low vegetation are detected as OOD points.

5 Conclusions

In this paper, we studied the performance of OOD detection in 3D semantic segmentation, by proposing two benchmarks for OOD benchmarking, one being Semantic3D-vs-S3DIS and the other dataset being Semantic3D-vs-Semantic3D without color. We trained a RandLA-Net model for 3D semantic segmentation on the Semantic3D dataset and ran inference on S3DIS and Semantic3D without color datasets individually. Overall, we observe that the OOD detection performance using Deep Ensembles is better than two other Bayesian methods. Finally, we conclude that OOD detection is relatively easy in case the OOD objects varying point geometry compared to training data. In case the OOD object has similar point geometry to training data, it is challenging as the RandLA-Net model hugely relies on point geometries. Additionally our results show that OOD detection is still challenging even in a simple setup like Benchmark A, requiring large computational costs. Future work should produce lightweight uncertainty quantification methods.

References

- [1] Iñigo Alonso, Luis Riazuelo, Luis Montesano, and Ana C. Murillo. 3d-mininet: Learning a 2d representation from point clouds for fast and efficient 3d lidar semantic segmentation. *IEEE Robotics and Automation Letters*, 5(4):5432–5439, 2020.
- [2] 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.
- [3] Swaroop Bhandary K, Nico Hochgeschwender, Paul Plöger, Frank Kirchner, and Matias Valdenegro-Toro. Evaluating uncertainty estimation methods on 3d semantic segmentation of point clouds. *arXiv e-prints*, pages arXiv–2007, 2020.
- [4] Tiago Cortinhal, George Tzelepis, and Eren Erdal Aksoy. Salsanext: Fast, uncertainty-aware semantic segmentation of lidar point clouds. In George Bebis, Zhaozheng Yin, Edward Kim, Jan Bender, Kartic Subr, Bum Chul Kwon, Jian Zhao, Denis Kalkofen, and George Baciuc, editors, *Advances in Visual Computing*, pages 207–222, Cham, 2020. Springer International Publishing.
- [5] Francesco D’Angelo and Christian Henning. Uncertainty-based out-of-distribution detection requires suitable function space priors. *CoRR*, abs/2110.06020, 2021.
- [6] Joshua V Dillon, Ian Langmore, Dustin Tran, Eugene Brevdo, Srinivas Vasudevan, Dave Moore, Brian Patton, Alex Alemi, Matt Hoffman, and Rif A Saurous. Tensorflow distributions. arxiv 2017. *arXiv preprint arXiv:1711.10604*.
- [7] 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.
- [8] Fredrik K Gustafsson, Martin Danelljan, and Thomas B Schon. Evaluating scalable bayesian deep learning methods for robust computer vision. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 318–319, 2020.
- [9] 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.
- [10] Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *arXiv preprint arXiv:1610.02136*, 2016.
- [11] 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.
- [12] Ryo Kamoi and Kei Kobayashi. Why is the mahalanobis distance effective for anomaly detection? *CoRR*, abs/2003.00402, 2020.
- [13] Deyvid Kochanov, Fatemeh Karimi Nejadasl, and Olaf Booij. Kprnet: Improving projection-based lidar semantic segmentation. *arXiv preprint arXiv:2007.12668*, 2020.
- [14] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. *arXiv preprint arXiv:1612.01474*, 2016.
- [15] Jinsol Lee and Ghassan AlRegib. Gradients as a measure of uncertainty in neural networks. In *IEEE International Conference on Image Processing, ICIP*, pages 2416–2420. IEEE, 2020.
- [16] 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.
- [17] Shiyu Liang, Yixuan Li, and Rayadurgam Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks. *arXiv preprint arXiv:1706.02690*, 2017.
- [18] Veriscian Pty Ltd. Cloud services for 3d scanning, -. [Online; accessed April 10, 2022].
- [19] Andres Milioto, Ignacio Vizzo, Jens Behley, and Cyrill Stachniss. Rangenet ++: Fast and accurate lidar semantic segmentation. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 4213–4220, 2019.

- [20] 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 (CVPR)*, July 2017.
- [21] Charles R 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.
- [22] Hang Su, Varun Jampani, Deqing Sun, Subhransu Maji, Evangelos Kalogerakis, Ming-Hsuan Yang, and Jan Kautz. Splatnet: Sparse lattice networks for point cloud processing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.
- [23] Yiyou Sun, Chuan Guo, and Yixuan Li. React: Out-of-distribution detection with rectified activations. *CoRR*, abs/2111.12797, 2021.
- [24] Joost van Amersfoort, Lewis Smith, Yee Whye Teh, and Yarin Gal. Uncertainty estimation using a single deep deterministic neural network. In *the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 9690–9700. PMLR, 2020.
- [25] Lei Wang, Yuchun Huang, Yaolin Hou, Shenman Zhang, and Jie Shan. Graph attention convolution for point cloud semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.
- [26] Yue Wang, Yongbin Sun, Ziwei Liu, Sanjay E. Sarma, Michael M. Bronstein, and Justin M. Solomon. Dynamic graph cnn for learning on point clouds. *ACM Transactions On Graphics (TOG)*, 38(5), October 2019.
- [27] Yeming Wen, Paul Vicol, Jimmy Ba, Dustin Tran, and Roger B. Grosse. Flipout: Efficient pseudo-independent weight perturbations on mini-batches. In *6th International Conference on Learning Representations, ICLR*, 2018.
- [28] Chenfeng Xu, Bichen Wu, Zining Wang, Wei Zhan, Peter Vajda, Kurt Keutzer, and Masayoshi Tomizuka. Squeezesegv3: Spatially-adaptive convolution for efficient point-cloud segmentation. In *European Conference on Computer Vision*, pages 1–19. Springer, 2020.
- [29] Yang Zhang, Zixiang Zhou, Philip David, Xiangyu Yue, Zerong Xi, Boqing Gong, and Hassan Foroosh. Polarnet: An improved grid representation for online lidar point clouds semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.
- [30] Xinge Zhu, Hui Zhou, Tai Wang, Fangzhou Hong, Yuexin Ma, Wei Li, Hongsheng Li, and Dahua Lin. Cylindrical and asymmetrical 3d convolution networks for lidar segmentation. *arXiv preprint arXiv:2011.10033*, 2020.

Appendix

A Experimental Setup

Deep Ensembles

We trained 20 instances of RandLA-Net by the same procedure as the authors described in [11]. The only change is made during the inference, we changed the inference pipeline to infer over all the points in the test set. The outputs of these 20 instances are averaged to extract the entropy and MSP values.

Flipout

We changed the last three classification layers of the RandLA-Net model to Flipout compatible using Tensorflow-probability API [6]. After exhaustive trial and error, the prior for Flipout layers is chosen as normal prior with unit standard deviation and zero mean. The same training hyperparameters as Deep Ensembles are used to train the Flipout style RandLA-Net. We used 20 inferences from this network to extract the entropy and MSP values.

Dropout

The RandLA-Net architecture proposed in [11] has a Dropout layer at the end of the classification layers. We also used the same setup with 0.5 as Dropout probability. Similar to the Flipout model we used 20 inferences to extract entropy and MSP values.

B OOD Detection Evaluation - AUROC Computation

In this section we describe how AUROC is computed for out of distribution detection in point cloud segmentation predictions.

We first computed the softmax predictions on each point in the test set of the ID dataset and similarly softmax predictions on each point in the test set of the OOD dataset. Then we computed the MSP and Entropy values from each point predictions in both ID and OOD datasets. These scores are used to compute AUROC metrics.

The computed per point MSP and Entropy values are used to compute the AUROC using the sklearn API. Each point in a ID point cloud receives a label $y = 0$, while each point in an OOD point cloud receives a label $y = 1$. AUROC is computed between all the point clouds in a given dataset combination. This is done because the number of points in each cloud is large, but the number of point cloud samples in each dataset is relatively small.

This means that the AUROC results presented in our paper indicate the ability of a model to discriminate individual points in the cloud as in-distribution or out of distribution. This is slightly different than the standard case of classification, where a whole sample is classified in the same way.

C RandLA-Net Evaluation

Deep Ensembles

In this section, we discuss the training results of RandLA-Net over 20 Deep Ensembles. Table 2 enumerates the meanIoU, per-class IoU, and Accuracy of RandLA-Net up to an ensemble size of 20. From the table, we observe that there is an improvement in performance with the increase in ensemble size. Also after the ensemble size of 10, the gains are very little to none. Figure 3 depicts the predictions of RandLA-Net with an ensemble size of 15. Typically, we observed that there is some amount of misclassification along the edges of the church, and trees.

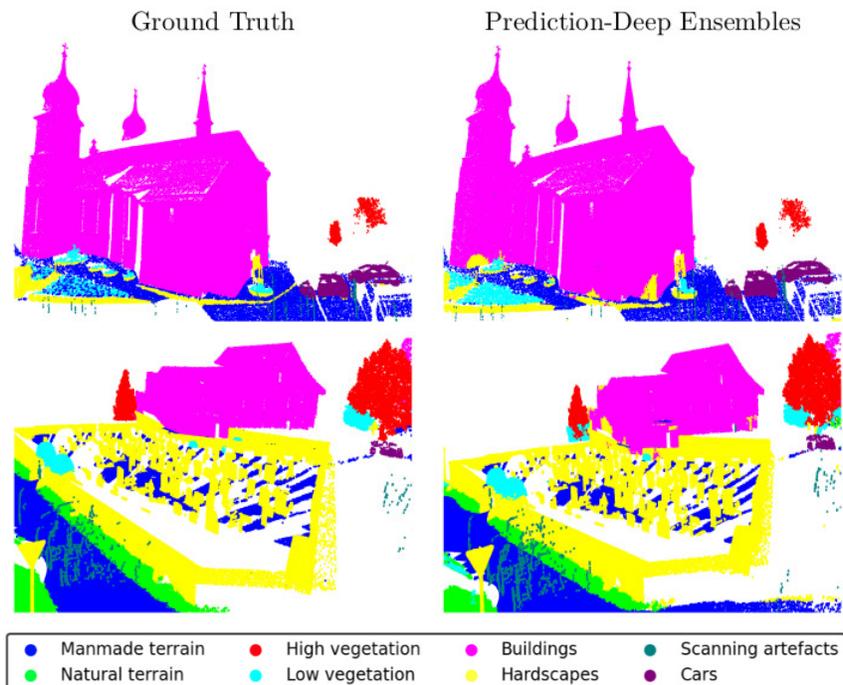


Figure 3: Predictions of RandLA-Net Deep Ensembles on Semantic3d dataset with ground truth in left column and predictions on right column. Predictions are computed with an ensemble size of 15.

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 2: 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.

Flipout

The following three layers highlighted in the red box in Figure 5 are changed to Flipout compatible. Table 3 depicts the mean IoU, per-class IoU and Accuracy of the Semantic3D trained Flipout style RandLA-Net. Even though the meanIoU, Accuracy and most of the classes IoU are similar to the results produced in [11]¹. Here we observe a significant improvement in the Hardscapes class (C6 in table) compared to Deep Ensembles. Unlike Deep Ensembles, with the increase in the number of forward passes we do not observe any performance improvements and judging by meanIoU and Accuracy, one can say that there is a slight decrease in overall performance with increase in the number of passes.

¹Since Semantic3D is an ongoing benchmark challenge, the test set is not public. We separated a part of the training set and used it as a test set. For reference, we evaluated the RandLA-Net on this test set with the weights provided by the authors.

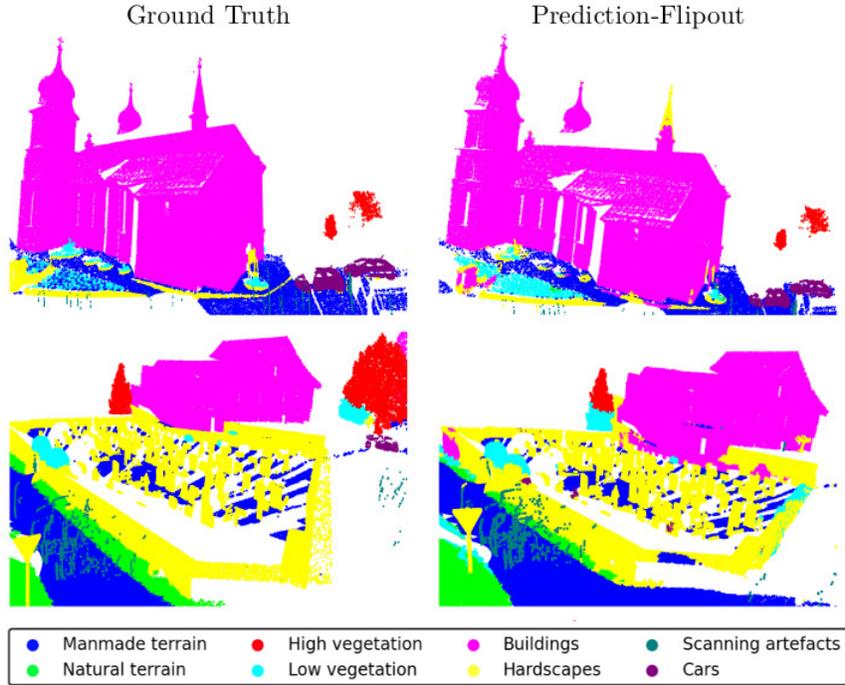


Figure 4: Predictions of Flipout style RandLA-Net on Semantic3d dataset with ground truth in left column and predictions on right column. Predictions are computed with 15 passes.

#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 3: Illustration of performance of Flipout style 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.

D OOD Dataset Predictions

Figure 6 shows the predictions of the RandLA-Net model on the S3DIS dataset (first OOD dataset). The first column represents the predictions on Deep Ensembles size of 15 and the second column with predictions from the Flipout model with 15 forward passes. In the case of Deep Ensembles, we

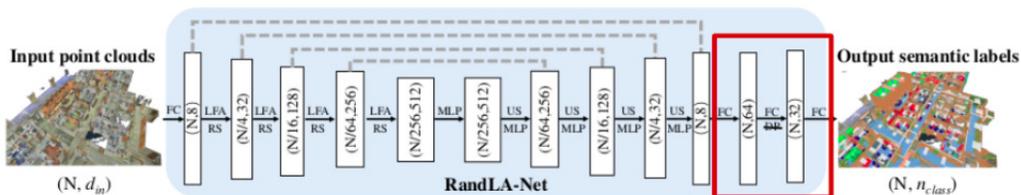


Figure 5: Flipout style RandLA-Net where the last three FC layers as depicted in red box are made Flipout-compatible.

observe that most of the walls are segmented as the building which is partly true but the other objects like cabinets, printer/wall posters, and chairs are misclassified. Whereas in the case of the Flipout model, most of the points are predicted as Hardscapes but both the Deep Ensembles and Flipout classify the cabinets and printer/wall posters as low vegetation. This is mostly because the feature vectors are near the feature vector of low vegetation. A prominent observation is that the predictions on the OOD dataset are triangular this is because of the inherent property of the data. S3DIS dataset is generated using a matterport scanner and according to [18] the matterport camera first generates a triangular mesh and then the points are extracted from these meshes.

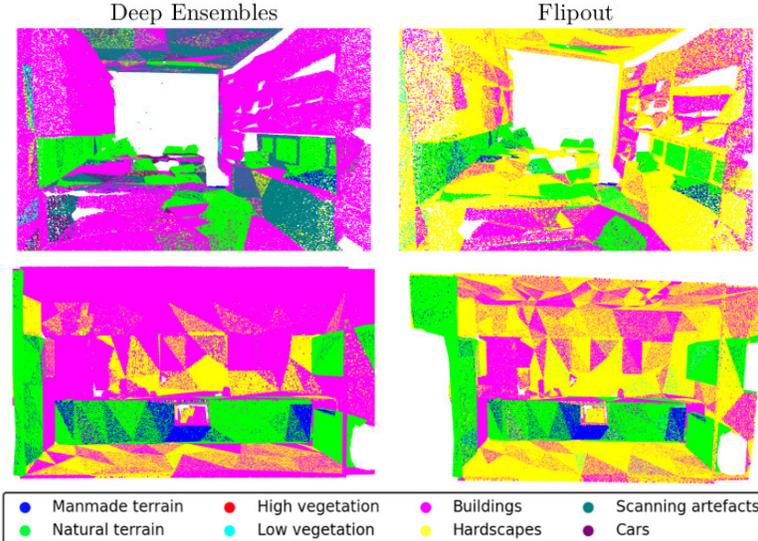


Figure 6: Predictions of RandLA-Net over S3DIS (OOD) dataset with first column using Deep Ensembles and second column using Flipout. Predictions are computed using ensemble size and number of passes as 15.

E ID-OOD Maps - Flipout

In this section, we present the ID and OOD points classified using threshold from MSP values extracted using Flipout style RandLA-Net with 10 passes. Figures 7a and 7b represent the ID points in green and OOD points in red for the first dataset (Semantic3D-vs-S3DIS). Similarly Figures 8a and 8b represent ID and OOD points for the second dataset.

F OOD Detection ROC Curves

In this section, we depict the ROC curves along with the optimal threshold MSP and Entropy values extracted from ROC curves. For both datasets, we use ensemble size and the number of passes of 10.

Semantic3D-vs-S3DIS

Figure 9 depicts the ROC curves for both MSP and Entropy for the first OOD dataset respectively. We observe that the Deep Ensembles' ROC curve is higher than other methods. Table 4 represents the extracted optimal thresholds from the above ROC curves.

Semantic3D-vs-Semantic3D without color

Similarly, Figure 10 represents the ROC curves for the second OOD dataset along with thresholds represented in Table 5. Here we observe that the ROC curves are smaller than the first OOD dataset. Also, the thresholds for MSP are higher and for Entropy are much lower. This kind of behaviour is expected because Semantic3D without color dataset has the same point geometry as the ID dataset.

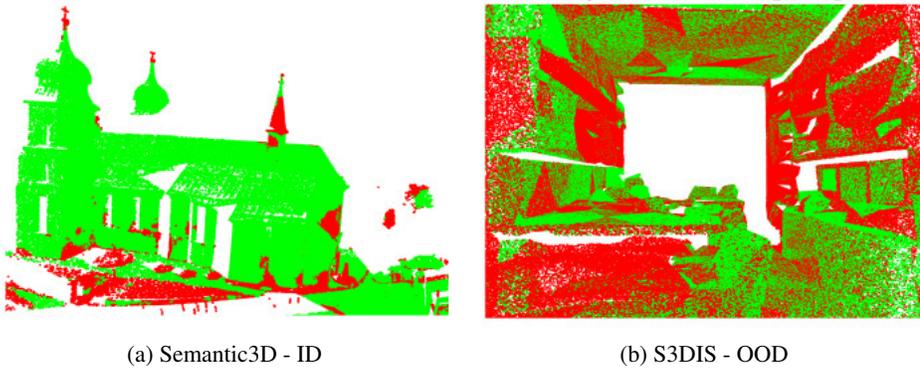


Figure 7: Images depicting the ID points in green and OOD points in red for Semantic3D (ID) dataset in (a) and S3DIS (OOD) dataset in (b). ID-OOD classification is made using Maximum Softmax Probability values generated from Flipout with 10 passes.

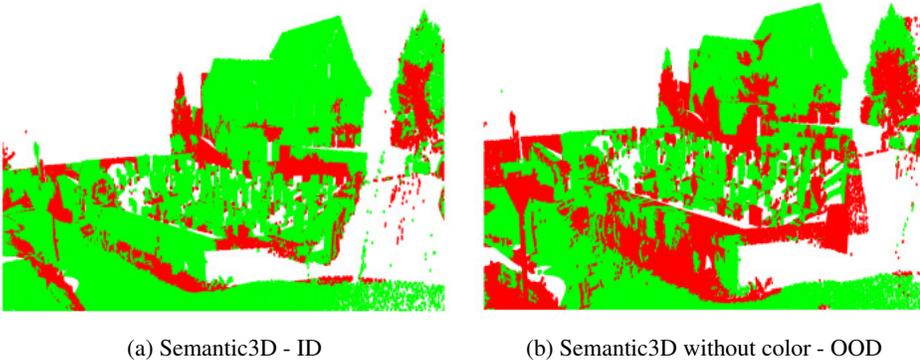


Figure 8: Images depicting the ID points in green and OOD points in red for Semantic3D (ID) dataset in (a) and Semantic3D without color (OOD) dataset in (b). ID-OOD classification is made using Maximum Softmax Probability values generated from Flipout with 10 passes

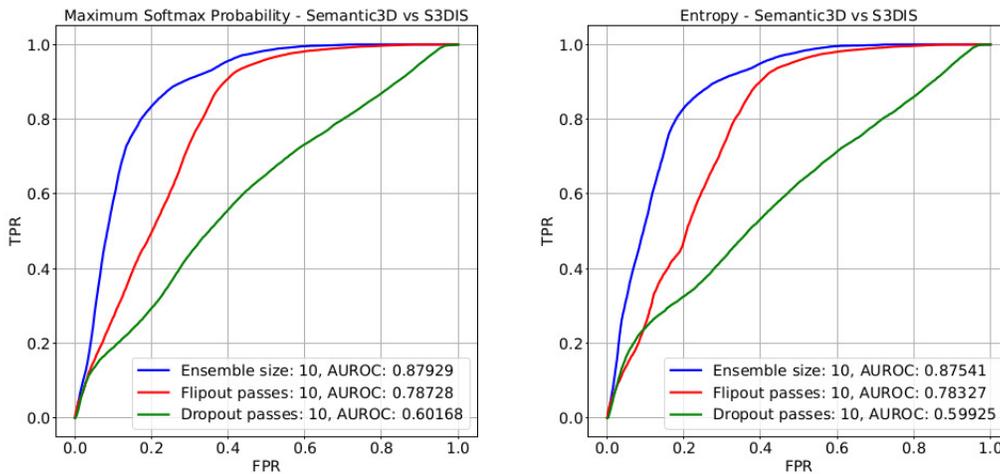


Figure 9: ROC curves of Semantic3D-vs-S3DIS for 10 Ensembles, 10 forward passes for Flipout and Dropout. Left image represents the ROC curves using Maximum Softmax Probability and right image represents ROC curves using Entropy.

OOD Benchmark	Method	MSP	Entropy
Semantic3D-vs-S3DIS	Ensembles	0.755	0.386
	Flipout	0.717	0.439

Table 4: MSP threshold and entropy threshold generated from the corresponding ROC curves for Semantic3D-vs-S3DIS datasets for MSP and entropy methods.

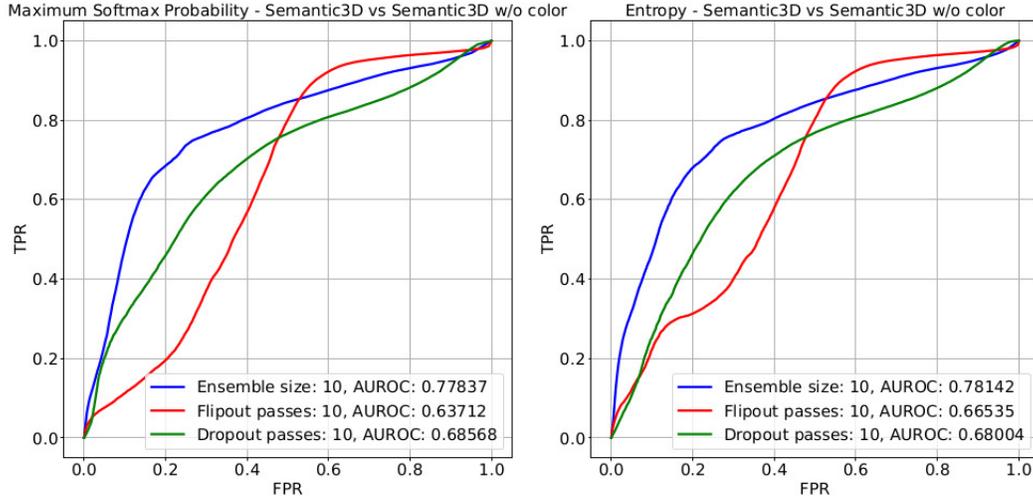


Figure 10: ROC curves of Semantic3D-vs-Semantic3D without color for 10 Ensembles, 10 forward passes for Flipout and Dropout. Left image represents the ROC curves using Maximum Softmax Probability and right image represents ROC curves using Entropy.

Whereas in the previous case, we observe lower thresholds for MSP and higher thresholds for Entropy.

OOD Benchmark	Method	MSP	Entropy
Semantic3D-vs-Semantic3D without color	Ensembles	0.790	0.323
	Flipout	0.796	0.131

Table 5: MSP threshold and entropy threshold generated from the corresponding ROC curves for Semantic3D-vs-Semantic3D without color datasets for MSP and entropy methods.