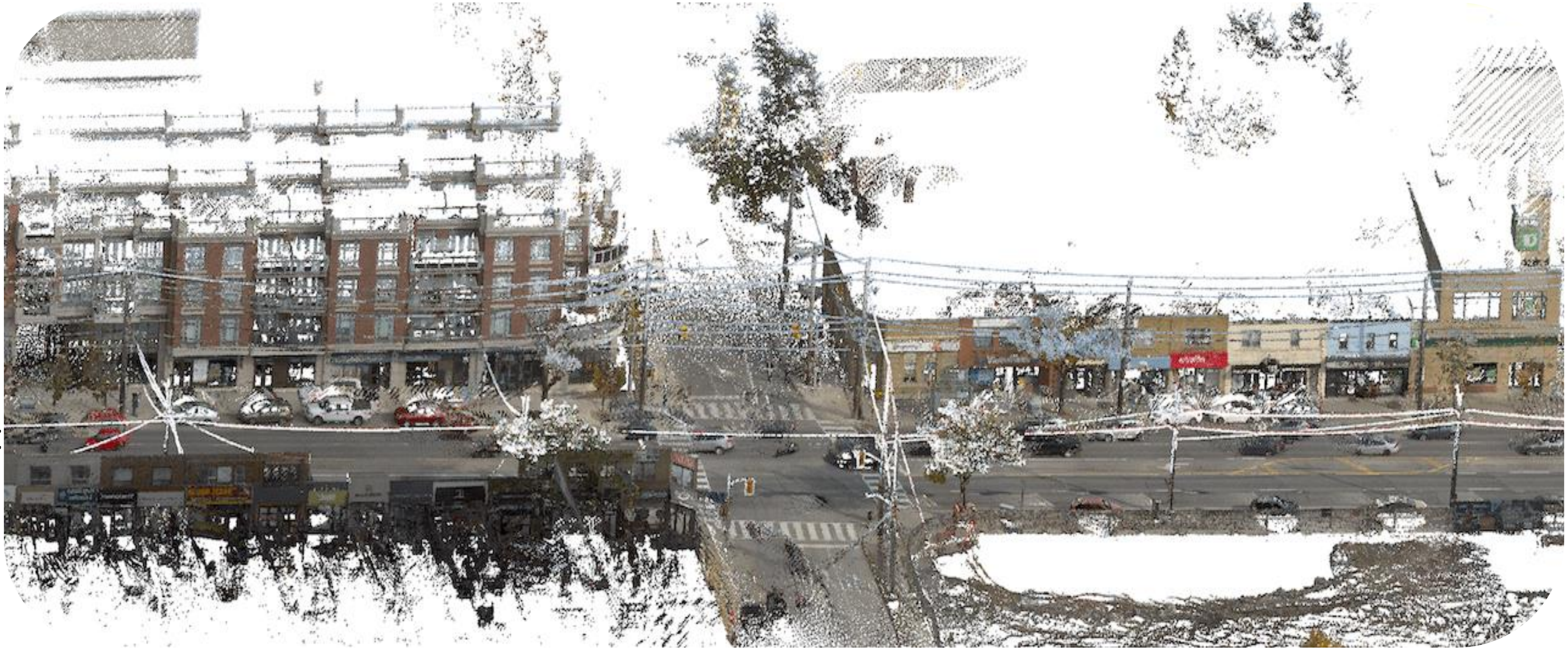
Active Learning for Machine Learning-based Classification of LiDAR Data: Mini Project

Contents

- Introduction
- Data description
- Methodology
- Results
- Classified Output
- References



Supervisor:
Dr.A.M.Ramiya,
Associate Professor

Presented by
R RAMBABU, SC22M075

Introduction

- This mini project aims to evaluate the efficacy of classification of LiDAR data, using the Random Forest (RF) algorithm in conjunction with active learning through the pool-based uncertainty sampling method
- In the LiDAR data classification tasks where labeling point clouds can be a labor-intensive and time-consuming process. It involves manually identifying and labeling different types of classes, such as vegetation, buildings, roads, and water bodies etc.
- Active learning is a semi-supervised machine learning technique that selects the most informative samples to label and add to the training set, with random forest machine learning algorithm as base estimator thereby reducing labeling effort while maintaining high classification accuracy.
- Active learning has broad applications across machine learning tasks, including the classification of satellite and aerial imagery such as hyperspectral and multispectral data, RGB photographic images in computer vision , text classification, and object detection

Dataset description

- The Toronto dataset (L004) is a large-scale urban outdoor point cloud dataset acquired by an MLS system in Toronto, Canada. This dataset covers approximately 0.25 km of road and consists of about 5.4 million points, each with 3D coordinates (x, y, z) and labeled with one of 8 classes[1].
- The labeled classes included categories such as Road (label 1), Road marking (label 2,) Natural (label 3), Building (label 4), Utility line (label 5), Pole (label 6), Car (label 7), Fence (label 8), unclassified (label 0).
- The dataset was highly imbalanced, with some classes having very few samples compared to others. this posed a challenge for training a machine learning model that could accurately classify all classes.
- To address this challenge, we used active learning to iteratively select the most informative examples to label and add to the training set, which helped improve the model's performance on the imbalanced dataset.

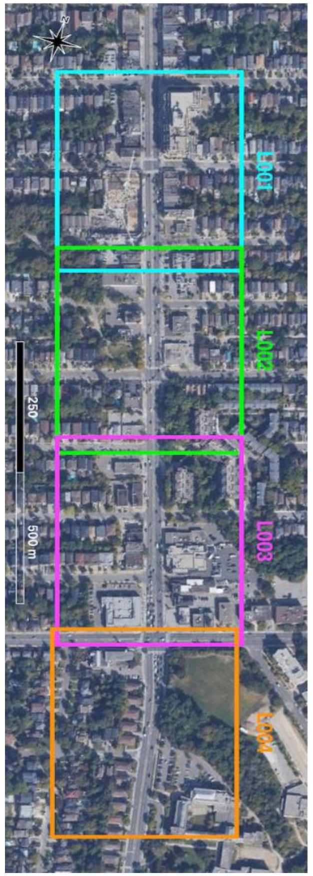


Fig. Toronto dataset (43°43'35.40"N, 79°25'2.30"W)

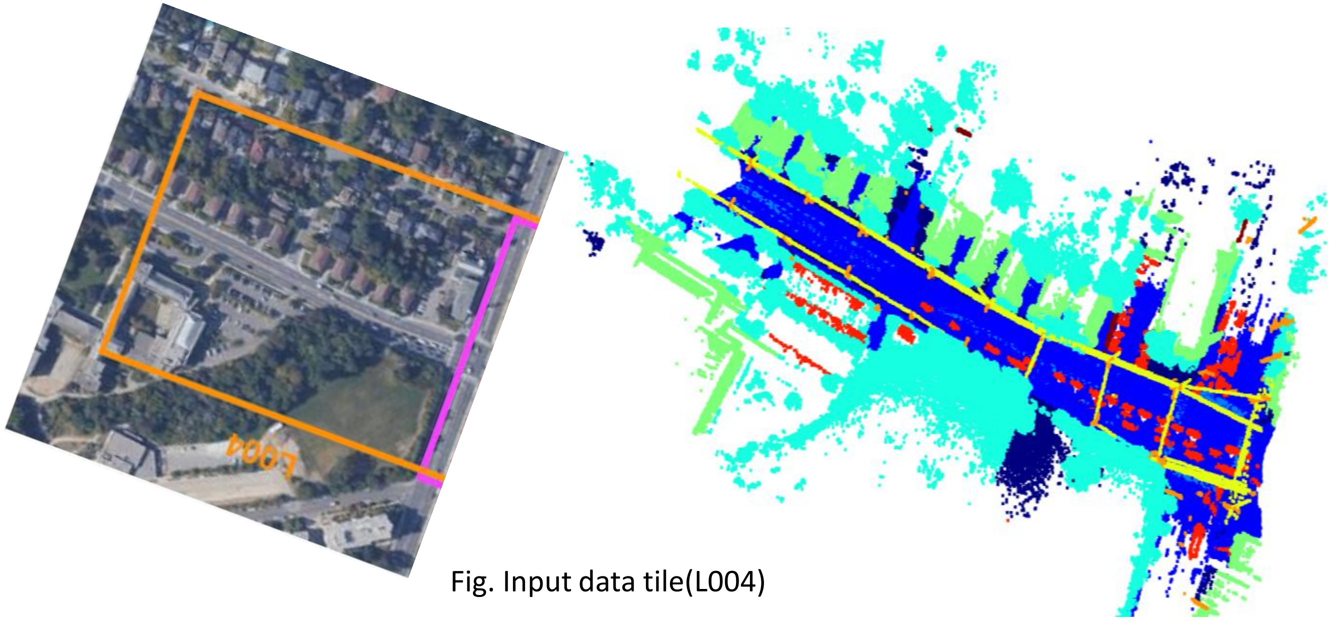


Fig. Input data tile(L004)

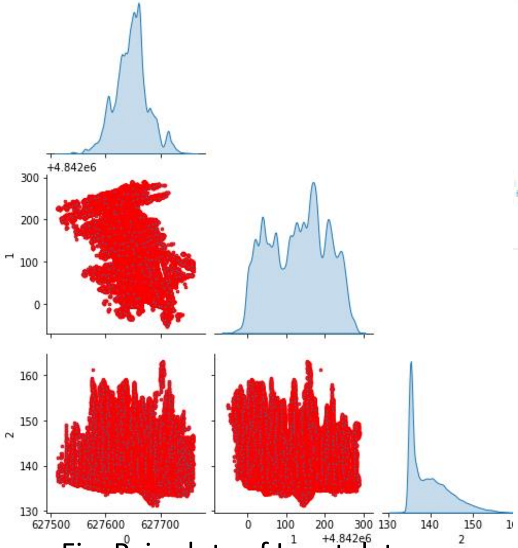


Fig. Pair plots of Input data

```
Element properties:
- x
- y
- z
- red
- green
- blue
- scalar_Intensity
- scalar_GPSTime
- scalar_ScanAngleRank
- scalar_Label
```

Fig. Features of point cloud

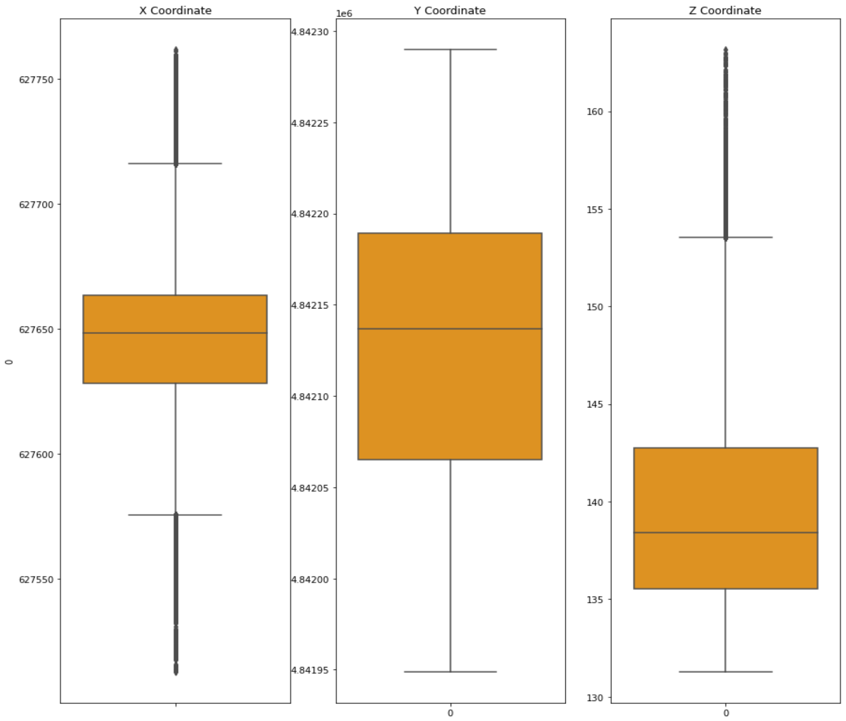


Fig. Box plots of Input data (X,Y,Z)

Methodology

Input data

- Data (Toronto dataset - L004)
- Labels

Initialisation

- Split Labelled, Pool data sets with proportion of 70%,30%
- Initialize the active learner with Random Forest with uncertainty sampling as query strategy with 10 iterations

Querying and Training

- Query the learner for most informative instances to labels
- Label the queried instance and add it to the labeled set and remove the queried instance from the pool set

Training on updated labels

- Reinitialize the learner with the updated labeled set
- Train the learner on the updated labeled set

Prediction on remaining labelled data

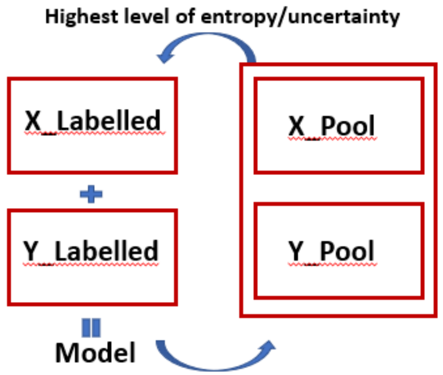
- Use the trained learner to predict the labels of the remaining unlabeled data
- Calculate the accuracy of the learner on the remaining unlabeled

Output Results

- Calculate the accuracy of the learner on the entire dataset
- Deriving the performance metrics

Fig. Flow Chart

- ModAL is a Python library that provides a framework for implementing pool-based active learning. One of the most common pool-based active learning strategies is uncertainty sampling.
- A pool-based active learning strategy, which involves iteratively selecting the most informative data points from an unlabeled pool and adding them to a labeled set, the pool-based active learning strategies tend to be more computationally efficient, as they only require a subset of the unlabeled data to be evaluated at each iteration
- the most informative samples are typically those with the highest uncertainty or the greatest potential to improve the classifier's performance added to the training set. One way to measure uncertainty is by using entropy, which is the measure of the amount of uncertainty in a probability distribution
- The labeled set is used to train a machine learning model, and the model's performance is evaluated on the remaining unlabeled data with 10 iterations




```
Iteration 1: Accuracy = 0.978
Iteration 2: Accuracy = 0.978
Iteration 3: Accuracy = 0.979
Iteration 4: Accuracy = 0.979
Iteration 5: Accuracy = 0.979
Iteration 6: Accuracy = 0.979
Iteration 7: Accuracy = 0.979
Iteration 8: Accuracy = 0.979
Iteration 9: Accuracy = 0.978
Iteration 10: Accuracy = 0.979
```

Fig. Accuracy obtained during the iterations

Results and Discussions

- From the derived performance metrics, it can be inferred that the classification model has performed well with high precision, recall, and F1-score values for most of the classes. The accuracy of the model is also high at 0.99. The Natural class has achieved a perfect score of 1.0 in all performance metrics, indicating that the model is very good at identifying this class.
- The Road_markings class has a slightly lower F1-score at 0.89, which could be due to the complexity of the class.
- Receiver Operating Characteristic (ROC) curve for a multiclass classification problem. The micro-average ROC curve has an area under the curve of 1.0, indicating a perfect classifier.
- The ROC curve for each individual class shows varying levels of classification accuracy, with some classes achieving a perfect score and others having a lower area under the curve. Overall, the classification model shows high accuracy in correctly identifying positive cases (TPR of 1.0) with relatively low false positives (FPR of 0.8 to 1.0) for most classes

```
Test Accuracy: 0.9937749296105115
Confusion Matrix:
[[ 37729    269      0    186     69      0      4     15      1]
 [   200 130796    264     71     53      1      4     72     17]
 [     0    862   4586      0      0      0      0      1      0]
 [   139     80      0 267625    278      3      6     19      7]
 [    61     63      0    194  73437      0      0      0      0]
 [     0      0      0     24      0   6979     37      0      0]
 [    37     13      0     34      0     29   3692      9      0]
 [    17    166      2     58      1      0      2  16183      1]
 [     2     18      0      3      4      0      0      0   1113]]
```

Fig. Confusion matrix

	precision	recall	f1-score	support
Unclassified 0	0.99	0.99	0.99	38273
Ground 1	0.99	0.99	0.99	131478
Road_markings 2	0.95	0.84	0.89	5449
Natural 3	1.00	1.00	1.00	268157
Building 4	0.99	1.00	1.00	73755
Utility_line 5	1.00	0.99	0.99	7040
Pole 6	0.99	0.97	0.98	3814
Car 7	0.99	0.98	0.99	16430
Fence 8	0.98	0.98	0.98	1140
accuracy			0.99	545536
macro avg	0.99	0.97	0.98	545536
weighted avg	0.99	0.99	0.99	545536

Fig. Performance metrics of classified data

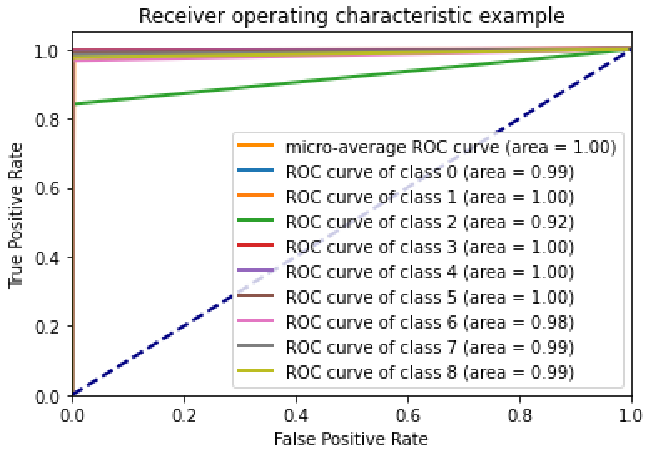


Fig. ROC curve

Output: Classified LiDAR point cloud

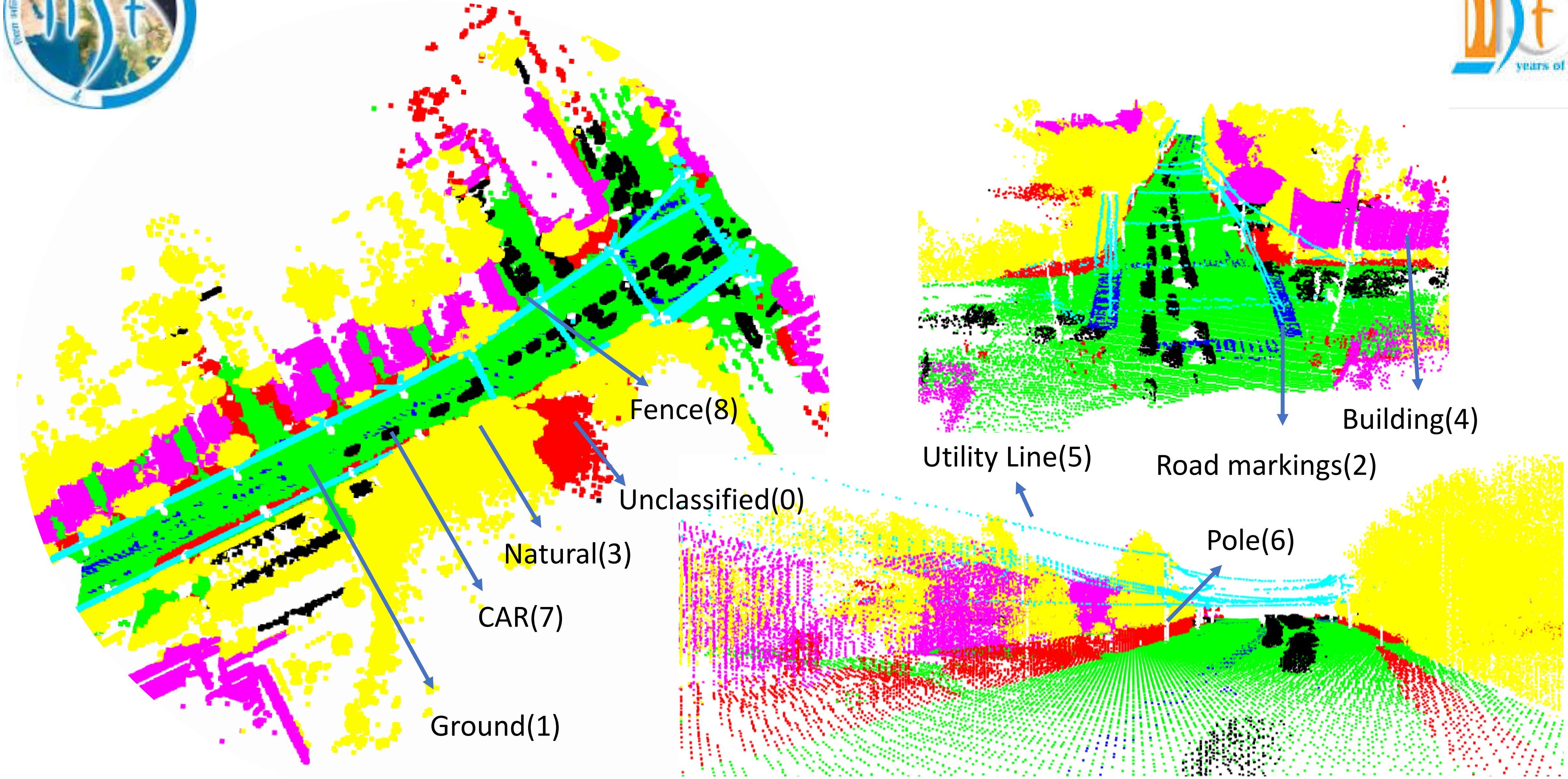


Fig. classified out puts

References

- [1] Tan, W., Qin, N., Ma, L., Li, Y., Du, J., Cai, G., Yang, K., Li, J. (2020). Toronto-3D: A Large-Scale Mobile LiDAR Dataset for Semantic Segmentation of Urban Roadways. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops
- [2]Toronto-3D belongs to Mobile Sensing and Geodata Science Lab, University of Waterloo. Toronto-3D is distributed under the CC BY-NC 4.0 License
- [3] https://github.com/WeikaiTan/Toronto-3D#classes
- [4] https://github.com/modAL-python/modAL
- [5] https://modal-python.readthedocs.io/en/latest/content/examples/pool-based_sampling.html
- [7] Sheykhmousa, M. (2020). Support Vector Machine Versus Random Forest for Remote Sensing Image Classification: A Meta-Analysis and Systematic Review.


Questions?
Suggestions!
THANK
YOU