LiDAR Point Cloud Classification by 3D Sparse CNN for large-scale Mobile Laser Scanning

Nan Li*[1], Harald Teufelsbauer[2], Florian Pöppl[2], Andreas Ullrich[2]

[1]RIEGL Research & Defense GmbH, 3580 Horn, Austria – nli@riegl.com
[2]RIEGL Laser Measurement Systems GmbH, 3580 Horn, Austria - hteufelsbauer@riegl.com; fpoeppl@riegl.com; aullrich@riegl.com;

KEYWORDS: MLS, Sparse Convolutional Neural Network, Point Cloud Classification, Data Augmentation

Abstract

Semantic classification is a fundamental step in Mobile Laser Scanning (MLS) point clouds processing, and remains a non-trivial task. In this work, we propose a classification framework based on a 3D Sparse Convolutional Neural Network (SparseCNN) for efficient processing of large-scale MLS data. A coarse-to-fine two-stage pipeline is introduced, where an essential model performs a classification for the entire scene, followed by a refinement stage for detailed ground-surface classes. To enhance robustness under diverse acquisition conditions, both point-wise and scene-wise data augmentation strategies are employed during the training, including rotation, jittering, density perturbation, noise injection, and patch swapping. To account for environmental and sensor variations, wavelength-specific models are trained for both urban and highway scenes. Experimental results on urban and highway datasets demonstrate strong performance, achieving over 90% accuracy for major classes, while ablation studies show that radiometric features are critical for distinguishing material dependent classes, such as traffic signs, and that the proposed augmentation strategies improve performance for challenging object categories, such as pedestrian, which is dynamic and structurally ambiguous.

1. Introduction

Mobile Laser Scanning (MLS) systems are powerful tools for acquiring high-resolution 3D point clouds of urban and roadway environments, offering rich spatial and radiometric information for applications such as urban mapping and infrastructure monitoring. Downstream tasks often require not only geometric information about the environment, but also semantic information, such as classification of points into specific categories. However, automatic and reliable classification of such point clouds remains a challenging task due to the massive data volume, varying point densities and scene complexity of MLS point clouds.

Recent developments in deep learning have provided an end-to-end classification solution for 3D point clouds, replacing traditional pipelines which require extracting handcrafted features with models that directly predict semantic labels from raw point clouds. The early pioneer work, such as PointNet (Qi et al., 2017) and PointNet++ (Qi et al., 2017) employs Multi-layer Perceptron (MLPs) to learn point-wise features directly from unordered points. Given the widespread application of Convolutional Neural Network (CNN) in image-based tasks, several approaches have adopted the CNN to point cloud, often by transforming irregular 3D points to regular representation, such as voxel grids. However, dense 3D convolution on 3D voxel grids is computationally expensive. To address this issue, Sparse Convolutional Neural Networks (SparseCNNs) (Graham and Van der Maaten, 2017) were proposed, where convolution is applied only to non-empty voxels and the sparsity is preserved to the next layer throughout the network to enable efficiently processing on large scale point cloud. Related frameworks such as Spconv (Yan et al., 2018) and the Minkowski Engine (Choy et al., 2019) follow similar principles with different implementations of sparse convolution on GPU. To avoid the intermediate representation of voxels, KPConv (Thomas et al., 2019) performs convolution directly on the points by using spatial distributed kernel points. Inspired by the success of Transformers in natural language processing (NLP), transformer-based architectures have been introduced to point cloud classification. Point Transformer (PTv1) (Zhao et al., 2020) employs vector self-attention to capture local geometric relationships. Its subsequent extension, PTv2 (Wu et al., 2022) optimizes the attention mechanism and neighbourhood aggregation to improve accuracy and efficiency. PTv3 (Wu et al., 2024) further improves scalability by introducing a serialized neighbourhood representation.

In this work, we employ a 3D Sparse Convolutional Neural Network (SparseCNN) as the architecture for large-scale MLS point clouds classification. Compared to point-based and transformer-based methods, SparseCNN provides a favorable balance between computational efficiency and scalability. To enhance model robustness under varying scanning conditions, we incorporate several augmentation strategies, including geometric transformations, density perturbations, noise injection and patch swap between scenes. Radiometric and echo attributes are used in addition to geometric attributes to increase classification performance.

To balance the efficiency and accuracy for large-scale MLS datasets, we propose a coarse-to-fine classification pipeline: an *essential model*, operating on a coarser spatial

resolution, first classifies the entire scene into broad categories. It is followed by a ground refinement model, which focuses on detailed ground-surface classes that need finer spatial resolution, such as road marking and curbstone.

2. Method

Our network is based on a SparseCNN backbone that operates on voxelized 3D space. Convolution operations are applied only to non-empty voxels, maintaining the sparse layout throughout all layers to preserve efficiency in deeper stages. Each non-empty voxel is represented as a position hash table and a feature matrix, and the input-output voxel index pairs at each kernel position are precomputed and stored. Thus, the SparseCNN converts the intensive 3D convolution into an efficient matrix multiplication.

In addition to the geometric features (XYZ coordinates), the input features include radiometric and echo attributes such as *reflectance*, *return number* and *number of returns*. These attributes provide valuable information to distinguish objects with similar geometry shapes but different material properties, for example, traffic signs in the essential classification and road markings in the ground refinement step.

To improve performance specifically for MLS datasets and enhance the model’s generalization ability across diverse scanning environments, we employ both point-wise and scene-wise augmentations during the training of the *essential model.*

Point-wise augmentations operate at the individual point level and simulate the varying local geometry, point density and noise level. The applied augmentations are listed as following:

- **Random Rotation** around the vertical axis to simulate different vehicle trajectories, for example, driving on sloped roads.
- **Jittering** of point coordinates to add positional noise and avoid overfitting the grid boundary.
- **Random Dropout** include the dropout from the whole scene to handle the varying point density and the dropout from a selected bounding box to simulate the occlusion.
- **Artificial noise** beneath the ground surface to enhance the robustness against false returns or “ghosting” artifacts caused by multi-bounce returns on highly reflective surface like windows and mirrors.

Scene-wise augmentation applies a spatial context modification at a larger sale. We employed a patch swap strategy, which was inspired by the idea proposed in (Xiao et al., 2022) where randomly selected local patches are exchanged between training tiles to increase the local context diversity and make the network more tolerant to structural and radiometric variations across different environments.

To further improve classification quality for ground features which require fine spatial resolution, we implemented a two-step pipeline:

1. Essential classification: The point cloud of the entire scene is processed with a voxel size of 8 cm, and is classified into broad scene-specific semantic categories like ground, building, vegetation, etc.

2. Ground refinement: Points identified as ground are subsequently passed into a ground refinement network with a finer resolution, voxel size of 2 cm. This classifies them into: drive way, side walk, curbstone, road marking and manhole covers.

This hierarchical approach ensures both efficiency and accuracy, as the high-resolution refinement is applied only to relevant subset of points.

3. Setup and Datasets

In real-world applications, the different environments often require scene-specific semantic class definitions. For example, highway structures such as noise barriers and gantries may be categorised as buildings in an urban scene application but are expected to be treated as distinct classes in highway scenarios. Thus, we trained two independent networks for two different scene types, where each model outputs a different set of semantic classes suited to the environment.

- Urban scene model: This model is trained on street-view MLS data containing common objects such as buildings, trees, cars and pedestrians. The Essential classification output classes include ground, low/high vegetation, parking/moving car, pedestrian & bike, fence, wire, pole, traffic sign, traffic light and others. In the Ground refinement model, ground points are classified into the 5 abovementioned classes: drive way, side walk, curbstone, road marking and manhole covers.
- Highway scene model: This model is trained on high-speed corridor MLS data containing features like guardrails, gantries and noise barriers. Output classes include ground, low/high vegetation, car, fence, wire, pole, traffic sign, traffic light, highway wall, highway camera, highway tunnel, highway gantry and others. For this scenario, the Ground refinement model only outputs 3 classes, ground, road marking and manhole, because sidewalks and curbstones are often absent in a highway scene.

Because reflectance responses vary with laser wavelength, we further trained two wavelength-specific model for each scene type. One model is trained on data from a 1550 nm system (e.g., VUX-1HA scanner), and the other on data from a 905 nm system (e.g., miniVUX-HA scanner). This separation is essential in allowing the network to adapt to wavelength-dependent radiometric behavior and maintain reliable classification accuracy across scanners.

The training data used in this study were annotated in-house. For the urban scene model, two datasets corresponding to different laser wavelengths were

prepared. For wavelength 1550 nm, approximately 1.7 km of MLS point cloud data was manually labeled, primarily collected in the city of Vienna. For wavelength 905 nm, a similar amount of data (~1.7 km) scanned on the streets of Hong Kong was labeled. The datasets were deliberately selected from different regions to ensure that the training samples covers a wide range of spatial contexts and object types. These include busy intersections with dense traffic signage, city center areas with tram infrastructure, and residential neighborhoods, thereby ensuring that the model is exposed to various object types and spatial configurations.

For the highway scene model, a larger amount of training data was required due to the sparser spatial distribution of objects compared to urban environments. To maintain sufficient diversity, we annotated approximately 9.7 km of highway point clouds for wavelength 1550 nm, collected from highways in the vicinity of Vienna. For wavelength 905 nm, approximately 5.7 km of highway data was labelled, covering highway sections between Horn and Vienna. Similarly, the samples are selected at different locations to include diverse highway infrastructure elements such as gantries, guardrails, and noise barriers.

4. Results

We conduct a quantitative evaluation of both essential classification and ground refinement for urban and highway scenes acquired using wavelength 1550 nm laser.

For the urban scene, evaluation was performed on an approximately 825 m MLS scan located in the central area of Vienna, which contains representative urban features such as tram stations and dense traffic infrastructures. A full classification result for a typical region is visualized in Figure 1, and the normalized confusion matrix for the essential classification and ground refinement are shown in Figure 2 (a) and (b), respectively.

Most urban object classes achieve an accuracy above 90%, including the dominant objects such as *ground*, *high vegetation* and *building*. Several misclassifications occur between *low* and *high vegetation*, as well as between *parking* and *moving cars*, due to their similar geometric and radiometric characteristics. In some cases, even human experts may find it difficult to distinguish between these categories. In addition, 15.1% of *traffic sign* points are misclassified as *poles*, which can be attributed to the difficulty in distinguishing between traffic signs and similarly shaped commercial boards based solely on point cloud geometric and radiometric features. *Traffic light* is another challenging class in complex urban environments: about 5.0% and 5.5% of *traffic light* points are misclassified as *wire* and *pole*, respectively. This is likely because traffic lights are either suspended from wires or mounted on poles, leading to confusion with surrounding structures or decorative elements.

For the ground refinement, 15% of *side walk* points are misclassified as *drive way*, as the boundary between these two categories is ambiguous in some case. Compared to *road marking* and *manhole*, *curbstone* achieves a higher accuracy, as they exhibit more distinctive geometric features. In contrast, the classification of *road marking* and *manhole* relies heavily on reflectance, which is influenced by environmental conditions such as weather and surface material. In particular, the reflectance values of manholes are only slightly different from the surrounding ground, and these values can vary significantly across different scans. This leads to relatively lower detection accuracy for this class.

For the highway scene model, Figure 3 presents a successful identification of gantries, guardrails (labelled as fence) and moving cars elongated in the point cloud due to the motion during scanning. Figure 4 (a) and (b) provides the confusion matrix of the essential and ground refinement classification, respectively.

As no manhole objects present in the evaluation dataset, only *ground* and *road marking* classes are considered in the evaluation. In addition, the number of *others* and *wire* points is very limited (28768 and 1260 points, accounting for approximately 0.01% and 0.001% of the total test data, respectively), making the corresponding accuracy highly sensitive and less reliable. This is reflected in the relatively low classification accuracy for the *others* and *wire* class.

Based on the confusion matrix, several confusions between certain classes can be observed. Misclassification between *low vegetation* and *ground*, which is possibly caused by that vegetation that is very close to the ground surface and the voxel resolution may limit the ability to accurately distinguish low vegetation from ground. 6.8% of *building* points are misclassified as *highway tunnel*, particularly in the case where wide bridge structures spans across the road and forms a covered region beneath it. Some *pole* points are misclassified as *building* and *gantry* as their vertical geometry is similar to the pillar-like support components of these structures. Despite these challenges, the model achieves high accuracy for several major classes, including *cars*, *highway walls*, *highway gantries*, and *traffic signs*. Additionally, the model demonstrates a strong performance on the classification of *ground* and *road marking*.

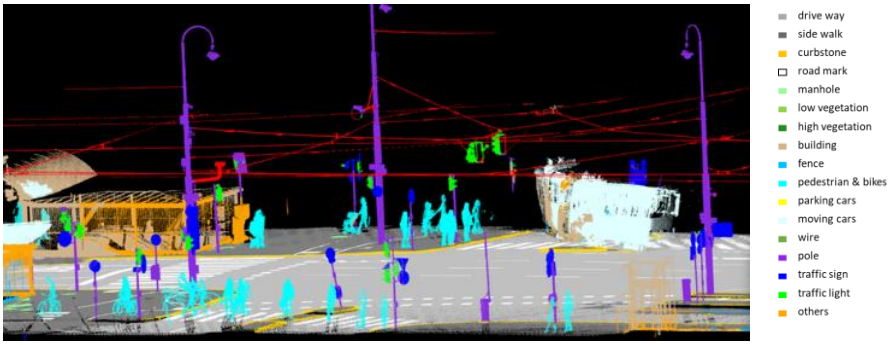
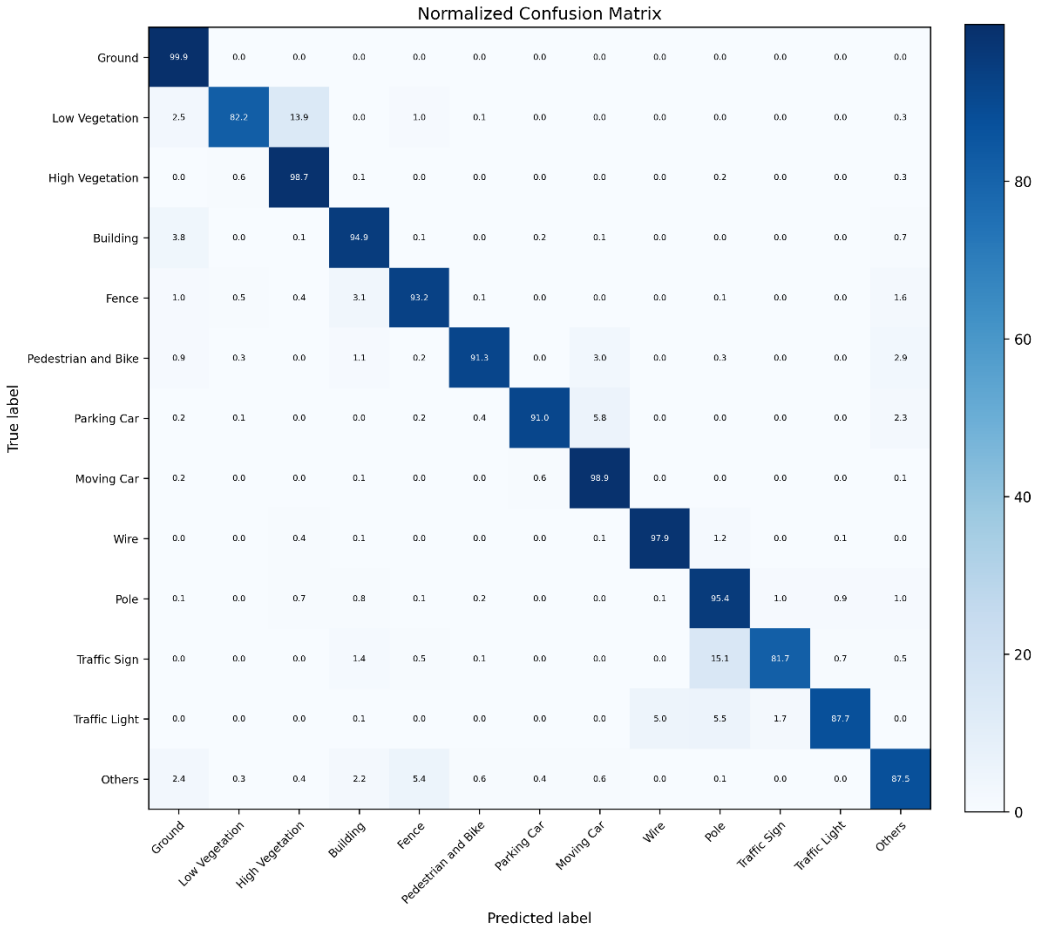
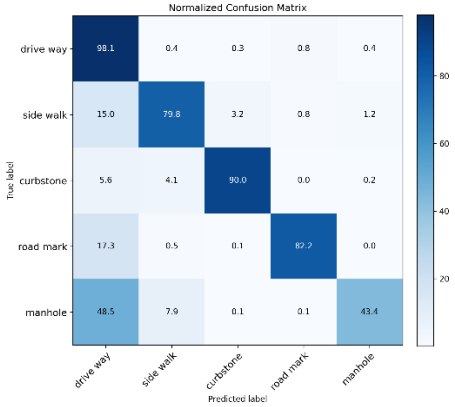


Figure 1. Classification in urban scene



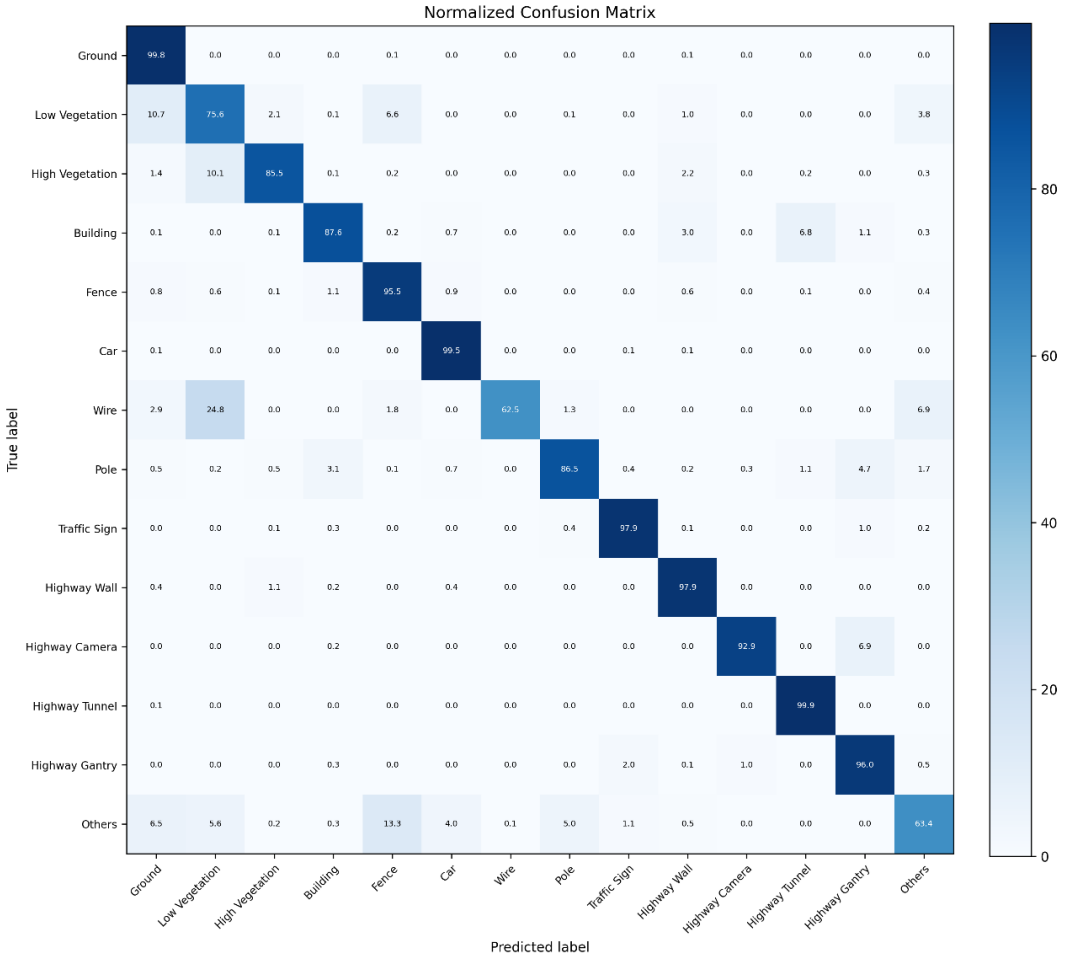
(a) Essential classification



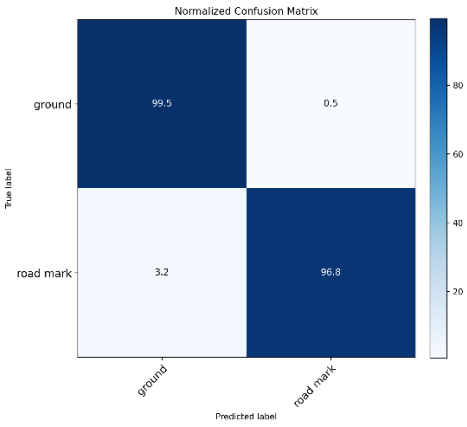
(b) Ground refinement

Figure 2. Normalized confusion matrix of the essential classification and ground refinement

Figure 3. Classification in highway scene



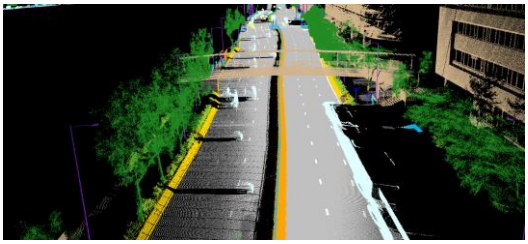
(a) Essential classification



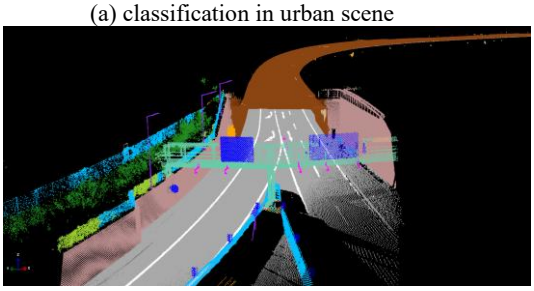
(b) Ground refinement

Figure 4. Normalized confusion matrix of the essential classification and ground refinement

For wavelength 905 nm datasets, both urban and highway scenes classification are visually inspected in Figure 5. Figure 5 (a) shows an urban scene collected from Hongkong with a bridge constructed across the street. (b) captures a tunnel and the adjacent highway walls. The quantitative evaluation has not yet been concluded and will be addressed in future work.



(a) classification in urban scene



(b) classification in highway scene

Figure 5. Classification for dataset collected at wavelength 905 nm laser

5. Discussion

To evaluate the impact of the proposed data augmentation strategies and the contribution of radiometric features, we conducted a series of experiments on the essential classification in the urban scene application. We define four training configurations as the following and Table 1 summarizes the per-class accuracy under each configuration.

- **Baseline:** Training without any augmentation.
- **PointAug:** Training with point-wise augmentations, including rotation, jittering, dropout, and artificial noise.
- **FullAug:** Training with both point-wise augmentations and scene-level augmentation (patch swap).
- **GeomOnly:** Augmentation configuration is the same as **FullAug**, but using only geometric features (XYZ), excluding all radiometric and echo attributes (*reflectance*, *return number* and *number of returns*).

Table 1. Per-class accuracy under different configurations

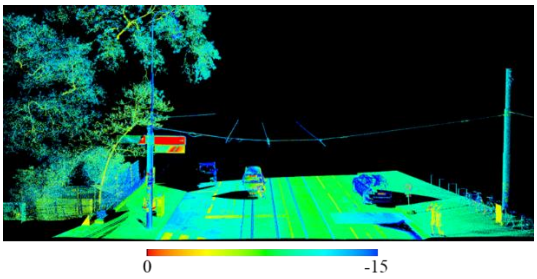
Class accuracy	Baseline	Point Aug	Full Aug	Geom Only
Ground	99.9%	99.8%	99.9%	99.8%
Low vegetation	81.7%	86.1%	82.2%	81.1%
High vegetation	98.1%	98.6%	98.7%	98.4%
Building	95.1%	98.8%	98.4%	92.4%
Fence	90.2%	91.9%	93.2%	91.7%
Pedestrian & Bike	87.2%	92.0%	91.3%	79.6%
Parking car	90.0%	89.2%	91.0%	93.9%
Moving car	99.4%	99.6%	98.9%	98.1%
Wire	96.4%	96.7%	97.9%	98.2%
pole	94.5%	95.0%	95.4%	94.8%
Traffic sign	67.1%	69.2%	81.7%	41.8%
Traffic light	87.0%	86.8%	87.7%	81.5%
Others	86.6%	85.4%	87.5%	89.5%

The radiometric features, particularly *reflectance*, play an important role in distinguishing objects represented more by material properties rather than by geometric structures. For instance, the identification of *traffic sign*, where the accuracy drops significantly from 81.7% by **FullAug** to 41.8% by **GeomOnly** when radiometric features are removed. A similar trend is observed for the *Pedestrian & Bike* class, where the performance decreases notably without radiometric features. In particular, 9.6% of these points are misclassified as *moving car* using the **GeomOnly** configuration. Both categories belong to the dynamic objects and often have an irregular or elongated geometric shapes in point clouds. In such cases, geometric features alone are insufficient, and reflectance becomes essential to distinguish between materials to improve the accuracy, for example, mental surfaces of cars or human clothing or skin.

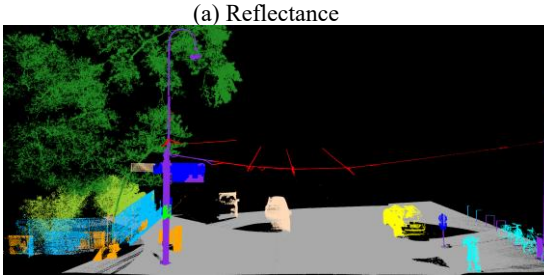
Regarding the effect of augmentations, the introduction of point-wise augmentations (**PointAug**) improves accuracy of several classes. Notably, the accuracy of *Pedestrian & Bike* increases from 87.2% to 92.0%, indicating that these augmentations help the network better handle dynamic objects, which often exhibit noisy and distorted geometries due to motion during scanning. The addition of patch swapping (**FullAug**) further enhances performance for specific classes, particularly *traffic signs*, where accuracy improves significantly from 69.2% to 81.7%. Misclassification of traffic signs often occurs due to inconsistent reflectance patterns across different scenes and sensors. By exchanging local regions between training tiles, patch swapping increases contextual diversity and improves robustness to such variations. An example is illustrated in Figure 6. In this case, as shown in Figure 6 (a) three traffic signs mounted on the same pole exhibit

different reflectance values. One sign shows strong reflectance contrast, while the other two have reflectance levels similar to surrounding objects such as the ground. Without scene-level augmentation, a large area of one sign (on the left-hand side of the pole) is misclassified as *building*. This is largely improved by apply the scene augmentation.

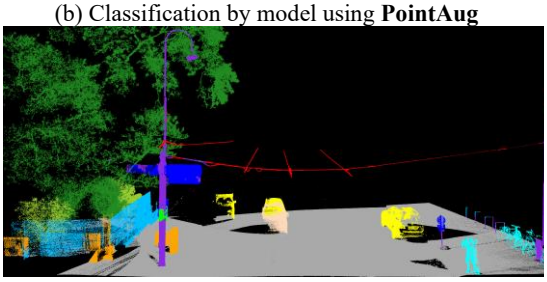
However, we also notice that **FullAug** tends to overly rely on the local spatial context. For example, parts of a bridge surface are occasionally misclassified as ground, suggesting that excessive contextual mixing can negatively affect class discrimination in certain cases, especially for large size objects.



(a) Reflectance



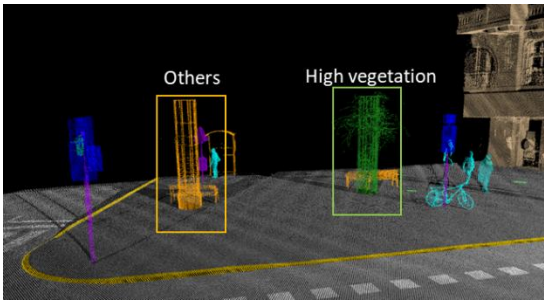
(b) Classification by model using **PointAug**



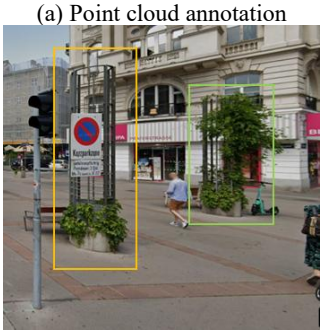
(b) Classification by model using **FullAug**

Figure 6. Reflectance and classification by different augmentation configurations

In addition, part of the observed errors can be attributed to ambiguities in annotation and class definitions. For example, distinctions between categories such as low versus high vegetation, or side walk versus drive way, are not always clearly defined. In some cases, it’s difficult to assign an object to a single category as illustrated in Figure 7, the metal street furniture on the left-hand side is labeled as *others*, whereas a similar object on the right one was labelled as *high vegetation* as it is covered by climbing vegetation. These ambiguities introduce inherent uncertainty in both training and evaluation.



(a) Point cloud annotation



(b) Street image from google map

Figure 7. Comparison between point cloud annotation and reference image

6. Conclusion

We present a robust MLS classification framework based on 3D SparseCNN with well-designed data augmentations and input features. The proposed two-stage pipeline provides an efficient solution for large-scale MLS point clouds classification, enabling a coarse semantic labelling of the entire scene and a refinement of ground-surface features.

The essential classification achieves strong performance in both urban and highway environments, with accuracy exceeding 90% for major object classes. In the ground refinement stage, road markings achieve an accuracy of approximately 80%, while manhole detection remains challenging, primarily due to weak and variable reflectance characteristics in point cloud data. Experiments show that radiometric features significantly improve the accuracy of material-dependent classes, such as traffic lights. And the proposed augmentations further enhance the robustness in the classification of objects with high intra-class variability or ambiguous geometric characteristics.

In future work, we plan to integrate image-based information to further improve classification performance, particularly for fine-grained ground elements such as distinguishing between driveways and sidewalks, as well as enhancing the detection of subtle features like manholes.

Reference

Choy, C., Gwak, J., Savarese, S., 2019: 4D spatio-temporal convolutional networks: Minkowski convolutional neural networks. *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*.

Graham, B., Van der Maaten, L., 2017: Submanifold sparse convolutional networks. *arXiv preprint*,

arXiv:1706.01307. Available at: https://arxiv.org/abs/1706.01307

Qi, C.R., Su, H., Mo, K., Guibas, L.J., 2017: PointNet: Deep learning on point sets for 3D classification and segmentation. *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*.

Qi, C.R., Yi, L., Su, H., Guibas, L.J., 2017: PointNet++: Deep hierarchical feature learning on point sets in a metric space. *Adv. Neural Inf. Process. Syst. (NeurIPS)*, 30.

Thomas, H., Qi, C.R., Deschaud, J.E., Marcotegui, B., Goulette, F., Guibas, L.J., 2019: KPConv: Flexible and deformable convolution for point clouds. *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*.

Wu, X., Lao, Y., Jiang, L., Liu, X., Zhao, H., 2022: Point Transformer V2: Grouped vector attention and partition-based pooling. *Adv. Neural Inf. Process. Syst. (NeurIPS)*, 35, 33330–33342.

Wu, X., Lao, Y., Jiang, L., Liu, X., Zhao, H., 2024: Point Transformer V3: Simpler, faster, stronger. *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*.

Xiao, A., Huang, J., Guan, D., Cui, K., Lu, S., Shao, L., 2022: PolarMix: A general data augmentation technique for LiDAR point clouds. *arXiv preprint*, arXiv:2208.00223. Available at: https://arxiv.org/abs/2208.00223.

Yan, Y., Mao, Y., Li, B., 2018: SECOND: Sparsely embedded convolutional detection. *Sensors*, 18(10), 3337. doi.org/10.3390/s18103337.

Zhao, H., Jiang, L., Fu, C.W., Jia, J., 2021: Point Transformer. *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*.