EXPLORING GROUND SEGMENTATION FROM LIDAR SCANNING-DERIVED IMAGES USING CONVOLUTIONAL NEURAL NETWORKS

M. L. R. Lagahit [1, 3] *, Z. Li [2, 3], K. Sakaguchi [2, 3], M. Matsuoka [1, 3]

[1] Department of Architecture and Building Engineering, Tokyo Institute of Technology, Tokyo, Japan
[2] Department of Electrical and Electronic Engineering, Tokyo Institute of Technology, Tokyo, Japan
[3] Tokyo Tech Academy for Super Smart Society, Tokyo Institute of Technology, Tokyo, Japan
* lagahit.m.aa@m.titech.ac.jp

KEY WORDS: Low-Cost LiDAR, Mobile Mapping, Ground Segmentation, Point Cloud-Derived Images, Deep Learning

ABSTRACT:

Recent works have attempted to extract features such as road markings from sparse mobile LiDAR scanning point cloud-derived images via convolutional neural networks (CNN). In this paper, the use of such methods for ground segmentation was explored. To begin, point clouds from each channel will be projected onto the y-z plane to generate the images that will be used for training and testing the CNN model. Then, for the main workflow, the following steps were performed for each channel: (1) point cloud-to-image conversion; (2) CNN classification; and (3) image-to-point cloud projection. Then utilizing multi-threading, each channel is processed in parallel to generate our ground-segmented sparse point cloud. Our findings have shown successful ground segmentation, achieving an f1-score of 98.9%. However, it performed 27.81% slower as compared to RANSAC. Overall, this initial investigation has demonstrated that ground segmentation from sparse point cloud-derived imagery is possible, and with further improvements to the CNN model, to make it faster, it has good potential to act as an alternative to conventional point cloud processing.

1. INTRODUCTION

1.1 Background

Recent research has attempted to extract features from sparse point cloud-derived images using convolutional neural networks (CNN) generated by low-cost mobile light detection and ranging (LiDAR) scanning. One example is the extraction of road markings such as lane lines and crossing marks that return relatively high intensity values (Lagahit & Matsuoka, 2023). The successful extraction of features from sparse point clouds enables the usage of low-cost LiDARs which leads to a more practical alternative for mobile mapping tasks, especially for those that monitor and track changes in dynamic environments.

An essential task for mobile mapping is the extraction of the ground surface. This enables the generation of digital terrain models (DTM), a digital elevation model (DEM) that represents the ‘bare earth’ (Guth et al., 2021). It also provides the necessary information for roadway-related tasks such as road surface extraction, object detection, and asset management, to name a few (Ma et al., 2018; Elhashash et al., 2022).

Common methods for ground surface extraction include successions of gridding and thresholding (Wu et al., 2016; Yadav et al., 2018; Lim et al., 2021). All of these, are related to iterations of forming a planar surface and determining which points fit or are close to it. In areas where surface (including vegetations and infrastructure) elevations greatly vary this approach faces difficulty and compensates by minimizing the sampling area to form the reference plane. As such, it is intriguing to see the performance of CNNs, which have been demonstrated to work well in situations where conventional methods struggle (Lagahit & Matsuoka, 2023).

1.2 Objective

This work explores ground segmentation from sparse point cloud-derived cross-sectional imagery using CNNs, conducting parallel classifications on each LiDAR channel. To accomplish this goal the following have been done: (1) The resulting CNN predictions on the images, using a variety of loss functions to improve performance, were evaluated; (2) The running time of the implemented procedure in comparison to conventional methods was examined; and (3) the resulting ground segmented point clouds were investigated.

2. METHODOLOGY

2.1 Dataset Gathering and Preparation

The point cloud scanning was obtained by mounting a Velodyne 16-channel LiDAR in front of a vehicle tilted 45 degrees downward. It was done on the roads of the Ookayama campus, Tokyo Institute of Technology. To reduce unnecessary points such as those of trees, the point cloud was filtered to retain only an area in front. It is then projected to the y-z plane with intensity as pixel values, to produce a cross-sectional or profile-view image of the road with a ground resolution of 4 by 4 centimeters and a size of 512 by 64 pixels, as shown in Figure 2-1.

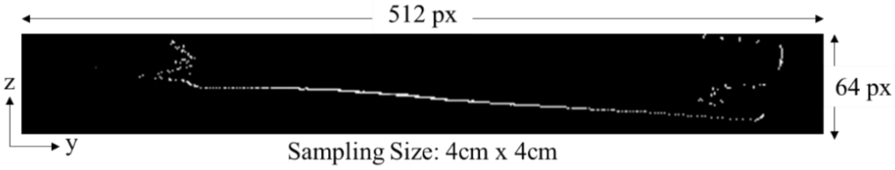


Figure 1-1. A sample generated point cloud-derived image.

https://doi.org/10.5194/isprs-archives-XLVIII-1-W1-2023-221-2023 | © Author(s) 2023. CC BY 4.0 License.

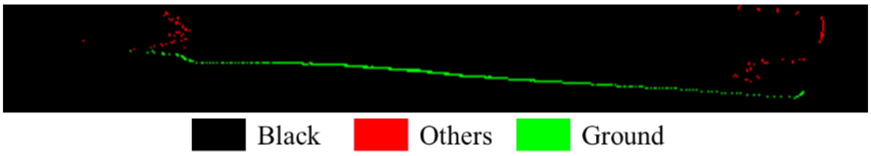


Figure 2-2. A sample labeled point cloud-derived image.

The images are then manually labeled into three classes: 'black', 'others', and 'ground', as can be seen in Figure 2-2. The 'ground' class, the green pixels, represents the target ground points. The 'others' class, the red pixels, represents all other point features. The 'black' class or pixels, represents the regions with no corresponding value in the point cloud.

Table 2-1 shows the distribution of images in the training, validation, and testing datasets, which are composed of pairs of intensity and labeled images as well as the number of class pixels. It can be observed that our target class accounts for roughly only 10% of the datasets while the black class, which has no value, makes up nearly 72%. This severe class imbalance can pose to be a challenge to the performance of the CNN model in obtaining good segmentation results.

Table 2-1. Dataset statistics.

Dataset	Number of Images	Number of Pixels per Class		
		Black	Others	Ground
Training	30,720	71.11%	17.78%	11.11%
Validation	7,680	71.51%	17.89%	10.60%
Testing	4,800	71.24%	17.81%	10.95%

2.2 Model Training

The Fast-SCNN model will be applied in this experiment. This CNN model, which is seen in Figure 2-3, has been developed for real-time segmentation utilizing recent developments, such as pyramid pooling, bottlenecks, and feature fusion (Poudel et al., 2019). It was able to outperform U-Net in terms of prediction speed by up to 15 times (Lagahit & Matsuoka, 2023).

Additionally, to improve performance on a severely class-imbalanced dataset, varying loss functions will also be tested. During model training, in general, the loss function guides the CNN by calculating differences between predictions and masks to adjust the network weights accordingly (Wang et al., 2022). The loss functions that have been implemented, such as the weighted cross-entropy, focal, focal dice, and combo loss functions, were developed to focus on harder class features.

A computer with an 11th Gen Intel i7 processor and 32 GB of memory has been used. A batch size of 16, an Adam optimizer, and a learning rate of 0.0001 were used in model training. Furthermore, a total of 100 epochs were performed for each test, using the model at an epoch with the lowest loss value.

2.3 Segmentation Workflow

The entire segmentation procedure is shown in Figure 2-4. It takes in the sparse point cloud and results in a classified, ground and non-ground sparse point cloud. Firstly, similar to how the dataset for CNN training and testing was prepared, geometric filters are applied to constrict the point cloud and remove noise. Then the point cloud for each LiDAR channel is projected to the y-z plane to generate cross-sectional images of the scanning. Then, multi-threading is employed to enable nearly simultaneous CNN predictions on each of the LiDAR channel-derived images for faster segmentation results. Finally, the classified images will be projected back to 3D space and collated to generate our classified sparse point cloud.

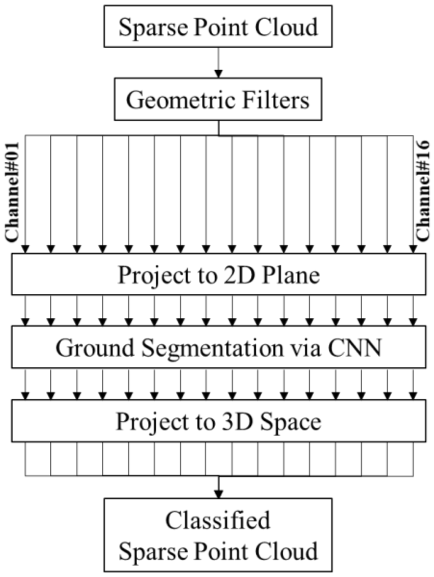


Figure 2-4. General Workflow.

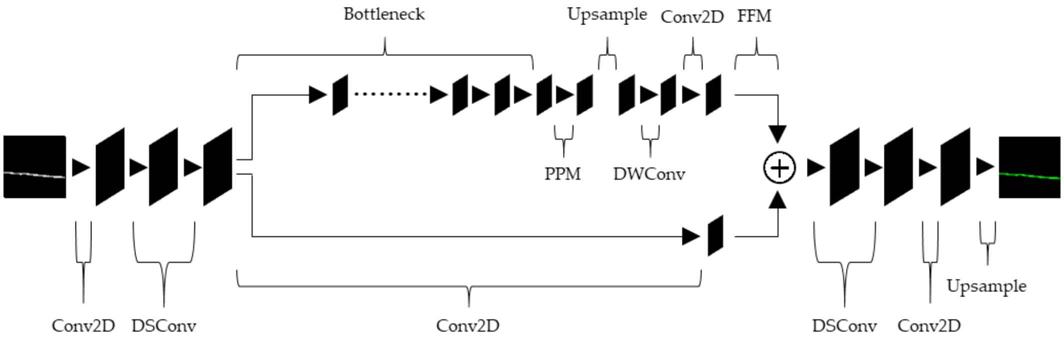


Figure 2-3. Fast-SCNN structure.

2.4 Assessment Criteria

The equations below are common evaluation metrics, derived from the confusion matrix, for assessing the resulting image segmentation. The proportion of actual positive cases that are correctly predicted as positive is known as recall; the proportion of predicted positive cases that are correctly predicted as positive is known as precision, and the harmonic mean between the two is known as the f1-score.

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative}, \qquad (1)$$

$$\text{Precision} = \frac{\text{True Positive}}{\text{True Positive+False Positive}}, \qquad (2)$$

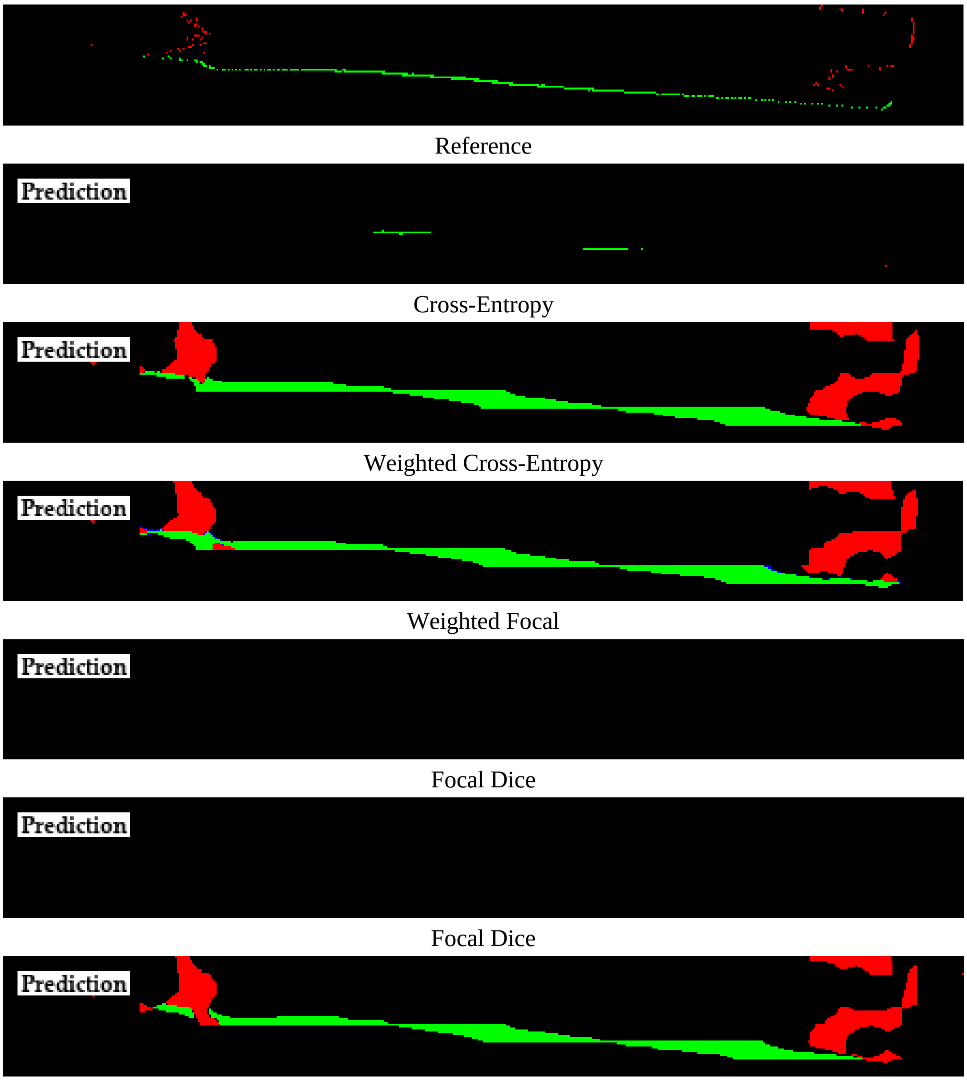
$$F1_{score} = \frac{2 \times Precision \times Recall}{Precision + Recall}, \qquad (3)$$

3. RESULTS AND DISCUSSION

3.1 Classified Images

Figure 3-1 shows the classification results on sparse point cloud-derived images using our proposed method on varying loss functions. It can be seen that models trained with cross-entropy and dice or focal dice (β=1) loss functions greatly struggled to detect features on the point cloud-derived image. Meanwhile, models trained with weighted cross-entropy, weighted focal (γ=2), and combo loss (α=0.75) functions gained prediction results that seemed to overreach and largely misclassify the surrounding pixels. However, the final objective is to project the classifications back into or to a point cloud so misclassifications in the black pixel regions can be omitted since they hold no corresponding point value (Lagahit & Matsuoka, 2023).

After masking out misclassifications in the black class regions, Figure 3-2 shows the resulting classification results on various loss functions. Huge improvements can easily be seen in the ground segmentation brought on by the masking process. In hindsight, it is clear that using weighted cross-entropy or weighted focal (γ=0) and combo loss (α=0.75) loss yielded the best results. Although, we can still spot that the model fails to identify the ground class along the edges of the image.



Reference

Cross-Entropy

Weighted Cross-Entropy

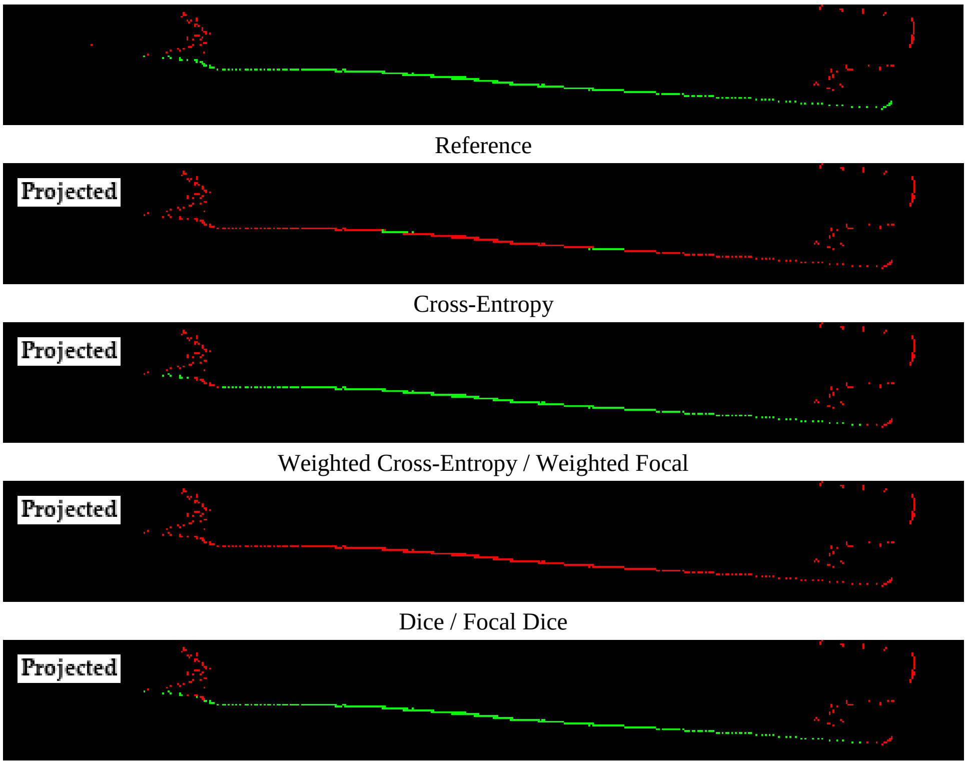
Weighted Focal

Focal Dice

Focal Dice

Combo

Figure 3-1. Sample resulting classified images.



Reference

Cross-Entropy

Weighted Cross-Entropy / Weighted Focal

Dice / Focal Dice

Combo

Figure 3-2. Sample resulting classified images after masking. (Results for some loss functions have been merged since they yielded the same sample results after masking.)

3.2 Ground Segmentation

Taking a look at the numerical evaluations or our resulting classifications as shown in Table 3-2, all our models performed poorly in precision as was reflected by the misclassifications seen in Figure 3-1. The model trained with a weighted focal loss (γ=2) [WF] performed best but it is not far from the results of models trained with a weighted cross-entropy [WCE] and combo (α=0.75) [C] loss, having 0.2 and 0.5 differences in the F1-score, respectively. However, it is important to note that recall from the aforementioned loss functions are high, meaning that a huge number of pixels based on the reference are correctly classified.

Table 3-1. Evaluation results of Fast-SCNN predictions for the target ground class. (%)

Loss Function	Recall	Precision	F1-Score
CE	11.5	57.3	19.1
WCE	**97.8**	18.8	31.5
WF	97.6	**18.9**	**31.7**
D/FD	5.9	0.9	1.5
C	97.7	18.5	31.2

Taking a look at numerical evaluations after the ‘black’ pixel omission shown in Table 3-3, while combo (α=0.75) [C] came close, weighted cross-entropy [WCE] or weighted focal (γ=0) [WF] outperformed all other loss functions in terms of f1-score for the ‘ground’ class. It can also be seen that the addition of weights, derived from class pixel ratios, had a significant impact on cross-entropy [CE] results. Furthermore, it has been observed that dice [D] or focal dice (β=1), which utilized the f1-score metric, performed poorly in comparison to the others even after masking.

Table 3-2. Evaluation results of Fast-SCNN predictions after masking for the target ground class. (%)

Loss Function	Recall	Precision	F1-Score
CE	11.5	**100.0**	20.6
WCE/WF	**97.8**	**100.0**	**98.9**
D/FD	5.9	**100.0**	11.1
C	97.7	**100.0**	98.8

3.3 Processing Speed

Table 3-3 shows the execution time for each component of the workflow presented in Figure 2-4. The pre-processing section contains the initial geometric filtering procedure. The exporting section contains the procedure of merging the classified points and saving them. The classification section contains all the other steps in-between. It can be seen that the classification step takes up the bulk of the processing time.

Table 3-3. Processing speeds of the workflow components**.**

Section	Time Taken (seconds)
Pre-Processing	0.178
Classification	1.037
Exporting	0.003
Total	**1.218**

Table 3-4 shows a comparison to RANSAC, a popular surface fitting method. The input point cloud was also downsampled with a voxel size of 4 cm for a fair comparison. Even after 1000 iterations, our method still falls behind by roughly 0.3 seconds or around 30%. This is largely caused by the slowness of the CNN classification procedure.

Table 3-4. Comparison to RANSAC.

Method	Iterations	Speed (seconds)
Ours	---	1.218
RANSAC (downsampled)	1000	**0.953**

3.4 Classified Sparse Point Cloud

What's more, as shown in Table 3-5, the number of ground points produced by our method is greatly reduced, even less than that of a downsampled point cloud undergoing RANSAC. Depending on how it is perceived, it can become advantageous in terms of file size and disadvantageous in terms of accuracy, considering a mean point distance of 2.3 cm to the reference.

Table 3-5. Generated ground point statistics.

Method	Number of Ground Points	Mean Point Distance
Ours	4,912	2.3 cm
RANSAC (downsampled)	5,550	3.2 cm
Reference	12,660	---

Figure 3-3 shows the resulting classified point cloud using our proposed method. Visually, there is very little difference between the ground segmentation between RANSAC and our method. Although, as was observed in the resulting classified images we can also see some point clouds misclassified as non-ground along the edges.

4. CONCLUSION

Our proposed method of ground segmentation through point cloud-derived cross-sectional images using CNN has proven to be successful in terms of accuracy, achieving an f1-score of 98.9%. However, it still has a long way to go in terms of speed, with a runtime that is 27.81% slower than the conventional RANSAC method. Furthermore, due to image conversion, less than half of the original points have also vanished, which can be concerning in terms of data loss but useful in terms of size reduction. The authors also acknowledge that huge improvements are still needed to be done in each of the workflow's sections, such as the CNN structure, to make the proposed method much faster and thus more practical than the conventional methods. It is quite unfortunate that our findings did not result in significant overall improvements, but this preliminary attempt has demonstrated that ground segmentation from sparse point cloud-derived imagery is possible and could be a potential alternative.

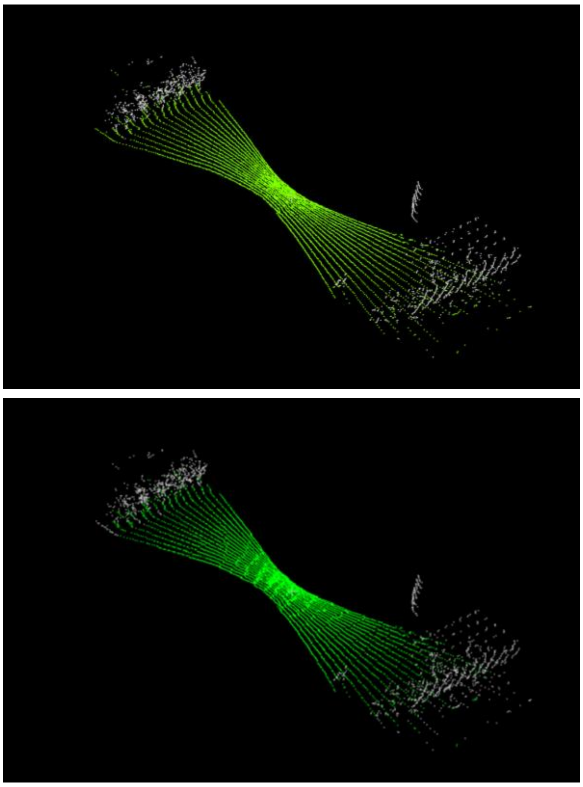


Figure 4-1. Resulting classified sparse point cloud. (Top) RANSAC and (Bottom) Ours.

REFERENCES

Elhashash, M., Albanwan, H., Qin, R. 2022. A Review of Mobile Mapping Systems: From Sensors to Applications. Sensors, 22, 4262.

Guth, P.L., Van Niekerk, A., Grohmann, C.H., Muller, J.-P., Hawker, L., Florinsky, I.V., Gesch, D., Reuter, H.I., Herrera-Cruz, V., Riazanoff, S., López-Vázquez, C., Carabajal, C.C., Albinet, C., Strobl, P. 2021. Digital Elevation Models: Terminology and Definitions. Remote Sensing, 13, 3581.

Lagahit, M.L.R., Matsuoka, M. 2023. Focal Combo Loss for Improved Road Marking Extraction of Sparse Mobile LiDAR Scanning Point-Cloud-Derived Images Using Convolutional Neural Networks. Remote Sensing, 15(3), 597.

Lim, H., Oh, M., Myung, H. 2021. Patchwork: Concentric Zone-Based Region-Wise Ground Segmentation With Ground Likelihood Estimation Using a 3D LiDAR Sensor. IEEE Robotics and Automation Letters, 6, 4, 6458-6465.

Ma, L., Li, Y., Li, J., Wang, C., Wang, R., Chapman, M. 2018. Mobile Laser Scanned Point-Clouds for Road Object Detection and Extraction: A Review. Remote Sensing, 10, 1531.

Poudel, R., Liwicki, S., Cipolla, R. 2019. Fast-SCNN: Fast Semantic Segmentation Network. arXiv.

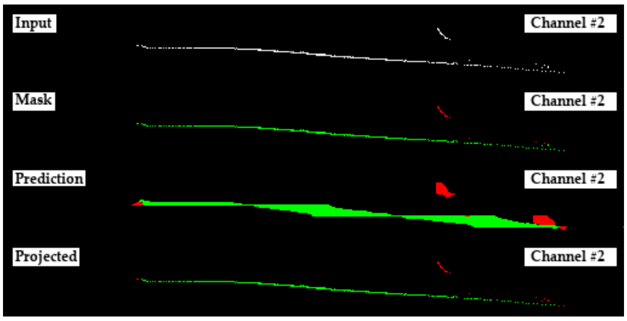
Wang, Q., Ma, Y., Zhao, K., Tian, Y. 2022. A Comprehensive Survey of Loss Functions in Machine Learning. Annals of Data Science, 9, 187-212.

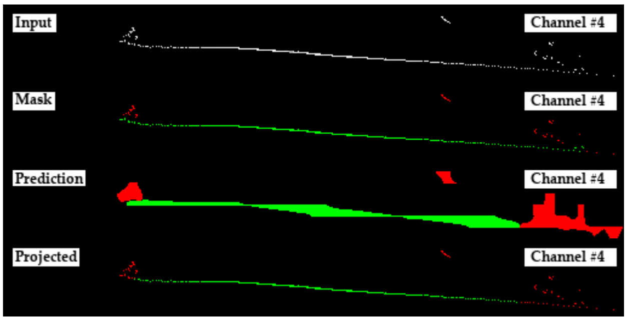
Wu, B., Yu, B., Huang, C., Wu, Q., Wu, J., 2016. Automated extraction of ground surface along urban roads from mobile laser scanning point clouds. Remote Sensing Letters, 7:2, 170-179.

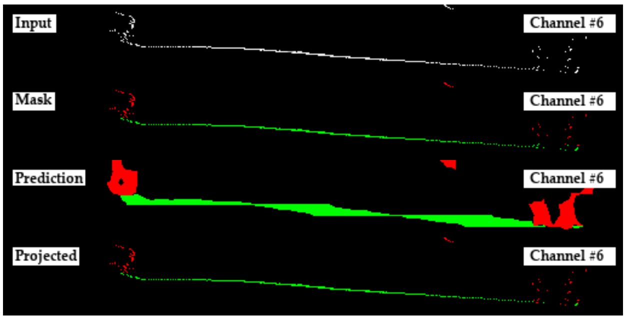
Yadav, M., Singh, A. K., Lohani, B. 2018. Extraction of Ground Surface along Roadway from Mobile LIDAR Data. ISPRS Ann. Photogramm. Remote Sens. Spatial Inf. Sci., IV-5, 103–107.

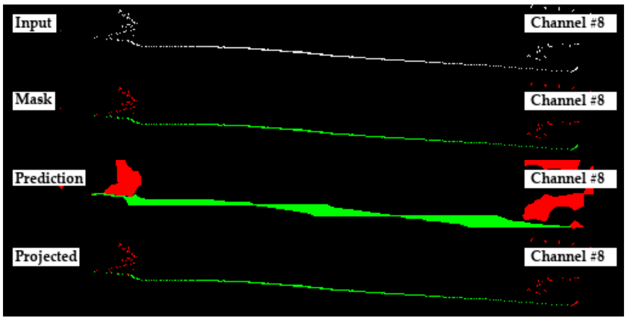
APPENDIX

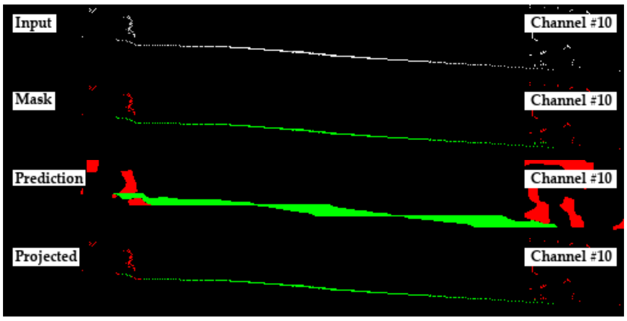
This appendix contains the classification results (raw predictions and projected versions) for every even LiDAR channel, together with their intensity (input) and labeled image (mask) counterparts, in one scanning to better visualize the results of the procedure shown in Figure 2-4. The intensity and labeled images should also provide more visual insights into the datasets used for training, validation, and testing.

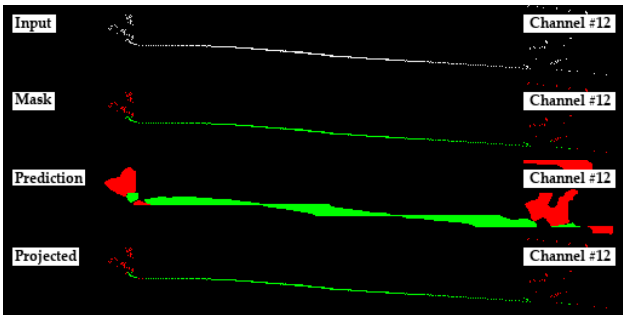

Input
Channel #2
Mask
Channel #2
Prediction
Channel #2
Projected
Channel #2

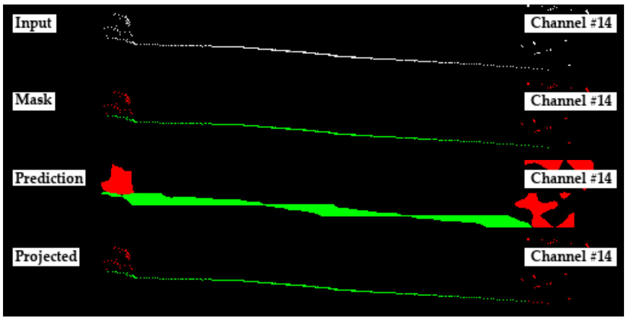

Input
Channel #4
Mask
Channel #4
Prediction
Channel #4
Projected
Channel #4

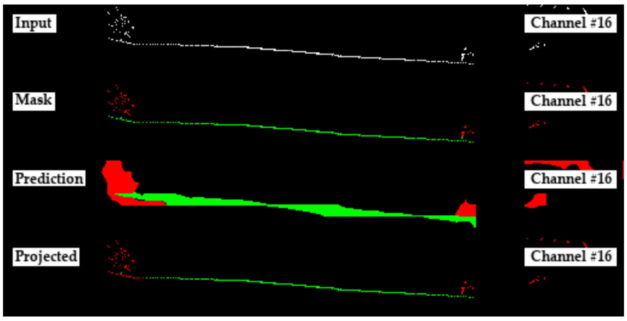

Input
Channel #6
Mask
Channel #6
Prediction
Channel #6
Projected
Channel #6


Input
Channel #8
Mask
Channel #8
Prediction
Channel #8
Projected
Channel #8


Input
Channel #10
Mask
Channel #10
Prediction
Channel #10
Projected
Channel #10


Input
Channel #12
Mask
Channel #12
Prediction
Channel #12
Projected
Channel #12


Input
Channel #14
Mask
Channel #14
Prediction
Channel #14
Projected
Channel #14


Input
Channel #16
Mask
Channel #16
Prediction
Channel #16
Projected
Channel #16