Reichler, Mikael; Taher, Josef; Manninen, Petri; Kaartinen, Harri; Hyyppä, Juha; Kukko, Antero

Semantic segmentation of raw multispectral laser scanning data from urban environments with deep neural networks

Published in:
ISPRS Open Journal of Photogrammetry and Remote Sensing

DOI:
10.1016/j.ophoto.2024.100061

Published: 01/04/2024

Document Version
Publisher's PDF, also known as Version of record

Published under the following license:
CC BY

Please cite the original version:
Reichler, M., Taher, J., Manninen, P., Kaartinen, H., Hyyppä, J., & Kukko, A. (2024). Semantic segmentation of raw multispectral laser scanning data from urban environments with deep neural networks. *ISPRS Open Journal of Photogrammetry and Remote Sensing*, *12*, 1-17. Article 100061. https://doi.org/10.1016/j.ophoto.2024.100061

ISPRS Open Journal of Photogrammetry and Remote Sensing 12 (2024) 100061

Contents lists available at ScienceDirect

ISPRS Open Journal of Photogrammetry and Remote Sensing

journal homepage: www.journals.elsevier.com/isprs-open-journal-of-photogrammetry-and-remote-sensing

Semantic segmentation of raw multispectral laser scanning data from urban environments with deep neural networks

Mikael Reichler [a,b], Josef Taher [a,b,*], Petri Manninen [a,c], Harri Kaartinen [a], Juha Hyyppä [a], Antero Kukko [a,d]

[a] *Finnish Geospatial Research Institute FGI, Vuorimiehentie 5, Espoo, 02150, Finland*
[b] *Aalto University, School of Science, Espoo, 02150, Finland*
[c] *Aalto University, School of Electrical Engineering, Espoo, 02150, Finland*
[d] *Department of Built Environment, Aalto University, School of Engineering, Espoo, 02150, Finland*

ARTICLE INFO

Keywords:
Multispectral point cloud
Mobile laser scanning
Semantic segmentation
Deep learning
Convolutional neural network
Real-time

ABSTRACT

Real-time semantic segmentation of point clouds has increasing importance in applications related to 3D city modelling and mapping, automated inventory of forests, autonomous driving and mobile robotics. Current state-of-the-art point cloud semantic segmentation methods rely heavily on the availability of 3D laser scanning data. This is problematic in regards of low-latency, real-time applications that use data from high-precision mobile laser scanners, as those are typically 2D line scanning devices. In this study, we experiment with real-time semantic segmentation of high-density multispectral point clouds collected from 2D line scanners in urban environments using encoder - decoder convolutional neural network architectures. We introduce a rasterized multi-scan input format that can be constructed exclusively from the raw (non-georeferenced profiles) 2D laser scanner measurement stream without odometry information. In addition, we investigate the impact of multispectral data on the segmentation accuracy. The dataset used for training, validation and testing was collected with multispectral FGI AkhkaR4-DW backpack laser scanning system operating at the wavelengths of 905 nm and 1550 nm, and consists in total of 228 million points (39 583 scans). The data was divided into 13 classes that represent various targets in urban environments. The results show that the increased spatial context of the multi-scan format improves the segmentation performance on the single-wavelength lidar dataset from 45.4 mIoU (a single scan) to 62.1 mIoU (24 consecutive scans). In the multispectral point cloud experiments we achieved a 71 % and 28 % relative increase in the segmentation mIoU (43.5 mIoU) as compared to the purely single-wavelength reference experiments, in which we achieved 25.4 mIoU (905 nm) and 34.1 mIoU (1550 nm). Our findings show that it is possible to semantically segment 2D line scanner data with good results by combining consecutive scans without the need for odometry information. The results also serve as motivation for developing multispectral mobile laser scanning systems that can be used in challenging urban surveys.

1. Introduction

Laser scanning in urban environments has multiple important and wide-ranging application areas, including flood hazard mapping (Costabile et al. (2021)), building 3D models of cities (Kada and McKinley (2009)), autonomous vehicle localization and mapping (Levinson and Thrun (2010)), compliance control of building sites (Bosché (2010)), map updating and change detection (Hyyppa et al. (2009)), planning of road environment illumination control (Maksimainen et al. (2020)) and improving the accuracy of predictive house price models (Helbich et al. (2013)). The aforementioned application areas share a commonality in the sense that they all benefit from the semantic segmentation of the point cloud data before proceeding with the subsequent processing steps, such as automated measurement of object attributes, feature extraction and visualization. In many cases, it is beneficial to provide the end results to the user with minimal delay, or even in real-time. Despite the recent developments in real-time capable point cloud semantic segmentation methods for 3D laser scanner data (Hu et al. (2020); Zhang

* Corresponding author. Department of Remote Sensing and Photogrammetry, Finnish Geospatial Research Institute FGI, The National Land Survey of Finland, Vuorimiehentie 5, FI-02150, Espoo, Finland.
E-mail address: josef.taher@nls.fi (J. Taher).

https://doi.org/10.1016/j.ophoto.2024.100061
Received 4 October 2023; Received in revised form 17 January 2024; Accepted 27 February 2024
Available online 1 March 2024

et al. (2020)), the semantic segmentation methods for high-density point clouds collected with survey-grade 2D laser scanners have not yet been extensively studied in online segmentation tasks.

The raw lidar data collected using 2D laser scanners is typically preprocessed by direct georeferencing into 3D point clouds before it is used in a downstream task. However, there are two problems related to this approach: First, the georeferencing process is carried out in an offline process, which by definition excludes all applications that require real-time capability. Second, the quality of the positioning solution substantially affects the quality of the georeferenced 3D point cloud. Mobile laser scanning systems can rely on direct georeferencing when the GPS visibility is good (Kukko et al. (2012)), but when the positioning satellite signal suffers from multi-path reflections, or is attenuated or obstructed by vegetation or buildings, the degraded positioning solution introduces substantial amounts of systematic error and inaccuracies into the georeferenced 3D point cloud (Kaartinen et al. (2015)).

It is also possible to create 3D point cloud from 2D scans through scan registration, but this approach is highly susceptible to errors when subsequent scans lack common features. In some cases, ambiguities present in the measurement scenario make it impossible to perform the task. This happens, for instance, when the measurement trajectory is perpendicular to the 2D scanning plane and the kinematic state estimate of the measurement system is inaccurate, or not available at all.

To overcome the challenges of data preprocessing via georeferencing or registration, Kaijaluoto et al. (2022) introduced a real-time capable deep learning based semantic segmentation method for raw 2D laser scanner data. In their method, each 360° rotation of the laser scanner mirror (individual scan line) is projected on a plane where the x-axis represents the rotation angle of the laser beam and y-axis represents the echo number (index of the returning laser pulse when multiple targets per each illumination pulse has been detected). The point-wise attributes, such as intensity, reflectance estimate, echo deviation and range, are encoded into the z-dimension using separate channels of the input matrix. Therefore, each 360° mirror rotation, or scan line, can be transformed in a straightforward manner into an input tensor that is then semantically segmented using the proposed LSSegNet fully convolutional encoder - decoder semantic segmentation architecture.

Kaijaluoto et al. (2022) et al. obtained results comparable to RandLA-Net (Hu et al. (2020)) 3D deep learning semantic segmentation network in spite of much simpler data preprocessing steps required by their method. However, they only experimented with single 2D scans in a forest environment and did not take into account multiple consecutive scans that might provide additional context information to the segmentation model. In urban environments the context information might be valuable, because the geometric texture of object categories along the scan line is less pronounced when compared to forest environment. Additionally, they only investigated data originating from a standard single-wavelength laser scanning system operating at the 1550 nm wavelength band and did not consider using multispectral lidar data with the network.

In this paper, we propose two things, the use of multiple consecutive scan lines and the use of multispectral laser scanning data to improve the accuracy of semantic segmentation in an urban environment. We use the raw 2D laser scanner data and compare the accuracy of semantic segmentation between multi-scan and a single-scan input format. Secondly, we compare single-wavelength laser scanner data to multispectral laser scanner data in a classification task. To enable real-time capability, we only experiment with deep learning-based semantic segmentation architectures that do not require georeferencing nor registration of the input data prior to semantic segmentation. Our contributions are as follows.

1. We introduce a projection-based convolutional neural network (CNN) input format for high-density raw 2D laser scanner data. The proposed input format utilizes multiple consecutive scans but does not require a data registration phase prior to inference.
2. We demonstrate that the use of multiple scans significantly improves the accuracy of semantic segmentation.
3. We show how the 2D multispectral laser scanning data can be used to improve the accuracy of semantic segmentation in an urban environment.

2. Related work

2.1. Real-time semantic segmentation of large-scale point clouds

Real-time semantic segmentation of point clouds is required as a preprocessing step in various applications for improving the end-user experience, such as in 3D modelling of urban environments (Babahajiani et al. (2017)) and in creating digital twins of cities (Lehtola et al. (2022)). In some applications the real-time semantic segmentation step is crucial for the correct operation of the whole data processing pipeline, such as in compressing high-definition point cloud maps for autonomous driving (Manninen et al. (2022)) and in improving the performance of the iterative closest point algorithm for lidar odometry (Parkison et al. (2018)). The semantic segmentation of point cloud data using deep learning can be divided roughly into three categories: point-based, voxel-based and projection-based methods. The categories and the methods in each category differ substantially in their capability of being used in real-time semantic segmentation tasks with large-scale point clouds produced by survey-grade 2D laser scanners.

Point-based semantic segmentation methods (Qi et al. (2017); Wang et al. (2019); Engel et al. (2021); Zeid et al. (2023); Hu et al. (2020)) operate directly on the permutation invariant set of 3D points and do not require the spatial hierarchy of the points to be passed on separately. They learn the geometrical relationships between the points and benefit from a wide field-of-view of the input data. Unfortunately, individual 2D lidar scan lines provide only a minimal spatial extent of the data in three-dimensions, unlike individual scans from 3D laser scanners, which might negatively affect the segmentation accuracy. Additionally, many of the methods suffer either from high memory usage with large-scale point clouds (Qi et al. (2017); Wang et al. (2019)) or from computational complexity that is quadratic with respect to the number of points (Zeid et al. (2023); Chen et al. (2023)). This renders the methods impractical for real-time semantic segmentation of mobile laser scanned 2D lidar data with the current hardware.

The concerns with slow inference speed and the potentially high memory usage typically related with point-based methods have been addressed in many voxel-based (Liu et al. (2019); Tang et al. (2020)) and projection-based methods. Projection-based methods are typically the fastest way to semantically segment large-scale point clouds which makes them an attractive option for real-time applications. The classification and semantic segmentation performance of projection-based models has been found to match (Goyal et al. (2021)), and even surpass (Kong et al. (2023)), that of more complex point- or voxel-based models further supporting their favourable position when compared to other methodologies.

A large body of literature has been published around the semantic segmentation of point cloud data at individual scan level, originating from rotating 3D laser scanners, using projection-based methods and various CNN architectures (Li et al. (2020); Zhang et al. (2020)). For example, Xie et al. (2021), experimented with real-time semantic segmentation of 3D laser scanning data by first projecting the points on an image plane via spherical projection and then applying a lightweight CNN model to the range image. They achieved 47.9 % mean Intersection over Union (mIoU) in the SemanticKITTI dataset (Behley et al. (2019); Geiger et al. (2012)). They also demonstrated the real-time capability of their method by performing inference on a FPGA platform. Wen et al. (2022) proposed a projection based method for point cloud semantic segmentation using a combination of data processing stages involving the patching of the input range image into a linear representation using Hilbert space filling curve and segmenting the input representation with an U-net style hybrid CNN-LSTM network architecture. They achieve 56.9 mIoU in the SemanticKITTI dataset Behley et al. (2019); Geiger et al. (2012). Further approaches for projection-based methods that use 3D laser scanner data have been listed in a comprehensive review paper by Li et al. (2020).

The literature concerning semantic segmentation of 2D laser scanner data at the scan line level, prior to registration or georeferencing, is very narrow. Previous works include, for example, deep learning based methods (Kaijaluoto et al. (2022); Balado et al. (2021)) and heuristic based methods (Lehtola et al. (2019)). They all demonstrate the semantic segmentation of raw survey-grade 2D laser scanner data into multiple classes in outdoor environments without registering or georeferencing the data prior to segmentation.

Kaijaluoto et al. (2022) introduced a series of projection-based LSSegNet CNN architectures that operated using only a single scan line and demonstrated the efficacy of the proposed architectures in forest environment by semantically segmenting points into ground, understorey, tree trunk and foliage classes. In their experiments they were able to achieve 80.1 % mIoU while a point-based RandLA-Net model (Hu et al. (2020)) achieved 80.6 % mIoU. This was despite the much more limited context information contained in the individual scan lines used by the LSSegNet model when compared to the RandLA-Net model which used a georeferenced point cloud of the same measurement scene as an input.

A rather minimalistic and innovative approach for the semantic segmentation of 2D laser scanner data at the scan line level was introduced by Balado et al. (2021), who proposed to semantically segment mobile laser scanning data using a Long Short-Term Memory (LSTM) based network architecture and a one-dimensional representation of the input point cloud where only the z-component (point height from ground level) of the coordinate values was used as point attribute for each scan line. They report overall accuracies as high as 97.3 % for three classes and 95.7 % for 11 classes. However, their data processing approach casts a doubt on the generalizability of the results. The dataset has been reported to have an average separation of only 2.7 cm between the scan lines along the driving direction which implies a high spatial correlation between the consecutive scan lines. The split into training and test sets has been performed by randomly sampling individual scan lines from the dataset with a 50 - 50 split, and therefore, the probability of obtaining highly correlating pairs of training and test scan lines is unreasonably high in order for the experimental setup to be considered well-constructed.

Lehtola et al. (2019) introduced a heuristic method, the preregistration classification (PRC), for classifying raw 2D laser scanner points based on the statistical properties of the point neighborhood. The benefit of PRC is that it is suitable for real-time computation on a CPU. However, PRCs disadvantage is that the obtained classification accuracy is sensitive to the methods control parameters, and additionally, the method does not incorporate a way for including radiometric information on the classification process.

2.2. Multispectral laser scanning and semantic segmentation of multispectral point clouds

There is a large volume of research describing the role of spectral information in remote sensing (Shaw and Burke (2003); Schott (2007); Khan et al. (2018)). Much of the literature has been concentrated around applications which utilize passive multi- or hyperspectral camera sensors that are used for retrieving the spectral information of targets from airborne platforms (Honkavaara et al. (2014); Näsi et al. (2018)) or from satellites (Van Mol et al. (2004); Pengra et al. (2007); Pignatti et al. (2013)). However, there is still a relatively small body of literature that is concerned with simultaneous spectral and geometric measurements of targets using multi- or hyperspectral laser scanners in mobile laser scanning setting. Despite this, interest in the field is increasing steadily in many applications (Kaasalainen (2019); Kaasalainen and Malkamäki (2021)) due to the multitude of motivating factors, such as improved estimation capability for various phenological traits (Wallace et al. (2012)) and greater distinguishability between various inorganic materials (Malkamäki et al. (2019)).

According to a 2019 survey (Kaasalainen (2019)) there were only ten research instruments worldwide for acquiring simultaneous three-dimensional point cloud data and spectral measurements of the points. Recently, there has been a surge in the number of companies and institutions publishing intellectual property related to multispectral laser scanners for use in autonomous vehicles, robotics, and surveying (Hall (U.S. Patent US8675181B2, Mar. 2014); Viswanathan and Xue (U. S. Patent US20190212447A1, Dec. 2020); Bozchalooi and Youcef-Toumi (U.S. Patent US10649072B2, May. 2020)). This trend is indicative of a growing interest in commercializing this technology and making the data ubiquitous.

Multi- and hyperspectral point clouds can be collected with devices and methods ranging from supercontinuum based hyperspectral lidars (Hakala et al. (2012); Taher et al. (2022)), and a combination of lidars operating at different wavelengths (Kukko et al. (2020); Vierhub-Lorenz et al. (2022); Hakula et al. (2023)), to hybrid methods combining actively illuminated or passive hyperspectral imaging and standard single-wavelength laser scanning (Suomalainen et al. (2011); Mitschke et al. (2022)).

Wehr et al. (2006) were one of the first to experiment with multispectral laser scanning. In their experiments, they demonstrated that multispectral lidar could be used for moisture detection of building surfaces and for inspection of damaged leaves in plants. In later publications, multi- and hyperspectral lidar data has been used for assessing the structural and physiological characteristics of forests, such as the Normalized Difference Vegetation Index (NDVI) in a simulation environment (Morsdorf et al. (2009)) and using a supercontinuum laser based hyperspectral lidar (Chen et al. (2010); Hakala et al. (2012)). Gong et al. (2015) compared single-wavelength lidar data to four wavelength multispectral lidar data in a classification task containing seven different object categories and were able to improve the classification accuracy by 39.2 percentage points when compared to single-wavelength lidar data. Similarly, Taher et al. (2022) illustrated the value of the spectral information in an autonomous driving related classification task using 30 channel single-photon sensitive hyperspectral lidar.

Airborne laser scanning with three wavelength multispectral lidar system, the Teledyne Optech Titan (van Rees (2015)), has been demonstrated in various applications: Matikainen et al. (2016) proposed to use the data in an automated map updating setup, Ekhtari et al. (2018), Matikainen et al. (2020) and Li et al. (2022) used the data to perform land cover classification, and Enayetullah et al. (2022) experimented with peatland tree type classification using the data. They all reported results that highlighted the importance of the spectral information in addition to the geometry of the objects. Previous studies, however, were conducted offline and did not consider the suitability of the methods for real-time applications.

The deep learning based semantic segmentation of multispectral lidar data collected from vehicle-mounted laser scanning platforms has been demonstrated previously using fully-convolutional encoder - decoder architecture by Taher (2019), using 3D RandLA-Net architecture and semi-supervised learning by Lőrincz et al. (2021), and using voxelized input representation and 3D U-Net type segmentation architecture by Vierhub-Lorenz et al. (2022). Our work differs from the previous work by not requiring registration of the data prior to the semantic segmentation.

3. Materials and methods

In this work, a high-density point cloud dataset was formatted to three different 2D raster formats to investigate the suitability of non-registered raw scan data in increasing the spatial field-of-view, and to examine the relationship of spectral information on the semantic segmentation performance. The performance was explored by comparing the input format proposed by Kaijaluoto et al. (2022) (referred to as the *single-scan* format) to formats containing multiple consecutive scan lines (referred to as *multi-scan* formats). The effect of spectral information was explored by developing a *multispectral* format which combines data from two scanners with different laser wavelengths, and comparing the results to corresponding monochrome results. These three input formats are discussed in more depth in sections 3.3.1-3.3.3.

3.1. Backpack laser scanning system

The point cloud dataset used in this work was collected with the AkhkaR4-DW backpack laser scanning system (Kukko et al. (2020)) shown in Fig. 1, developed by the Finnish Geospatial Research Institute (FGI). The backpack includes two survey-grade lidar instruments, the Riegl VUX-1HA (VUX) and miniVUX-1UAV (miniVUX), as well as a positioning system consisting of a NovAtel Pwrpak7 GNSS receiver with a GNSS-850 antenna and a ISA-100C inertial measurement unit (IMU). The system receives signals from four satellite constellation satellites (GPS, GLONASS, GALILEO and BEIDOU), which are processed offline together with data from the IMU in a tighly coupled manner to produce position and attitude estimates at 200 Hz.

The laser scanners operate at wavelengths of 1550 nm (VUX) and 905 nm (miniVUX), and on-line processing facilitates the discerning of multiple echoes per pulse. In addition to range and reflectance

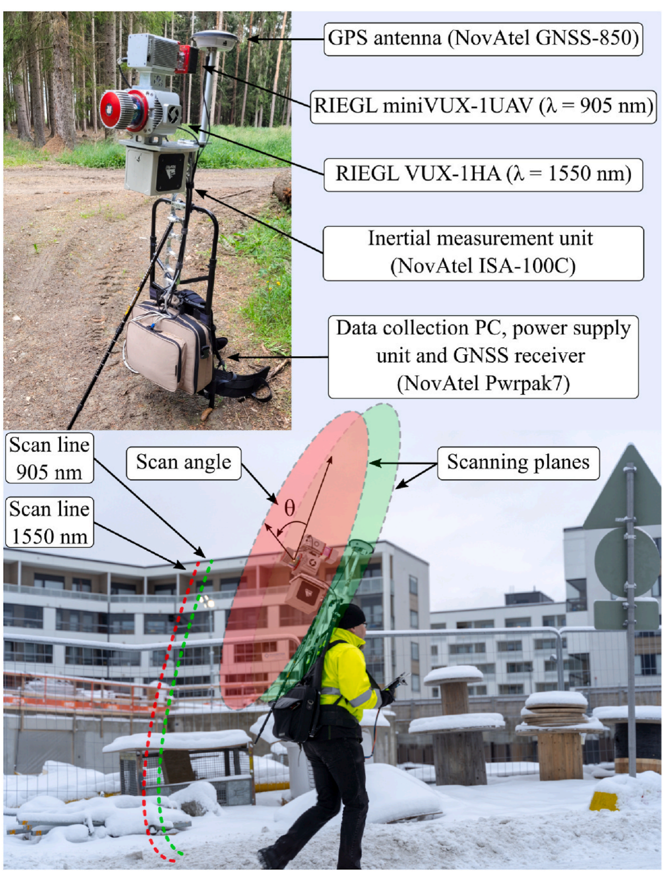


Fig. 1. Overview of the AkhkaR4-DW backpack laser scanning system Kukko et al. (2020). The system enables the simultaneous capture of high-density point clouds with two separate wavelengths ($\lambda_1 = 905$ nm and $\lambda_2 = 1550$ nm). The RIEGL VUX-1HA laser scanner provides $1.017 \cdot 10^6$ points/second, while the miniVUX-1UAV provides $0.1 \cdot 10^6$ points/second. Each point has the target reflectance value $r(\lambda)$ and the echo deviation σ as their attributes, in addition to the geometrical information $\mathbf{x} \in \mathrm{R}^3$.

measurements, the scanners record echo deviation values (ratio between the durations of received and transmitted pulses) for each received return signal. During the data collection, the VUX was set to a pulse repetition rate of 1 million pulses per second and a mirror frequency of 250 scans (complete revolutions of the mirror) per second, while the corresponding values for the miniVUX were set to 0.1 million pulses per second and 100 scans per second.

As shown in Fig. 1, the scanners are positioned on the backpack such that the scanning planes are parallel to each other and slightly tilted forward with respect to the walking direction. This results in two helix-shaped scanlines, with the miniVUX scanning plane preceding the VUX scanning plane by roughly 20 cm.

3.2. Dataset

The dataset was collected on 20th April 2020, a dry spring day, from the village of Masala (60.156 N, 24.532 E) located in southern Finland, a mixed urban environment containing buildings of different size, streets and street intersections, residential yards and various vegetation types. Laser scanning data was acquired while walking at a speed of 5 km/h and covering a total distance of 4 km, from which data collected during a trajectory section of 595 m was selected for manual labeling. The labelled point cloud shown in Fig. 2 spans a geographical area of roughly 36′800 m^2, containing 228 million points in total, of which 212 million were captured with the VUX and 16 million with the miniVUX.

The points in the point cloud were labelled manually to 13 classes: asphalt, soft ground, brick paving, building details, brick wall, vegetation, plastered wall, other man-made objects, curb, concrete wall, car, road marking and noise. The noise class contains points that do not correspond to any meaningful objects in the scene, typically isolated air points or artifacts from window reflections. An excerpt from the VUX dataset is shown in Fig. 3, showing some of the visual differences in reflectance and echo deviation between the various classes. Fig. 4 shows the benefits of using a multispectral laser scanning system. The characteristic reflectance of surface materials found in our dataset can be seen to differ considerably from class-to-class between the measurement wavelengths, therefore providing important cues about the class affiliation.

For supervised training, the dataset was split into training, validation and test sets as shown in Fig. 2. The training set contains 160 million points, while the validation and test sets contain 39 million and 29 million points, respectively. Both the validation and test sets were chosen such that they represent the complete dataset as good as possible, while minimizing overlap with the training test set. Additionally, to rule out any contamination between the training and test data, the test set comprises a point cloud section that is completely separate from the training and validation sets. The class-wise point distribution for each set is shown in Fig. 5.

3.3. Point cloud data processing

To enable the manual annotation of ground truth labels and multispectral dataset generation, data was transformed into a point cloud format. First, the GNSS and IMU data recorded during the data acquisition was post-processed with the NovAtel Waypoint Inertial Explorer software (version 8.9, NovAtel Inc., Canada) to produce an initial trajectory estimate. Post processed kinematic (PPK) correction was applied by placing a virtual GNSS base station from Trimnet service (RINEX 3.04) near the center of the trajectory. This trajectory was then combined with the raw point records from the laser scanners using the Riegl RiPROCESS software to produce an optimized trajectory and georeferenced point clouds. As a final step before manual labelling, the point cloud was prefiltered to contain only points with reflectance values between −20 dB and +5 dB (the minimum and maximum values in the raw data were −25 dB and 17.5 dB, respectively) and measurement range values between 2 m and 40 m.

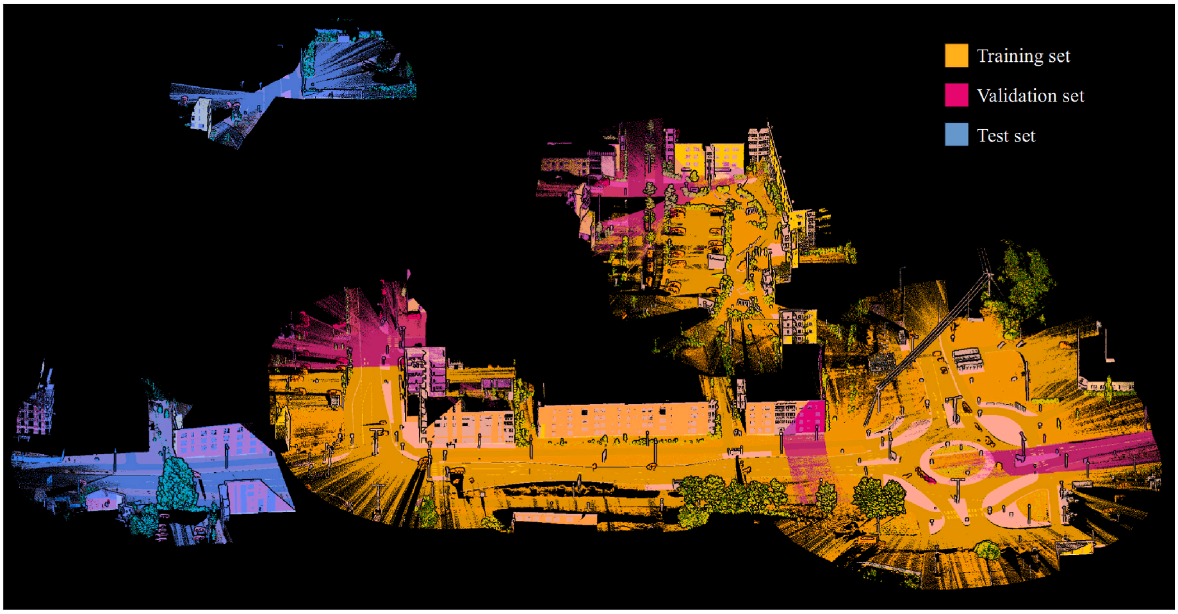
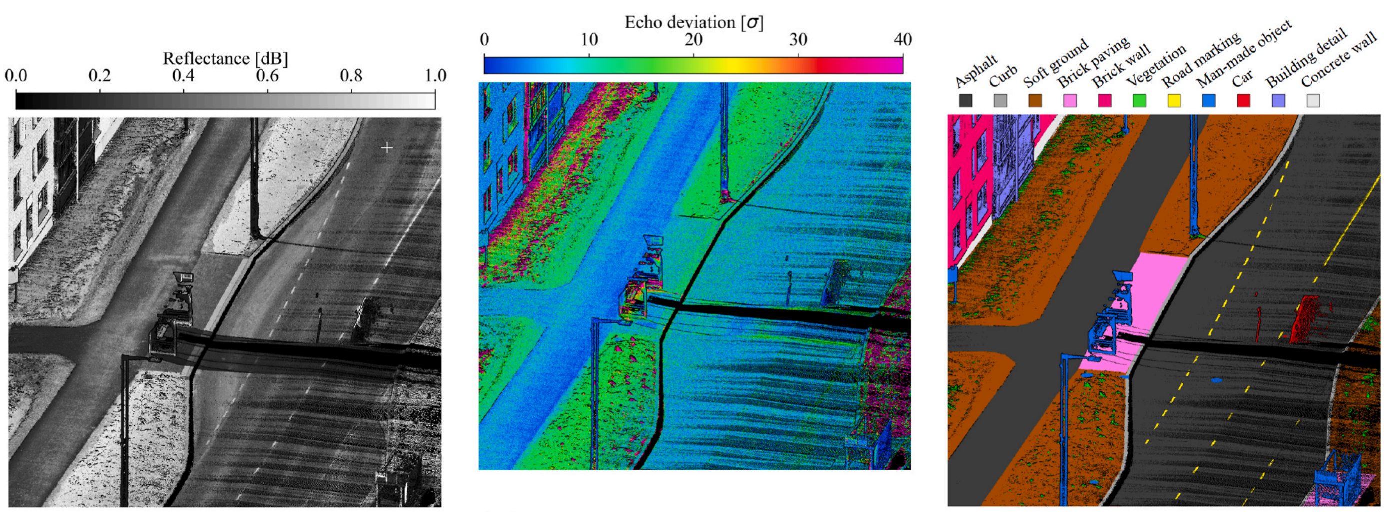


Fig. 2. The labelled dataset with color overlays indicating training, validation, and test sets. The test set, which represents 13% of the total data, is entirely isolated from the training (70% of total data) and validation (17% of total data) sets. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)



(a) Reflectance channel (b) Echo deviation channel (c) Class labels

Fig. 3. Excerpt from the VUX dataset, showing the differences between the reflectance and echo deviation channels. In addition to surface reflectance values provided by the reflectance channel, the echo deviation channel can give information on attributes such as surface roughness and object size with respect to the laser footprint size. Small echo deviation values are characteristic of points belonging to classes such as light poles and cars that are typically flat perpendicular to the laser beam and thus reflect a temporally compact return pulse. Large echo deviation values are produced by classes such as vegetation that cause the echo waveform to be more temporally dispersed.

Next, ground truth labels were manually annotated to the remaining points to enable supervised learning. The time consuming task of manual labelling was performed using the Terrascan tool (version 022.014) from Terrasolid running on Spatix. First, ground points were extracted from the point clouds by using Terrascan's iterative Hard Surface-routine based on a planar fit algorithm, which classifies dominant median surface points that form local horizontal planes. Then, non-ground classes were extracted by selecting 2D vertical slices of the point cloud and selecting points by drawing closed polygon shapes. Finally, the extracted ground points were classified to different ground classes using the same polygon drawing technique.

Some misclassifications especially between the vegetation and soft ground classes is to be expected, since the cut between these classes is often hard for even humans to make. Additionally, reflected points appearing inside buildings were prevalent due to the large number of building windows along the trajectory, which presented additional difficulty for the manual labelling process even with aggressive prefiltering.

The labelled points were then formatted to three different cylindrical projection formats; the single-scan, multi-scan and multispectral format,

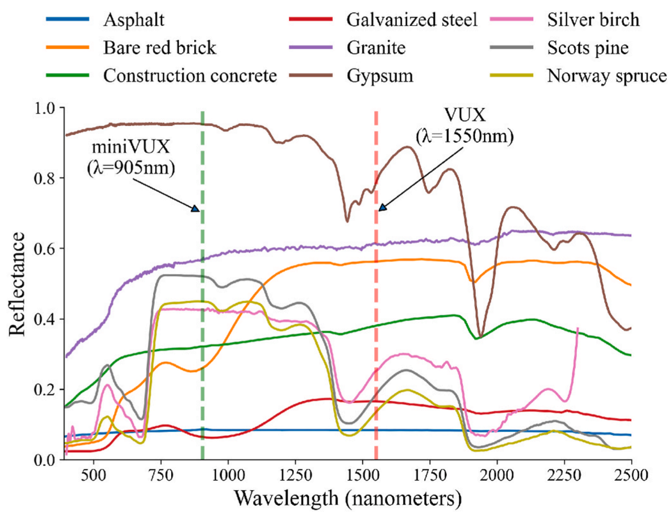


Fig. 4. Reflectance spectra for various organic and inorganic objects/materials present in our dataset. The reflectance data has been obtained from the ASTER spectral library (Baldridge et al. (2009)) and from the 25 boreal tree species spectral library (Hovi et al. (2017))). The wavelengths used by VUX and miniVUX laser scanners have been annotated in the plot. The ratios of the reflectance values at the measurement wavelength bands (905 nm and 1550 nm) differ considerably between the materials, which translates to distinct input data class distributions that are easier for the deep learning models to learn.

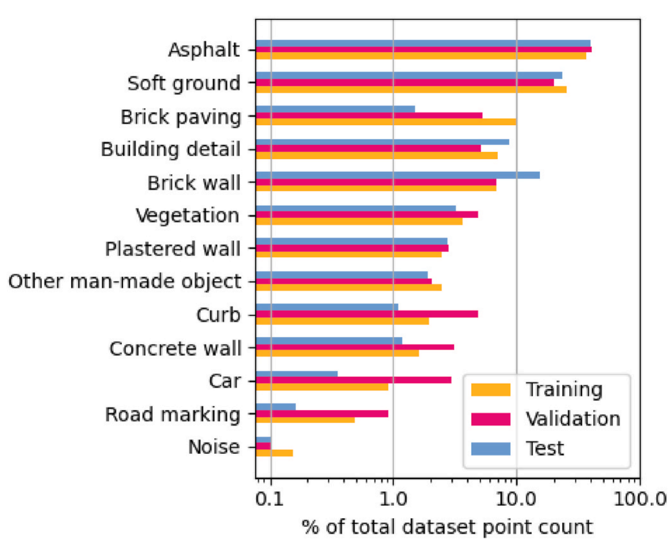


Fig. 5. Class distributions in the training, validation and test sets. Note the logarithmic scale.

the details of which have been summarized in Table 1. First, the point clouds were divided to individual scans by assigning each point a scan and pulse index using point-wise timestamp values provided by the laser scanner. Simultaneously, dummy points were added to regions with no echos, such as the sky, in order to keep the cylindrical projection undistorted. These points, as well as points that were omitted during the prefiltering, were included in the final input format but left unlabelled. Due to small fluctuations in the scanning mirror rotation frequency, scan extraction using timestamps only causes the scan starting points to drift slighly which essentially introduces random translations to the training data set, with an effect of reducing overfitting and making the trained model more robust.

3.3.1. *Single-scan input format*

The single-scan format was constructed by organizing each 1550 nm (VUX) scan into a 2D raster format with the echo number on the first axis, the pulse index on the second axis and the feature channels on the third axis. Although the maximum number of echoes per pulse is not limited, in our urban dataset the maximum number of echoes from one transmitted laser pulse was six. With three feature channels, namely range, reflectance and echo deviation, and a single target class per point, the input data format has the dimensions of (6, 5760, 3) and the output data format the dimensions of (6, 5760, 1), where 5760 is the maximum number of pulses per scan plus a safety margin due to scan length fluctuations. The training data was augmented by randomly flipping 20 % of the arrays along the second axis.

Algorithm 1. Nearest Neighbor search to construct a multispectral *Cloud* from a sparse $Cloud_{905nm}$ and a dense $Cloud_{1550nm}$. The difference in backward and forward time offsets is due to the offset of the scanners in the backpack (see Fig. 7a). The time offsets and maximum distance were chosen as conservatively as possible while finding a suitable nearest neighbor for 95% of the $Cloud_{905nm}$ points.

Table 2
Statistical properties of the measurement geometry parameters for the multispectral dataset as defined in Fig. 7. The values have been computed for each pair of points that form the multispectral point cloud. r is the mean of the two measurement ranges between the scanners and the point. S is the distance between the scanners during multispectral point acquisition. D is the distance of points that are fused together. φ is the angle between the scanning boresights of the two scanners during multispectral point acquisition.

Parameter	Mean	Standard deviation
r	8.47 m	7.49 m
S	0.188 m	0.194 m
D	8.22 mm	6.74 mm
φ	1.76°	2.23°

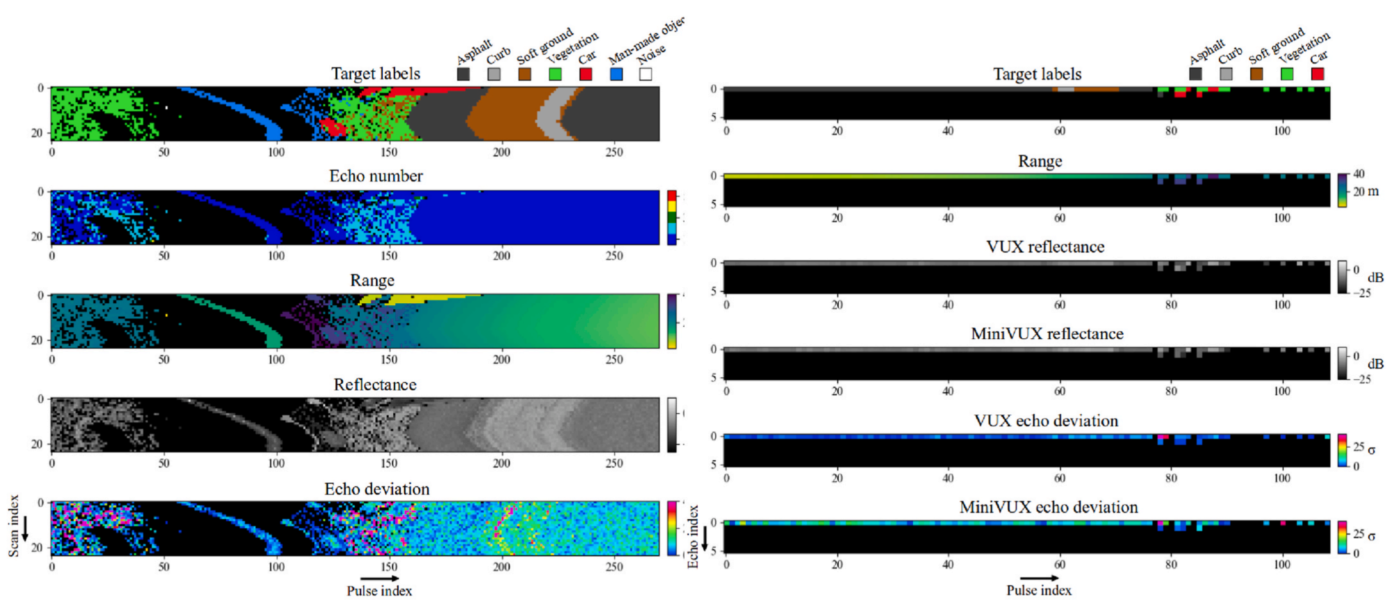
Table 1
Details of the neural network input formats used in this work. The single-scan and multi-scan formats contain only points captured with VUX ($\lambda = 1550$ nm), whereas the multispectral format contains data captured with both VUX and miniVUX ($\lambda = 905$ nm). The range r, reflectance $\mathbf{R}_\lambda$ and echo deviation σ_λ channels are used in all three input formats, while the number of returns (# of ret.) is only used in the multi-scan formats. The dataset size denotes the total number of input arrays constructed from the point cloud data and "Width" is the number of consecutive scans in each input array.

			Channels						
Format	Width	Input shape	r	$\mathbf{R}_{1550nm}$	$\mathbf{R}_{905nm}$	σ_{1550nm}	σ_{905nm}	# of ret.	# of unique input arrays
Single-scan	1	(6, 5760, 3)	✓	✓		✓			71066
Multi-scan	8	(8, 5760, 4)	✓	✓		✓		✓	8883
	16	(16, 5760, 4)	✓	✓		✓		✓	4441
	24	(24, 5760, 4)	✓	✓		✓		✓	2961
	32	(32, 5760, 4)	✓	✓		✓		✓	2220
Multispectral	1	(6, 1152, 5)	✓	✓	✓	✓	✓		28440

$Cloud \leftarrow$ empty data structure with dimensions of $Cloud_{905nm}$
$\Delta_{t+} \leftarrow 0.65$ s ▷ Maximum forward time offset
$\Delta_{t-} \leftarrow 0.45$ s ▷ Maximum backward time offset
$D_{max} \leftarrow 4$ cm ▷ Maximum distance
for $i \leftarrow 0, Cloud_{905nm}.size$ **do**
  $p \leftarrow Cloud_{905nm}[i]$
  $t_{min} \leftarrow p.timestamp - \Delta_{t-}$
  $t_{max} \leftarrow p.timestamp + \Delta_{t+}$
  $Candidates \leftarrow \left\{P \in Cloud_{1550nm} \mid P.timestamps \in [t_{min}, t_{max}]\right\}$
  $n \leftarrow$ nearest neighbor of p from $Candidates$ within D_{max} ▷ Get nearest neighbor
  if $p.label == n.label$ **then**
    $Cloud[i] \leftarrow (concatenate(p, n))$
  end if
end for

3.3.2. Multi-scan input format

The multi-scan format was constructed by stacking multiple consecutive 1550 nm scans together and including only the first return per pulse to keep the data two-dimensional. Thus, the pulse index is on the first axis, the scan index on the second axis and the feature channels on the third axis. To retain information about target laser cross section, the number of returns per pulse was added as an additional channel (small objects such as vegetation branches and leaves usually do not cover the whole laser footprint, resulting in detection of multiple echoes per pulse, whereas large objects such as planar surfaces return only one echo). With four feature channels, namely range, reflectance, echo deviation and number of returns, and a single target class per point, the input data format has the dimensions of (n, 5760, 4) and the output data format the dimensions of (n, 5760, 1), where n (hereafter referred to as width) is the number of consecutive scans stacked together. The different widths used in this work are provided in Table 1. The rationale behind the multi-scan format is to provide the convolutional neural network a spatially larger field-of-view (FOV) when compared to the narrow FOV of the single-scan format. This should improve model performance especially for complex datasets containing numerous classes and only a limited set of learnable features. To increase the amount of multi-scan training data, the training set was constructed by starting new arrays at intervals of *width*/2, which overlaps the consecutive 2D arrays by 50 % and doubles the training data. The training data was augmented by randomly flipping 20 % of the arrays along the second axis. For additional augmentation, randomly selected 30 % of the training data was sheared by rolling consecutive scans in an input array by 5 points with respect to each other along the second axis. A short section of a single multi-scan input array is shown in Fig. 6a, it clearly shows spatial relationships between the points of various classes.



(a) Multi-scan input format (b) Multispectral input format

Fig. 6. The feature array of the multi-scan input format (a) includes 4 channels, while the feature array of the multispectral input format (b) consists of 5 channels. The black pixels represent missing return pulses. Only a short section of a complete input array is shown, since a multispectral input array consists of 1152 pulses and a multi-scan input array consists of 5760 pulses which makes the arrays extremely elongated. The sparsity of the multispectral input format is due to the rarity of pulses that create multiple echos in an urban environment.

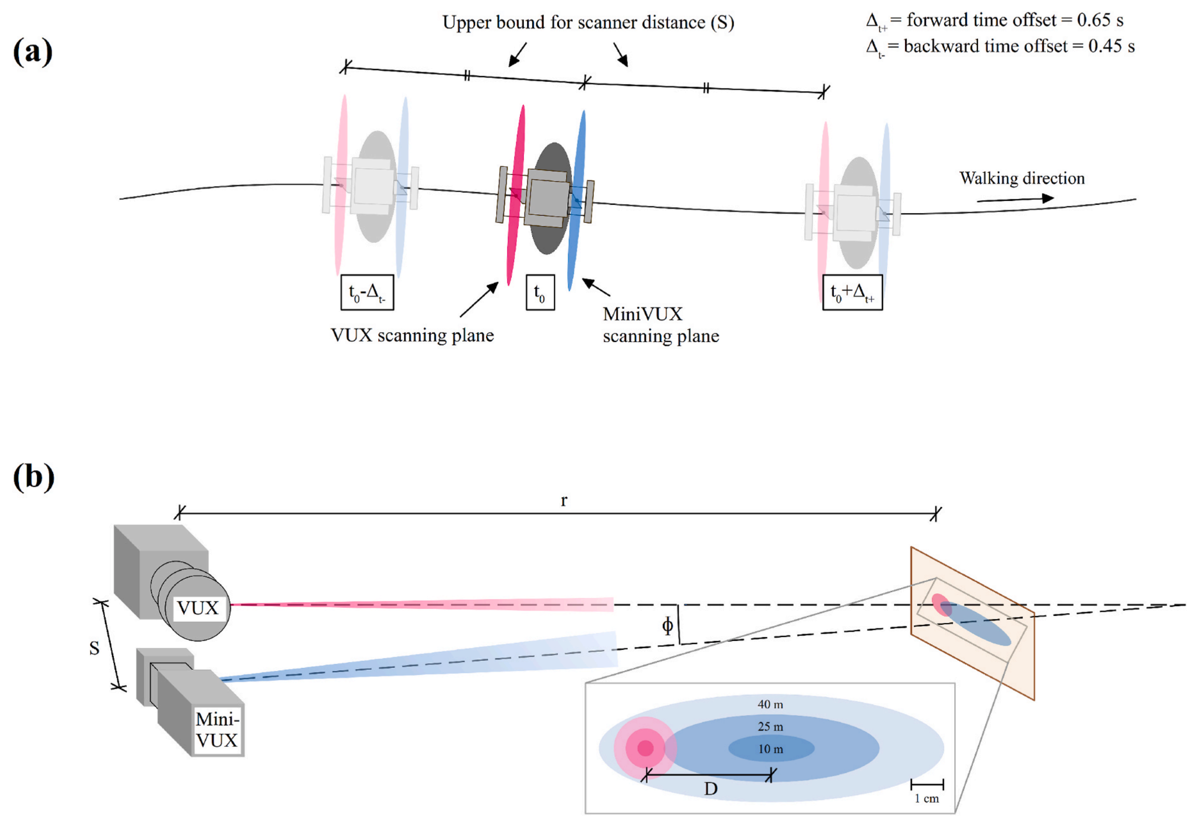


Fig. 7. Not drawn to scale. (a) Visualization of temporal filter procedure used in algorithm 1. The time offsets constrict the maximum distance between the miniVUX and VUX scanners (S) during the nearest neighbor search. (b) The measurement geometry of the multispectral setup and laser footprints at different distances. By limiting the maximum distance between the scanners (S), the minimum acquisition range (r) and point distance (D) during multispectral point acquisition, the dual scanner setup can be used to approximate a single multispectral lidar. Dataset statistics for the variables r, S, D and φ are listed in Table 2. The overlap of the laser beam footprints between the two scanners as a function of range is shown in Fig. 8.

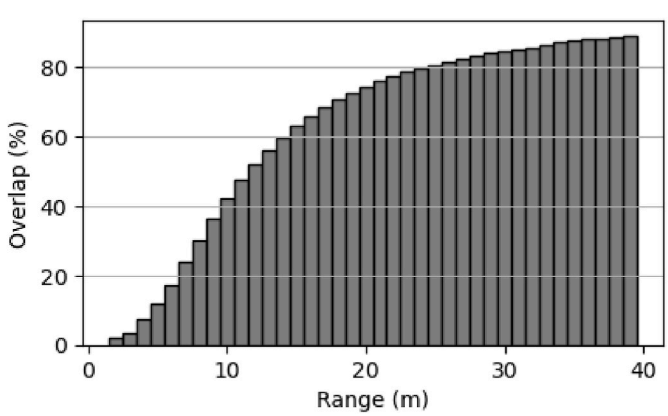


Fig. 8. The overlap of the laser beam footprints between the two scanners as a function of range (percentage of the area of the 1550 nm footprint covered by the 905 nm footprint). The overlap is substantial already at a 15 m distance from the backpack laser scanning system. Therefore, each nearest neighbor point pair in the multispectral dataset should correctly approximate the material specific reflectance spectrum when both points fall on a surface where the feature sizes perpendicular to the illumination direction are equivalent or larger than the transverse beam profile dimensions of the 905 nm laser scanner.

3.3.3. *Multispectral input format*

The multispectral format was created by combining the sparse 905 nm (miniVUX) point cloud with the much more dense 1550 nm (VUX) point cloud. By taking advantage of the large point cloud density difference between the two scanners, their close proximity in the backpack and their close to identical scanning patterns (see Fig. 1), for each point in the georeferenced 905 nm point cloud, it is highly probable to find a corresponding point from the georeferenced 1550 nm point cloud that is a close spatial neighbor and is acquired from an almost identical vantage point. By satisfying these two requirements in an otherwise static environment, the experimental setup proposed here can be used to approximate a single multispectral laser scanner that measures both wavelength channels via the same optical boresight. As shown in Figs. 7b and 8, this approximation quickly improves with growing measurement range due to the reducing angle φ between the scanning beams and larger overlap of the beam footprints.

Algorithm 1 was used to find the nearest neighbors to every 905 nm point from the 1550 nm point cloud. The temporal filter approach used in Algorithm 1 is visualized in Fig. 7a. If the maximum nearest neighbor distance and temporal filter window criteria were met, the two points were fused together to form multispectral points containing reflectance and echo deviation values from the both scanners. The statistics for scanner distances (S), measurement ranges (r), point distances (D) and beam angles (φ) as defined in Fig. 7 are listed in Table 2 for the whole dataset.

The fused multispectral points, containing 5 channels each, are placed in arrays almost identical to the single-scan format, with the echo number on the first axis, the pulse index on the second axis and the feature channels on the third axis (see Table 1). As with the single-scan format (1550 nm), also the 905 nm data contains at most 6 echoes per pulse. Thus, the input data format has the dimensions of (6, 1152, 5) and the output data format the dimensions of (6, 1156, 1), where 1156 is the maximum number of pulses per scan plus a safety margin due to scan length fluctuations. The training data augmentation was performed identically to the single-scan format. A short section of the multispectral input format is shown in Fig. 6b, showing the rather sparse structure of

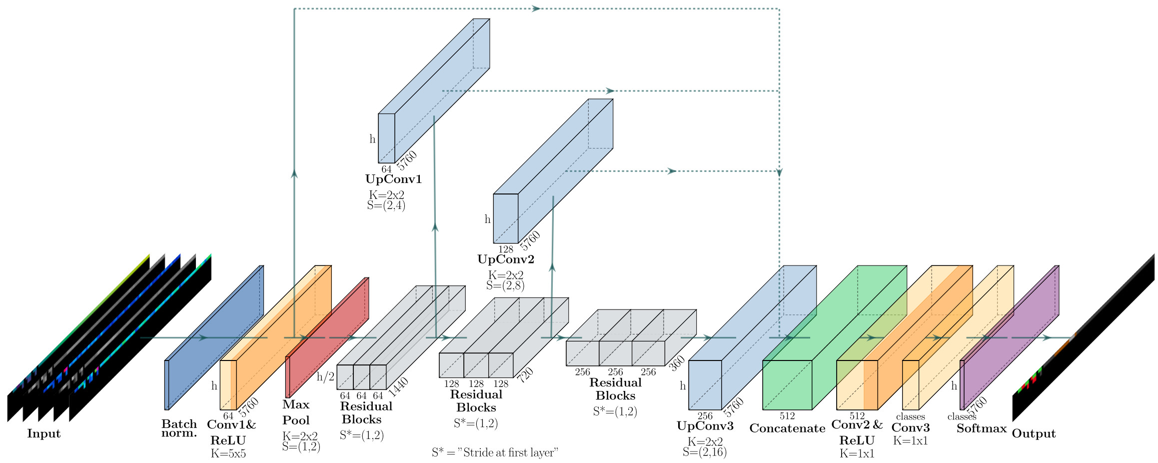


Fig. 9. A visualization of the Net1 point cloud semantic segmentation architecture. Fig. 10 provides a detailed visualization of the residual block.

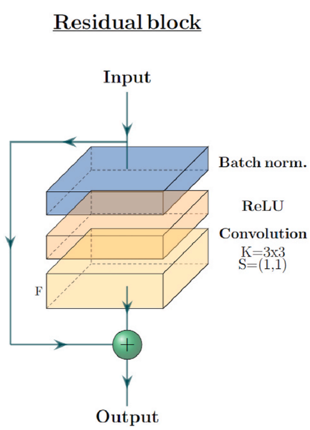


Fig. 10. A visualization of the residual block used in the Net1 architecture.

the data due to the rarity of multiple echoes per pulse in our dataset.

To allow comparing multispectral results to the monochrome results, two reference monochrome formats were created from the multispectral format. The 905 nm and 1550 nm reference monochrome formats were constructed by simply removing the channels corresponding to the other wavelength from the input arrays, and are otherwise exactly identical to the multispectral format.

3.4. Deep learning semantic segmentation models

A major advantage of the raw laser scanning input formats is that they are easily processed with projection based 2D convolutional deep learning models. We implemented two semantic segmentation architectures based on the LSSegNet1 and LSSegNet2 CNN architectures presented in Kaijaluoto et al. (2022):

Net1: visualized in Figs. 9 and 10. The Net1 model is closely based on the LSSegNet1 - architecture. In preliminary testing, we found out that the mean intersection over union (mIoU) metric on the validation set was maximized by adding a batch normalization layer to the input and omitting the spatial dropout and second batch normalization, relu and convolution layers from the original residual block architecture used in Kaijaluoto et al. (2022). Due to our different input format dimension the following changes were also done to the original architecture: The stride of the first MaxPool layer was set to S=(1,2), the stride of the convolution layer in the first residual block with F = 128 was set to S=(1,2) and the strides of the second and third UpConv2 layers were set to S=(2, 8) and S=(2,16), respectively. Other than these changes, the architecture of our Net1 architecture was kept equivalent to the original LSSegNet1 architecture.

Net2: visualized in Figs. 11 and 12. The Net2 model is identical to the LSSegNet2 architecture. Since our dataset contains more classes than the dataset used with the original LSSegNet2 architecture, we experimented with expanded architectures containing 5 or 6 encoder-decoder pairs as opposed to those 4 in LSSegNet2, in addition to standard grid search hyperparameter optimization described in Section 3.4.1. However, we did not find any mIoU performance improvements by deviating from the

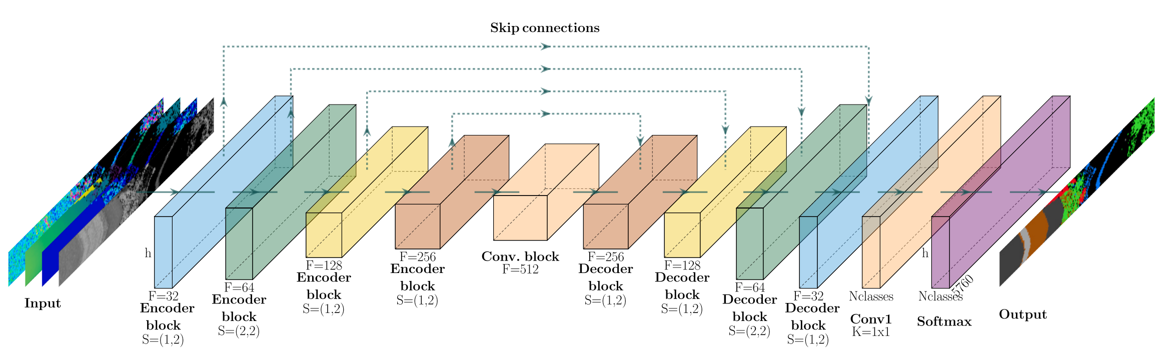


Fig. 11. A visualization of the Net2 point cloud segmentation architecture. The structure of the convolutional, encoder- and decoder blocks have been visualized in Fig. 12.

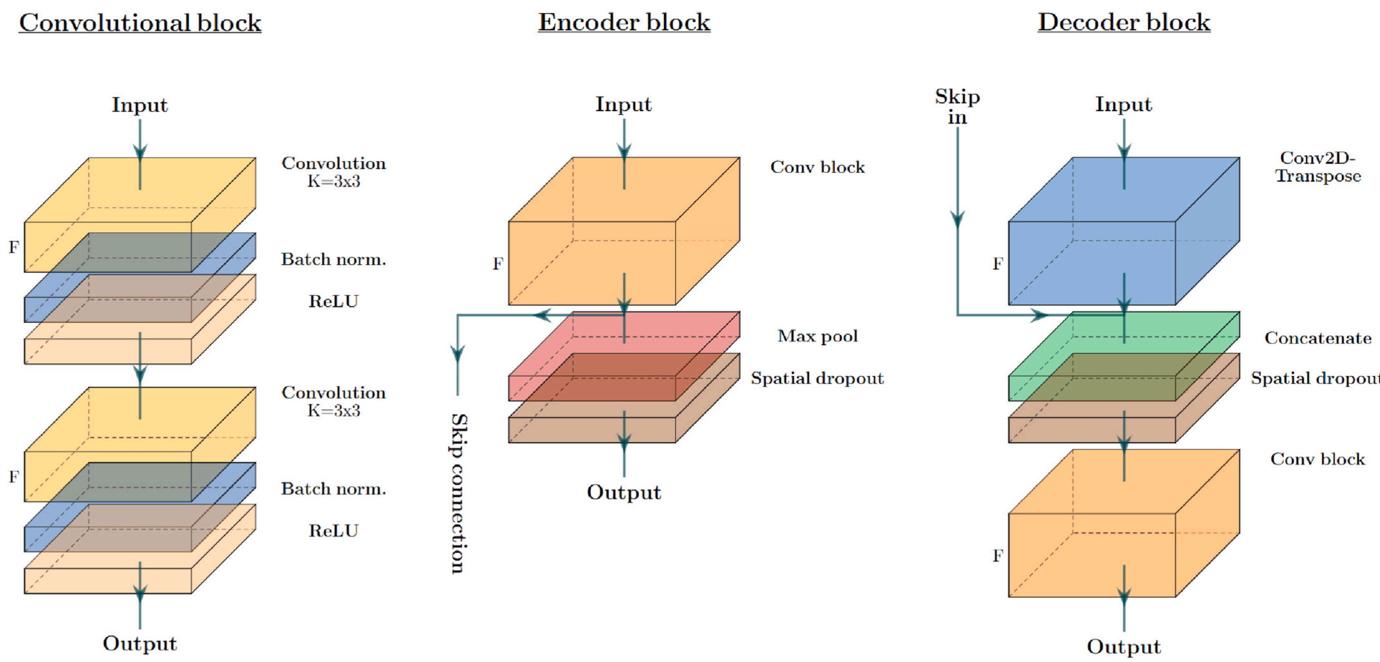


Fig. 12. A visualization of the Net2 architecture building blocks. The convolutional block (left-hand side), the encoder block (middle) and the decoder block (right-hand side).

original architecture.

Kaijaluoto et al. (2022) demonstrated in their paper that the LSSegNet2 architecture was capable of achieving equivalent, or better semantic segmentation performance in forest context when compared to the RandLA-Net (Hu et al. (2020)) point-wise 3D semantic segmentation architecture. Due to this observation, we do not extend our experiments to cover other deep learning architectures, but rather use the results obtained by Kaijaluoto et al. as a performance reference for the two implemented deep learning models.

Additionally, the purpose of this work is to provide comparison between single-scan and multi-scan formats and to examine the influence of the spectral information on the segmentation results, which can be accomplished without explicit comparisons with other semantic segmentation frameworks.

In preliminary testing, the following model architecture - input format pairs were found out to give the best mIoU performance, and were thus used for the experiments: Net1 for the single-scan and the multispectral experiments. Net2 for the multi-scan experiments.

3.4.1. Model training and hyperparameters

Tensorflow 2.9.1 with Keras was used as the deep learning framework to train and test the models. The Adam optimizer was used to minimize cross-entropy loss

$$\mathscr{L}(y,\hat{y}) = -\sum_{c=1}^{K} y_c \log(\hat{y}_c), \tag{1}$$

where y is the true and $\hat{y}$ is the predicted one-hot encoded class label. The cyclic learning rate policy used by Kaijaluoto et al. (2022) was not found to increase performance compared to a constant learning rate, which was set to 0.0004. The decay rates for the first and second moment estimates were left as the default $\beta_1 = 0.9$ and $\beta_2 = 0.999$, respectively.

For each network, a manual grid search was conducted to optimize hyperparameters (number of convolutional blocks and convolutional layers, number of encoder-decoder pairs and number of upconvolution-convolution pairs) with the objective to maximize the mean intersection over union (mIoU) metric over the validation set. Models were trained for a maximum of 40 epochs, and training was stopped if the cross-entropy loss over the validation set did not improve for at least 10 epochs. Finally, the weights yielding the lowest loss for the validation set were used to evaluate model performance on the test set. The full details of the computational hardware have been listed in Table 3.

Table 3
Details of the computational hardware and software.

Feature	Definition
CPU	Intel Xeon Gold 6234 (16 physical cores @3.30 GHz)
RAM	256 GB (DDR4 @3200 MHz)
GPU	Nvidia Quadro RTX 6000 (24 GB of memory)
Python version	3.10.4
CUDA version	11.7
Tensorflow version	2.9.1
Operating system	Ubuntu 20.04

4. Results and discussion

Three experiments were conducted to establish the feasibility of using deep learning and raw laser scanning data input format for semantic segmentation of objects in urban environment. First, we evaluated the importance of context information by comparing single scan inputs to multiple scan inputs. Second, we evaluated the usefulness of spectral information in a single scan semantic segmentation task. Finally, we characterized the computational performance and memory requirements of the implemented deep learning models. An example view of the semantically segmented test set is provided in Fig. 13.

We did not experiment with 3D convolution although suggested by Kaijaluoto et al. (2022). Since our dataset contains only a relatively small number of pulses with more than one echo, this would make the 3D input format extremely sparse compared to the multi-scan format. However, for a dataset containing a larger number of multi-echo pulses, a 3D convolution approach could be a sensible extension to the multi-scan format where only the first echoes are included.

4.1. Results of single-scan and multi-scan experiments

The class-wise intersection over union (IoU) and mean intersection

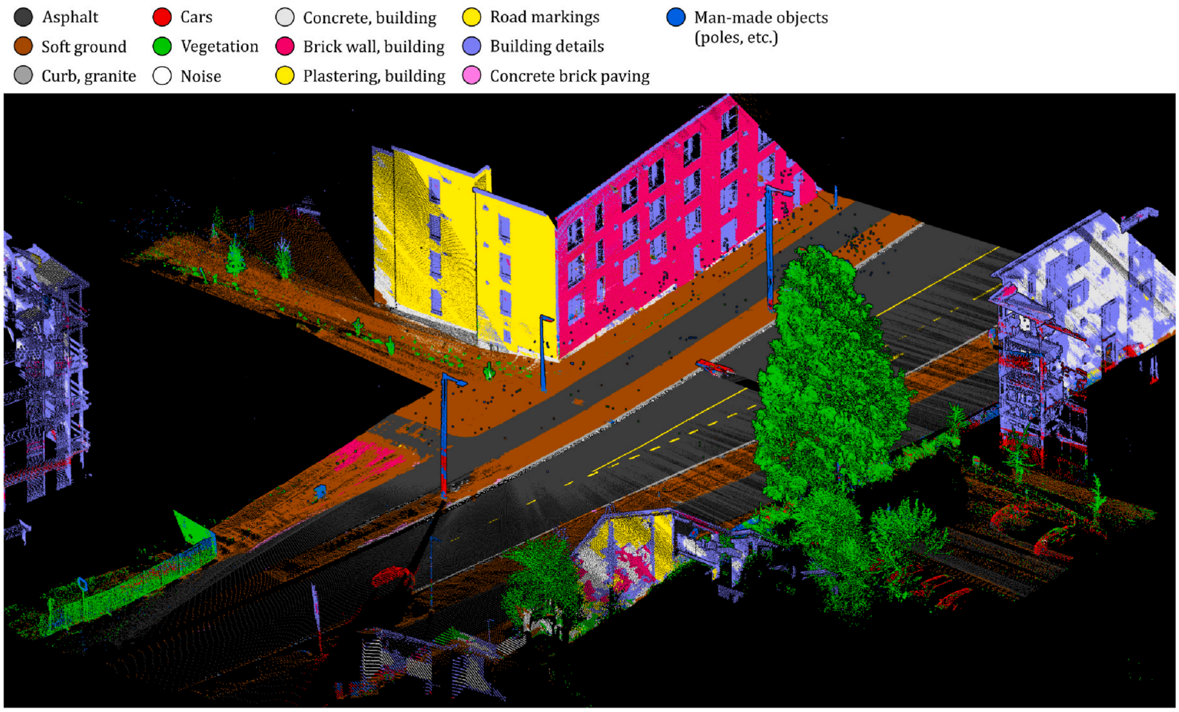


Fig. 13. General view of the test set environment with class predictions given by the Net2 architecture (width = 24). Most of the majority classes, such as building walls, asphalt and vegetation, have been segmented with good results. Misclassifications can be observed in geometrically challenging locations where the point density is low, for example, on street light poles.

over union (mIoU) values shown in Table 4 suggest that models trained with the multi-scan format outperform models trained with the single-scan format in every class except the noise class. Furthermore, the multi-scan format with width 24 achieves clearly better IoU and mIoU values than formats of width 8, 16 or 32. The multi-scan formats seem to benefit from increased width in classes that are prevalent in the dataset, such as asphalt and soft ground. However, classes that are more rare in the dataset, such as noise, car and other man-made objects, do not consistently perform better with growing multi-scan width. This could be due to the small sample size leading to increased variability, or because most of these classes display features at a small scale, reducing the advantages attainable by a large field-of-view. Conversely, model performance for classes that form large planar surfaces, such as brick paving, plastered wall and road marking, seem to benefit most from the increasing spatial FOV of the wider multi-scan formats, which might indicate that the neural network models are especially well suited to learn from linear features in the data.

Detailed views of the predicted test set are shown in Fig. 14. As already noted in Kaijaluoto et al. (2022), the context-deprived single--scan format suffers most notably from so called "salt and pepper" misclassification (see Fig. 14a), where individual points are sporadically misclassified, even though the majority are classified correctly. As can be seen from the multi-scan predictions in Fig. 14b, this mode of misclassification is much more rare when the model can leverage a larger spatial context and thus larger spatial features. Also, class boundaries are much less vague in the multi-scan predictions when compared to the single-scan predictions. However, while the added spatial context reduces the misclassification of random, spatially dispersed points, it tends to misclassify larger patches of neighboring points (highlighted area in Fig. 14b), which might be more difficult to detect using simple metrics such as classification probability.

The imbalanced class distribution of the dataset is clearly visible in the classification results. This is especially evident from the confusion matrices in Fig. 15 with most off-diagonal misclassified entries being in the lower left of the matrices, where the minority classes reside. Minority classes in the dataset are often misclassified to more prevalent classes, whereas more prevalent classes are only seldom misclassified to minority classes. Interestingly, the single-scan format is the strongest of the formats in classifying noise points. Most of the noise points are window reflections, which are spatially close to points acquired from windows and window frames, which belong to the building details class. The confusion matrices in Fig. 15 show, that the multi-scan formats often misclassify noise points to the building details class. We hypothesize that this is caused by the multi-scan models classifying points primarily by their spatial neighborhood, whereas the single-scan models are forced to learn more from the point-wise channels (range, reflectance, echo deviation and return number), that might be better indicators for points acquired via a window-reflected path.

The multi-scan classification accuracy does not deteriorate significantly in areas that are captured during large angular rate of changes in the trajectory, such as the lower part of Fig. 13. Since the data was collected from a moving backpack, the trajectory is not linear, but includes constant angular movements, which from the lidars point of view accumulate to large jerky translations that quickly increase with growing measurement range. Since the multi-scan format is constructed by stacking consecutive scans together without correcting for these considerable non-linear movements, individual input arrays are randomly skewed and stretched. It could be argued that this intrinsic "augmentation" of the input data makes the trained models more robust to variations in the data and reduces the assumptions that need to be made concerning the type of data fed to the network. To explore this hypothesis, we fed raw 1550 nm (VUX) data collected from a similar urban environment using the Roamer-R4DW mobile laser scanning system (El Issaoui et al. (2021)) to the network. The obtained results were by qualitative comparison very close to the results observed in the backpack test set, although the network was trained only with the backpack data. These results are not presented here due to their preliminary nature, but exploring the invariance of the method to the data collection platform certainly presents an interesting direction for further research.

Table 4
The results from the experiments with the single-scan and the multi-scan models. The increased spatial FOV in the multi-scan models increases the segmentation performance when compared to the original single-scan model.

Format	Width	Network	mIoU	Asphalt	Soft ground	Brick paving	Building detail	Brick wall	Vegetation	Plastered wall	Other man-made object	Curb	Concrete wall	Car	Road marking	Noise
Single	1	Net1	45.4	86.6	70.7	36.3	43.3	58.4	61.1	74.5	24.0	22.3	6.6	13.8	37.1	**55.3**
Multi	8	Net2	59.2	93.1	85.8	75.4	65.7	76.5	**73.3**	76.8	**32.7**	38.6	**35.3**	49.1	47.0	20.2
	16	Net2	51.9	90.7	82.9	38.3	64.8	73.0	72.8	56.9	30.9	38.2	29.9	**50.3**	33.3	13.1
	24	Net2	**62.1**	**96.1**	**89.8**	**90.1**	**70.5**	76.2	71.8	**83.2**	32.3	**58.3**	15.7	48.8	57.7	16.2
	32	Net2	60.7	94.9	89.6	82.1	67.4	**77.9**	71.1	80.3	28.7	53.8	10.2	46.4	**66.8**	19.3

4.2. Results of multispectral experiments

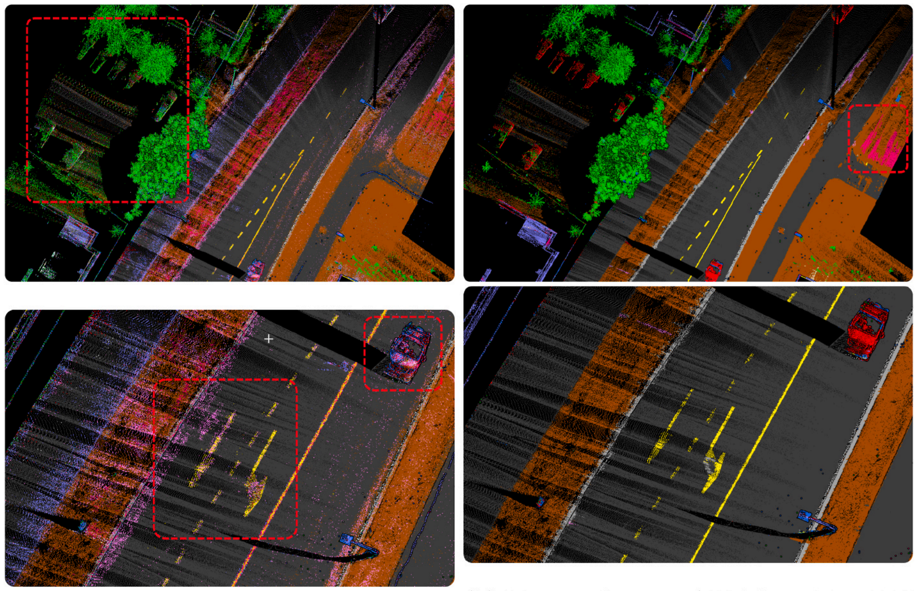
The IoU results for the multispectral input format outperform the corresponding reference monochrome formats in almost every class as shown in Table 5. The advantages of the spectral data are most clearly visible in the classification accuracy of large, planar classes, such as asphalt, where the multispectral format achieves an IoU value of 81.9 % as compared to 72.8 % (1550 nm) and 46.7 % (905 nm), or soft ground, where an IoU value of 72.0 % is achieved as compared to 59.5 % (1550 nm) and 42.9 % (905 nm). Since the geometric features of the aforementioned classes differ little from each other, the models using only one wavelength contain/base on limited amount of point-wise information, and thus struggle more at discerning between the classes. Conversely, the models utilizing both wavelengths can in addition learn from the differences in reflectance spectra between these classes. This is also shown in the confusion matrices in Fig. 16, where the asphalt, soft ground, brick paving, brick wall, concrete wall and road marking classes are mostly classified correctly with the multispectral model but mixed up frequently by the monochrome models.

Correspondingly, the classification performance for classes, such as vegetation and other man-made objects, that display more unique geometric properties show less improvement with the multispectral format. However, even these classes perform better since the limited contextual information included in the single-scan-based format means that the models can learn only from limited geometric information. This demonstrates the critical role that point-wise information plays in our context-deprived input format. For example, a recent study by Vierhub-Lorenz et al. (2022), using a point-based 3D U-Net architecture on a multispectral point cloud semantic segmentation task in a similar urban environment as this work, demonstrated that the spectral information increased IoU classification performance for planar ground classes, but found no significant increase in the IoU values for classes with more unique geometric features since the 3D CNN architecture was able to leverage the geometric properties significantly better than our approach, reducing the relative contribution of the multispectral data.

The excerpts from predicted point clouds in Fig. 17 show, that the classification accuracy of the road marking class visibly improves with the multispectral data. Road markings exhibit very little spatial features that discern them from the underlying road surface, and thus the segmentation networks have to rely heavily on the reflectance and echo deviation channels. Most likely, the road marking paint in our dataset has substantial differences in the reflectance values between the 905 nm and 1550 nm wavelengths, and therefore, the models can leverage the additional spectral information effectively.

Since the multispectral format is identical to the single-scan format up to the additional reflectance and echo deviation channels, the predictions in Fig. 17 display similar spatially randomly misclassified points as the single-scan predictions in Fig. 14.

The reference results for the 1550 nm wavelength alone generally achieve higher IoU values than the 905 nm wavelength alone, even though the point distributions are exactly the same. A contributing factor for this difference could be the smaller laser footprint of the 1550 nm scanner beam, which results in a higher spatial precision compared to the 905 nm scanner beam. Also, the differences between the laser pulse intensities might play a role. A recent study by Taher et al. (2022) on single-photon sensitive hyperspectral lidar demonstrated that weak return pulses estimate the target reflectance values with lower accuracy than strong, high-intensity return pulses. Due to the more strict eye-safety criterions at the miniVUX's 905 nm wavelength, the return pulse intensities at the VUX's 1550 nm wavelength range are most probably substantially higher, which in turn reduces the uncertainty in the estimated target reflectance values. Thirdly, the differences in the material specific reflectance spectra for classes used in this work might be more pronounced at the 1550 nm wavelength range than in the 905 nm wavelength range, which could mean that the 1550 nm wavelength used by the VUX is more effective in discerning between classes in an



(a) Net1 single-scan (mIoU = 45.4)

(b) Net2 multi-scan (Width = 24, mIoU = 62.1)

Fig. 14. Visual comparison of the single-scan model and the multi-scan model predictions. The multi-scan model performs better with the class specific local structures, the red boxes indicate a few chosen examples of these errors. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

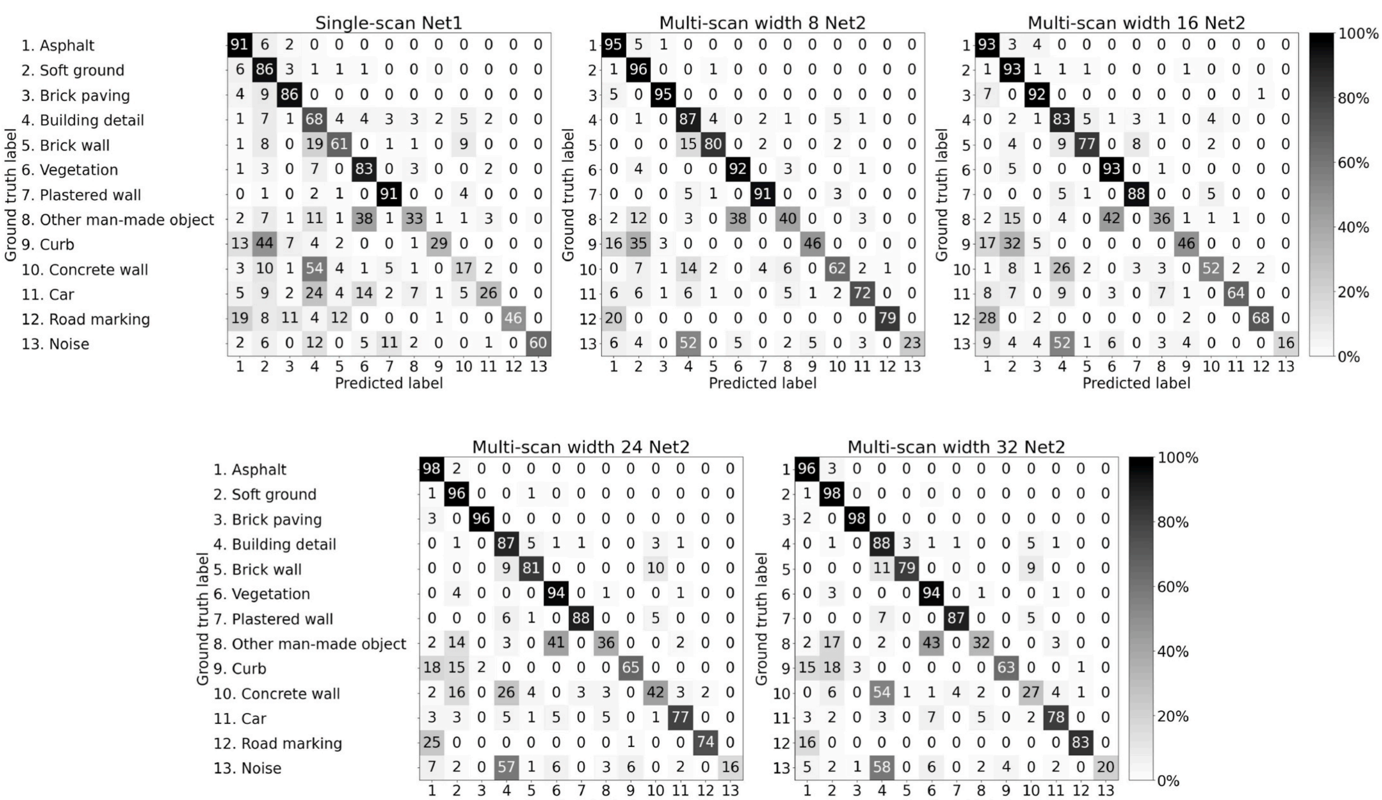


Fig. 15. Confusion matrices of the single-scan and multi-scan test results. The classes are presented in descending order according to the number of points. The normalization is over the predicted labels.

urban environment.

Tests where the multispectral format was combined with the multi-scan format were also performed using both of the network architectures, but these yielded clearly worse results than any of the results represented here. We concluded that this was caused by the limited size of the dataset, which became unacceptably small when only using the first returns from the already sparse 905 nm point cloud. However, this approach could be explored further especially with a denser and larger

Table 5
Segmentation results in the test set for the multispectral experiments. Net1 segmentation architecture was used in the experiments.

Channels	mIoU	Asphalt	Soft ground	Brick paving	Building detail	Brick wall	Vegetation	Plastered wall	Other man-made object	Curb	Concrete wall	Car	Road marking	Noise
1550 nm	34.1	72.8	59.5	16.4	42.7	43.7	53.2	56.4	17.7	15.5	5.1	**8.4**	36.6	15.7
905 nm	25.4	46.7	42.9	5.2	37.0	17.7	54.2	45.1	11.1	10.6	13.5	1.2	35.9	9.4
905 nm + 1550 nm	**43.5**	**81.9**	**72.0**	**29.4**	**48.0**	**51.1**	**57.7**	**70.6**	**20.5**	**24.8**	**15.4**	7.4	**65.8**	**20.8**

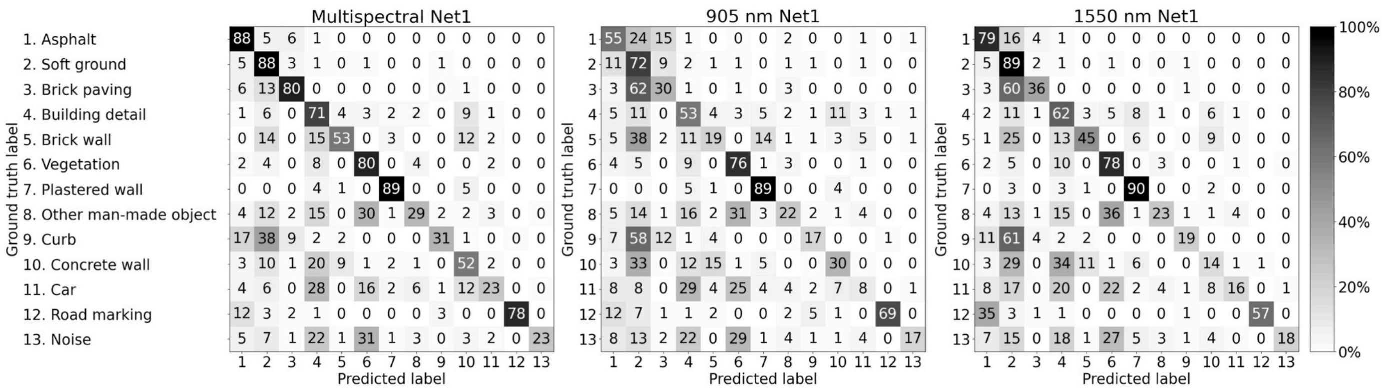


Fig. 16. Confusion matrices of the multispectral test results. The classes are presented in descending order according to the number of points. The normalization is over the predicted labels.

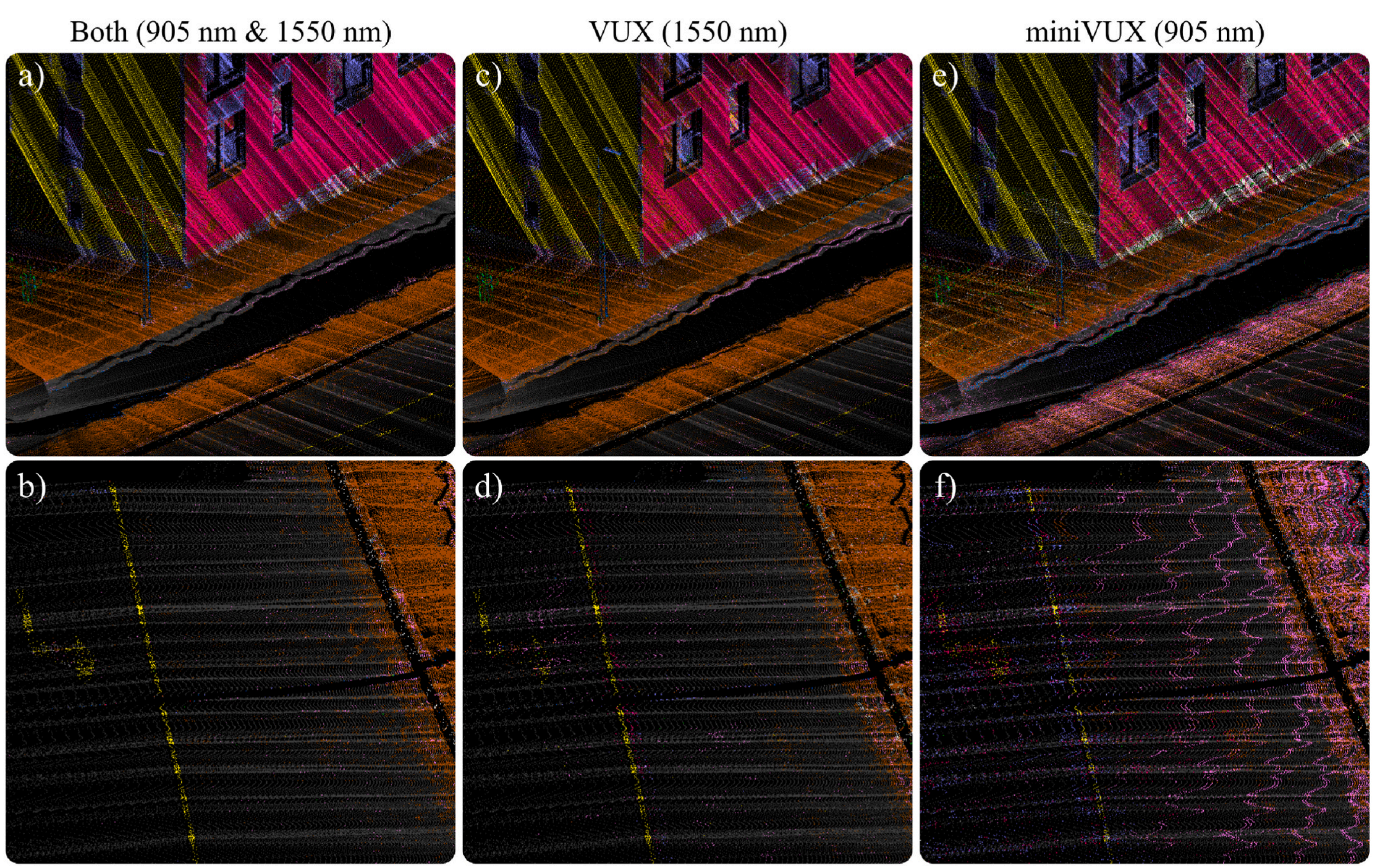


Fig. 17. Visual comparison of the class predictions given by the Net1 architecture between the multispectral point cloud (left-hand side), the VUX point cloud at a wavelength of 1550 nm (middle) and the miniVUX point cloud at a wavelength of 905 nm (right-hand side). The multispectral input data reduces the amount of class mixing in local neighbourhoods, which is evident on the road markings in Figures d) and f). Figure b) shows clear and concise road marking boundaries when compared to either of the single-wavelength input formats.

Table 6
Model sizes, inference speeds and average forward pass times. The inference speed is the inference frequency of the GPU when a batch of input arrays, approximately 1/10th of a second worth of data, is passed through the model once. The average forward pass time is the reciprocal of the inference speed.

Network	Parameters	Model size (MB)	Input format	Inference speed (scans/s)	Avg forward pass time (ms/scan)
Net1 (3 channels)	2 962 730	11.30	Single-scan	1333.3	0.75
Net2	7 773 514	29.65	Multi-scan (width 8)	484.8	2.06
			Multi-scan (width 16)	883.9	1.13
			Multi-scan (width 24)	1471.3	0.68
			Multi-scan (width 32)	1693.1	0.59
Net1 (5 channels)	2 965 930	11.31	Multispectral	617.3	1.61

multispectral dataset.

4.3. Results for model performance

The number of parameters, sizes and prediction speeds for the best performing models for each input format are illustrated in Table 6. The parameter count includes the trainable and non-trainable parameters and the size represents the GPU memory consumption of the model using 32-bit single-precision floating-point parameters. To enable speed comparisons between the models, the inference speed metric was determined by using an inference batch size of approximately 1/10th (±36% depending on the input format) of a second worth of data, which corresponds to a realistic prediction frequency for most real-time applications. Since the single-scan and multi-scan formats consist of VUX data that operates at 250 scans per second and the multispectral format represents miniVUX data that operates at 100 scans per second, all the models presented in the table are capable of real-time processing with our computational hardware and a 1/10th scans per second batch size. If the input arrays are fed to the GPU one-by-one, only the multi-scan-based formats are capable of real-time processing. This is due to the inclusion of only the first returns per pulse in the multi-scan format, which decreases the number of input array elements by $7/9^{\text{ths}}$ when compared to the single-scan format.

Our results clearly outperform inference speed results of under 60 Hz as reported by Kaijaluoto et al. (2022), mostly due to more powerful computational hardware and the more compact and less sparse multi-scan format.

When considering portable onboard hardware applications, model sizes and inference speeds could be further reduced by using mixed- or half-precision numerical formats that store the model and perform inference calculations at least partially using a 16-bit floating-point representations instead of the 32-bit floating-point representation used in this work. Compared to semantic segmentation using point-wise 3D approaches, our method reduces the complexity of the semantic segmentation pipeline drastically by not requiring any trajectory estimate and point cloud georeferencing, speeding up the process even further. Further speed benefits downstream could be achieved for real-time/time critical applications by using a lower capacity deep learning model and outputting only classes relevant for the application for further processing.

5. Conclusions

Our work showcases the possibility of performing real-time semantic segmentation on non-georeferenced 2D laser scanner data in urban environments using both single-wavelength and multispectral lidar data. We introduce deep learning input formats containing increased spatial context (multi-scan format) and data from two scanners with different wavelengths (multispectral format), and compare their performance against the single-scan input format presented by Kaijaluoto et al. (2022). Our best performing method achieves a mean intersection over union (mIoU) value of 62.1 for a test dataset consisting of 13 distinct classes. This level of prediction accuracy is high when considering the unprocessed format of the 2D input data but the result stands comparison even against state-of-the-art methods for real-time semantic segmentation of 3D point clouds proposed by Hu et al. (2020) and Kong et al. (2023).

After conducting experiments with various input data representations and combinations, we have arrived at the following conclusions.

1. Stacking multiple consecutive scan lines together into a 2D raster format, with the goal of increasing the spatial context, improved the semantic segmentation results substantially. Test set mIoU values increased from 45.4 with the single-scan format to 62.1 with the multi-scan format.
2. Multispectral lidar data, that was obtained by combining measurements from separate laser scanners operating at two different wavelengths, increased semantic segmentation accuracy when compared to single-wavelength reference results. The models using multispectral lidar data with wavelengths 905 nm and 1550 nm achieved a test set mIoU value of 43.5, which is significantly higher than either of the corresponding single-wavelength results of 34.1 (1550 nm) or 25.4 (905 nm).

All tested models demonstrated the capability of real-time inference. Our methods enable semantic segmentation pipelines that do not rely on a georeferenced point cloud, making them computationally effective since they require only minimal preprocessing of the raw scanner data and eliminate the need for trajectory estimate. This makes our methods well suited for real-time semantic segmentation of single line laser scanning data, such as data produced by survey scanners, especially in settings where a precise trajectory estimate is not instantaneously available.

Author contributions

Data collection, A.K. and H.K; data processing, M.R., A.K. and H.K.; methodology, M.R.; conceptualization, M.R. and J.T.; writing and editing, M.R., J.T. and P.M.; visualizations, M.R., J.T., P.M. and A.K.; funding, J.H., A.K. and H.K.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

We gratefully thank Academy of Finland project “UNITE Forest-Human-Machine Interplay - Building Resilience, Redefining Value Networks and Enabling Meaningful Experiences” (decision number 337656) and Henry Ford Foundation for funding.

References

Babahajiani, P., Fan, L., Kämäräinen, J.-K., Gabbouj, M., 2017. Urban 3D segmentation and modelling from street view images and LiDAR point clouds. Mach. Vis. Appl. 28 (7), 679–694.

Balado, J., Van Oosterom, P., Díaz-Vilariño, L., Arias, P., 2021. Semantic segmentation of mobile laser scanning point clouds with long short-term memory networks: preliminary results. Int. Arch. Photogram. Rem. Sens. Spatial Inf. Sci. 43, 123–130.

Baldridge, A.M., Hook, S.J., Grove, C., Rivera, G., 2009. The ASTER spectral library version 2.0. Rem. Sens. Environ. 113 (4), 711–715.

Behley, J., Garbade, M., Milioto, A., Quenzel, J., Behnke, S., Stachniss, C., Gall, J., 2019. SemanticKITTI: a dataset for semantic scene understanding of LiDAR sequences. In: Proc. Of the IEEE/CVF International Conf. on Computer Vision (ICCV).

Bosché, F., 2010. Automated recognition of 3D CAD model objects in laser scans and calculation of as-built dimensions for dimensional compliance control in construction. Adv. Eng. Inf. 24 (1), 107–118.

Bozchalooi, I.S., Youcef-Toumi, K., U.S. Patent US10649072B2, 2020. LiDAR Device Based on Scanning Mirrors Array and Multi-Frequency Laser Modulation.

Chen, G., Wang, M., Yang, Y., Yu, K., Yuan, L., Yue, Y., 2023. Pointgpt: Auto-Regressively Generative Pre-training from Point Clouds arXiv preprint arXiv:2305.11487.

Chen, Y., Räikkönen, E., Kaasalainen, S., Suomalainen, J., Hakala, T., Hyyppä, J., Chen, R., 2010. Two-channel hyperspectral LiDAR with a supercontinuum laser source. Sensors 10 (7), 7057–7066.

Costabile, P., Costanzo, C., De Lorenzo, G., De Santis, R., Penna, N., Macchione, F., 2021. Terrestrial and airborne laser scanning and 2-D modelling for 3-D flood hazard maps in urban areas: new opportunities and perspectives. Environ. Model. Software 135, 104889.

Ekhtari, N., Glennie, C., Fernandez-Diaz, J.C., 2018. Classification of airborne multispectral LiDAR point clouds for land cover mapping. IEEE J. Sel. Top. Appl. Earth Obs. Rem. Sens. 11 (6), 2068–2078.

El Issaoui, A., Feng, Z., Lehtomäki, M., Hyyppä, E., Hyyppä, H., Kaartinen, H., Kukko, A., Hyyppä, J., 2021. Feasibility of mobile laser scanning towards operational accurate road rut depth measurements. Sensors 21 (4), 1180.

Enayetullah, H., Chasmer, L., Hopkinson, C., Thompson, D., Cobbaert, D., 2022. Identifying conifer tree vs. deciduous shrub and tree regeneration trajectories in a space-for-time boreal peatland fire chronosequence using multispectral lidar. Atmosphere 13 (1), 112.

Engel, N., Belagiannis, V., Dietmayer, K., 2021. Point transformer. IEEE Access 9, 134826–134840.

Geiger, A., Lenz, P., Urtasun, R., 2012. Are we ready for autonomous driving? The KITTI vision benchmark suite. In: Proc. Of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), pp. 3354–3361.

Gong, W., Sun, J., Shi, S., Yang, J., Du, L., Zhu, B., Song, S., 2015. Investigating the potential of using the spatial and spectral information of multispectral LiDAR for object classification. Sensors 15 (9), 21989–22002.

Goyal, A., Law, H., Liu, B., Newell, A., Deng, J., 2021. Revisiting point cloud shape classification with a simple and effective baseline. In: International Conference on Machine Learning. PMLR, pp. 3809–3820.

Hakala, T., Suomalainen, J., Kaasalainen, S., Chen, Y., 2012. Full waveform hyperspectral LiDAR for terrestrial laser scanning. Opt Express 20 (7), 7119–7127.

Hakula, A., Ruoppa, L., Lehtomäki, M., Yu, X., Kukko, A., Kaartinen, H., Taher, J., Matikainen, L., Hyyppä, E., Luoma, V., et al., 2023. Individual tree segmentation and species classification using high-density close-range multispectral laser scanning data. ISPRS Open J. Photogrammet. Rem. Sens. 9, 100039.

Hall, D.S., U.S. Patent Us8675181B2, 2014. Color LiDAR Scanner.

Helbich, M., Jochem, A., Mücke, W., Höfle, B., 2013. Boosting the predictive accuracy of urban hedonic house price models through airborne laser scanning. Comput. Environ. Urban Syst. 39, 81–92.

Honkavaara, E., Hakala, T., Markelin, L., Jaakkola, A., Saari, H., Ojanen, H., Pölönen, I., Tuominen, S., Näsi, R., Rosnell, T., et al., 2014. Autonomous hyperspectral uas photogrammetry for environmental monitoring applications. In: ISPRS Archives. International Society for Photogrammetry and Remote Sensing (ISPRS).

Hovi, A., Raitio, P., Rautiainen, M., 2017. A spectral analysis of 25 boreal tree species. Silva Fenn. 51 (4).

Hu, Q., Yang, B., Xie, L., Rosa, S., Guo, Y., Wang, Z., Trigoni, N., Markham, A., 2020. Randla-net: efficient semantic segmentation of large-scale point clouds. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 11108–11117.

Hyyppa, J., Jaakkola, A., Hyyppa, H., Kaartinen, H., Kukko, A., Holopainen, M., Zhu, L., Vastaranta, M., Kaasalainen, S., Krooks, A., et al., 2009. Map updating and change detection using vehicle-based laser scanning. In: 2009 Joint Urban Remote Sensing Event. IEEE, pp. 1–6.

Kaartinen, H., Hyyppä, J., Vastaranta, M., Kukko, A., Jaakkola, A., Yu, X., Pyörälä, J., Liang, X., Liu, J., Wang, Y., et al., 2015. Accuracy of kinematic positioning using global satellite navigation systems under forest canopies. Forests 6 (9), 3218–3236.

Kaasalainen, S., 2019. Multispectral terrestrial LiDAR: state of the art and challenges. Laser Scan. 5–18.

Kaasalainen, S., Malkamäki, T., 2021. Hyperspectral lidar: a progress report. Opt Photon. News 32 (11), 38–43.

Kada, M., McKinley, L., 2009. 3D building reconstruction from LiDAR based on a cell decomposition approach. Int. Arch. Photogram. Rem. Sens. Spatial Inf. Sci. 38 (Part 3), W4.

Kaijaluoto, R., Kukko, A., El Issaoui, A., Hyyppä, J., Kaartinen, H., 2022. Semantic segmentation of point cloud data using raw laser scanner measurements and deep neural networks. ISPRS Open J. Photogrammet. Rem. Sens. 3, 100011.

Khan, M.J., Khan, H.S., Yousaf, A., Khurshid, K., Abbas, A., 2018. Modern trends in hyperspectral image analysis: a review. IEEE Access 6, 14118–14129.

Kong, L., Liu, Y., Chen, R., Ma, Y., Zhu, X., Li, Y., Hou, Y., Qiao, Y., Liu, Z., 2023. Rethinking Range View Representation for Lidar Segmentation arXiv preprint arXiv: 2303.05367.

Kukko, A., Kaartinen, H., Hyyppä, J., Chen, Y., 2012. Multiplatform mobile laser scanning: usability and performance. Sensors 12 (9), 11712–11733.

Kukko, A., Kaartinen, H., Osinski, G., Hyyppä, J., 2020. Modelling permafrost terrain using kinematic, dual-wavelength laser scanning. ISPRS Annal. Photogrammet. Rem. Sens. Spatial Inform. Sci. 5 (2).

Lehtola, V.V., Koeva, M., Elberink, S.O., Raposo, P., Virtanen, J.-P., Vahdatikhaki, F., Borsci, S., 2022. Digital twin of a city: review of technology serving city needs. Int. J. Appl. Earth Obs. Geoinf., 102915

Lehtola, V.V., Lehtomäki, M., Hyyti, H., Kaijaluoto, R., Kukko, A., Kaartinen, H., Hyyppä, J., 2019. Preregistration classification of mobile LiDAR data using spatial correlations. IEEE Trans. Geosci. Rem. Sens. 57 (9), 6900–6915.

Levinson, J., Thrun, S., 2010. Robust vehicle localization in urban environments using probabilistic maps. In: 2010 IEEE International Conference on Robotics and Automation. IEEE, pp. 4372–4378.

Li, D., Shen, X., Guan, H., Yu, Y., Wang, H., Zhang, G., Li, J., Li, D., 2022. AGFP-Net: attentive geometric feature pyramid network for land cover classification using airborne multispectral LiDAR data. Int. J. Appl. Earth Obs. Geoinf. 108, 102723.

Li, Y., Ma, L., Zhong, Z., Liu, F., Chapman, M.A., Cao, D., Li, J., 2020. Deep learning for lidar point clouds in autonomous driving: a review. IEEE Transact. Neural Networks Learn. Syst. 32 (8), 3412–3432.

Liu, Z., Tang, H., Lin, Y., Han, S., 2019. Point-voxel cnn for efficient 3D deep learning. Adv. Neural Inf. Process. Syst. 32.

Lőrincz, S.-B., et al., 2021. Contrastive Learning for LIDAR Point Cloud Segmentation.

Maksimainen, M., Vaaja, M.T., Kurkela, M., Virtanen, J.-P., Julin, A., Jaalama, K., Hyyppä, H., 2020. Nighttime mobile laser scanning and 3D luminance measurement: verifying the outcome of roadside tree pruning with mobile measurement of the road environment. ISPRS Int. J. Geo-Inf. 9 (7), 455.

Malkamäki, T., Kaasalainen, S., Ilinca, J., 2019. Portable hyperspectral lidar utilizing 5 ghz multichannel full waveform digitization. Opt Express 27 (8), A468–A480.

Manninen, P., Hyyti, H., Kyrki, V., Maanpää, J., Taher, J., Hyyppä, J., 2022. Towards high-definition maps: a framework leveraging semantic segmentation to improve NDT map compression and descriptivity. In: 2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). IEEE, pp. 5370–5377.

Matikainen, L., Hyyppä, J., Litkey, P., 2016. Multispectral airborne laser scanning for automated map updating. Int. Arch. Photogram. Rem. Sens. Spatial Inf. Sci. 41, 323–330.

Matikainen, L., Karila, K., Litkey, P., Ahokas, E., Hyyppä, J., 2020. Combining single photon and multispectral airborne laser scanning for land cover classification. ISPRS J. Photogrammetry Remote Sens. 164, 200–216.

Mitschke, I., Wiemann, T., Igelbrink, F., Hertzberg, J., 2022. Hyperspectral 3d point cloud segmentation using randla-net. In: International Conference on Intelligent Autonomous Systems. Springer, pp. 301–312.

Morsdorf, F., Nichol, C., Malthus, T., Woodhouse, I.H., 2009. Assessing forest structural and physiological information content of multi-spectral LiDAR waveforms by radiative transfer modelling. Rem. Sens. Environ. 113 (10), 2152–2163.

Näsi, R., Honkavaara, E., Blomqvist, M., Lyytikäinen-Saarenmaa, P., Hakala, T., Viljanen, N., Kantola, T., Holopainen, M., 2018. Remote sensing of bark beetle damage in urban forests at individual tree level using a novel hyperspectral camera from uav and aircraft. Urban For. Urban Green. 30, 72–83.

Parkison, S.A., Gan, L., Jadidi, M.G., Eustice, R.M., 2018. Semantic iterative closest point through expectation-maximization. In: BMVC, p. 280.

Pengra, B.W., Johnston, C.A., Loveland, T.R., 2007. Mapping an invasive plant, phragmites australis, in coastal wetlands using the eo-1 hyperion hyperspectral sensor. Rem. Sens. Environ. 108 (1), 74–81.

Pignatti, S., Palombo, A., Pascucci, S., Romano, F., Santini, F., Simoniello, T., Umberto, A., Vincenzo, C., Acito, N., Diani, M., et al., 2013. The prisma hyperspectral mission: science activities and opportunities for agriculture and land monitoring. In: 2013 IEEE International Geoscience and Remote Sensing Symposium-IGARSS. IEEE, pp. 4558–4561.

Qi, C.R., Su, H., Mo, K., Guibas, L.J., 2017. Pointnet: deep learning on point sets for 3D classification and segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 652–660.

Schott, J.R., 2007. Remote Sensing: the Image Chain Approach. Oxford University Press on Demand.

Shaw, G.A., Burke, H.K., 2003. Spectral imaging for remote sensing. Linc. Lab. J. 14 (1), 3–28.

Suomalainen, J., Hakala, T., Kaartinen, H., Räikkönen, E., Kaasalainen, S., 2011. Demonstration of a virtual active hyperspectral LiDAR in automated point cloud classification. ISPRS J. Photogrammetry Remote Sens. 66 (5), 637–641.

Taher, J., 2019. Deep Learning for Road Area Semantic Segmentation in Multispectral Lidar Data. Ph.D. thesis, Master Thesis. Aalto University.

Taher, J., Hakala, T., Jaakkola, A., Hyyti, H., Kukko, A., Manninen, P., Maanpää, J., Hyyppä, J., 2022. Feasibility of hyperspectral single photon LiDAR for robust autonomous vehicle perception. Sensors 22 (15), 5759.

Tang, H., Liu, Z., Zhao, S., Lin, Y., Lin, J., Wang, H., Han, S., 2020. Searching efficient 3d architectures with sparse point-voxel convolution. In: European Conference on Computer Vision. Springer, pp. 685–702.

Van Mol, B., Ruddick, K., et al., 2004. The compact high resolution imaging spectrometer (chris): the future of hyperspectral satellite sensors. imagery of oostende coastal and inland waters. In: Proceedings of the Airborne Imaging Spectroscopy Workshop. Brugge.

van Rees, E., 2015. The first multispectral airborne lidar sensor. Geoinformatics 18 (1), 10.
Vierhub-Lorenz, V., Kellner, M., Zipfel, O., Reiterer, A., 2022. A study on the effect of multispectral LiDAR data on automated semantic segmentation of 3d-point clouds. Rem. Sens. 14 (24), 6349.
Viswanathan, P.S., Xue, B., U.S. Patent US20190212447A1, 2020. Scanning 3D Imaging Device with Power Control Using Multiple Wavelengths.
Wallace, A., Nichol, C., Woodhouse, I., 2012. Recovery of forest canopy parameters by inversion of multispectral LiDAR data. Rem. Sens. 4 (2), 509–531.
Wang, Y., Sun, Y., Liu, Z., Sarma, S.E., Bronstein, M.M., Solomon, J.M., 2019. Dynamic graph cnn for learning on point clouds. ACM Trans. Graph. 38 (5), 1–12.
Wehr, A., Hemmleb, M., Maierhofer, C., 2006. Multi-spectral laser scanning for inspection of building surfaces: state of the art and future concepts. In: Proceedings of the 7th International Conference on Virtual Reality. Archaeology and Intelligent Cultural Heritage, pp. 147–154.
Wen, S., Wang, T., Tao, S., 2022. Hybrid cnn-lstm architecture for lidar point clouds semantic segmentation. IEEE Rob. Autom. Lett. 7 (3), 5811–5818.
Xie, X., Bai, L., Huang, X., 2021. Real-time LiDAR point cloud semantic segmentation for autonomous driving. Electronics 11 (1), 11.
Zeid, K.A., Schult, J., Hermans, A., Leibe, B., 2023. Point2vec for Self-Supervised Representation Learning on Point Clouds arXiv preprint arXiv:2303.16570.
Zhang, Y., Zhou, Z., David, P., Yue, X., Xi, Z., Gong, B., Foroosh, H., 2020. 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, pp. 9601–9610.