remote sensing

MDPI

Article

TUM-MLS-2016: An Annotated Mobile LiDAR Dataset of the TUM City Campus for Semantic Point Cloud Interpretation in Urban Areas

Jingwei Zhu [1], Joachim Gehrung [1,2], Rong Huang [1], Björn Borgmann [1,2], Zhenghao Sun [1], Ludwig Hoegner [1], Marcus Hebel [2], Yusheng Xu [1,*] and Uwe Stilla [1]

1 Photogrammetry and Remote Sensing, Technical University of Munich (TUM), 80333 Munich, Germany; jingwei.zhu@tum.de (J.Z.); joachim.gehrung@iosb.fraunhofer.de (J.G.); rong.huang@tum.de (R.H.); bjoern.borgmann@iosb.fraunhofer.de (B.B.); zhenghao.sun@tum.de (Z.S.); ludwig.hoegner@tum.de (L.H.); stilla@tum.de (U.S.)

2 Fraunhofer IOSB, Ettlingen, Fraunhofer Institute of Optronics, System Technologies and Image Exploitation, Gutleuthausstr. 1, 76275 Ettlingen, Germany; marcus.hebel@iosb.fraunhofer.de

* Correspondence: yusheng.xu@tum.de

Received: 30 April 2020; Accepted: 2 June 2020; Published: 9 June 2020

Abstract: In the past decade, a vast amount of strategies, methods, and algorithms have been developed to explore the semantic interpretation of 3D point clouds for extracting desirable information. To assess the performance of the developed algorithms or methods, public standard benchmark datasets should invariably be introduced and used, which serve as an indicator and ruler in the evaluation and comparison. In this work, we introduce and present large-scale Mobile LiDAR point clouds acquired at the city campus of the Technical University of Munich, which have been manually annotated and can be used for the evaluation of related algorithms and methods for semantic point cloud interpretation. We created three datasets from a measurement campaign conducted in April 2016, including a benchmark dataset for semantic labeling, test data for instance segmentation, and test data for annotated single 360° laser scans. These datasets cover an urban area of approximately 1 km long roadways and include more than 40 million annotated points with eight classes of objects labeled. Moreover, experiments were carried out with results from several baseline methods compared and analyzed, revealing the quality of this dataset and its effectiveness when using it for performance evaluation.

Keywords: MLS point clouds; semantic labeling; instance segmentation

1. Introduction

Geospatial data plays a vital role in a wide variety of urban applications like road mapping, field navigation, and building reconstruction [1]. Recently, the spreading uses of 3D point clouds generated from Light Detection and Ranging (LiDAR) systems or multi-view stereo vision provide more diverse options of using geospatial data, with accurate geometric and rich radiometric information [2]. In particular, point clouds measured with the LiDAR system mounted on a mobile platform (MLS) can directly map large-scale urban areas with accurate and detailed 3D measures [3,4]. However, for applying this appealing type of geospatial data in practical jobs, a semantic interpretation of the acquired point clouds is often an obligatory procedure [5]. In this regard, in the past decade, a vast amount of strategies, methods, and algorithms have been developed to explore the semantic interpretation of 3D point clouds for extracting desirable information. To assess the performance of the developed algorithms or methods, public standard benchmark datasets should invariably be introduced and used, which serve as an indicator and ruler in the evaluation and comparison [6].

However, although some remarkable datasets have been proposed and popularized, the creation of publicly accessible large-scale annotated datasets is still in its early life. In this work, we introduce and present large-scale Mobile LiDAR point clouds acquired at the city campus of the Technical University of Munich for the evaluation of related algorithms and methods for semantic point cloud interpretation. The entire point cloud covers an urban area of approximately 0.2 km^2, with around 1 km long roadways. For two major tasks in the semantic interpretation, namely the semantic labeling and object segmentation, we created three datasets from a measurement campaign in April 2016, including a benchmark dataset for semantic labeling, test data for instance segmentation, and test data for annotated single 360° laser scans. The benchmark dataset for semantic labeling contains more than 40 million points with eight classes of objects labeled. We also provide an annotated dataset with instance segmentation of more than 1000 labeled objects in this scene. Moreover, more than 17,000 single 360° scans have been partially annotated, which could be of great use for more challenging investigations like LiDAR-based Semantic SLAM. An example of a small area of the presented benchmark dataset for semantic labeling is shown in Figure 1.

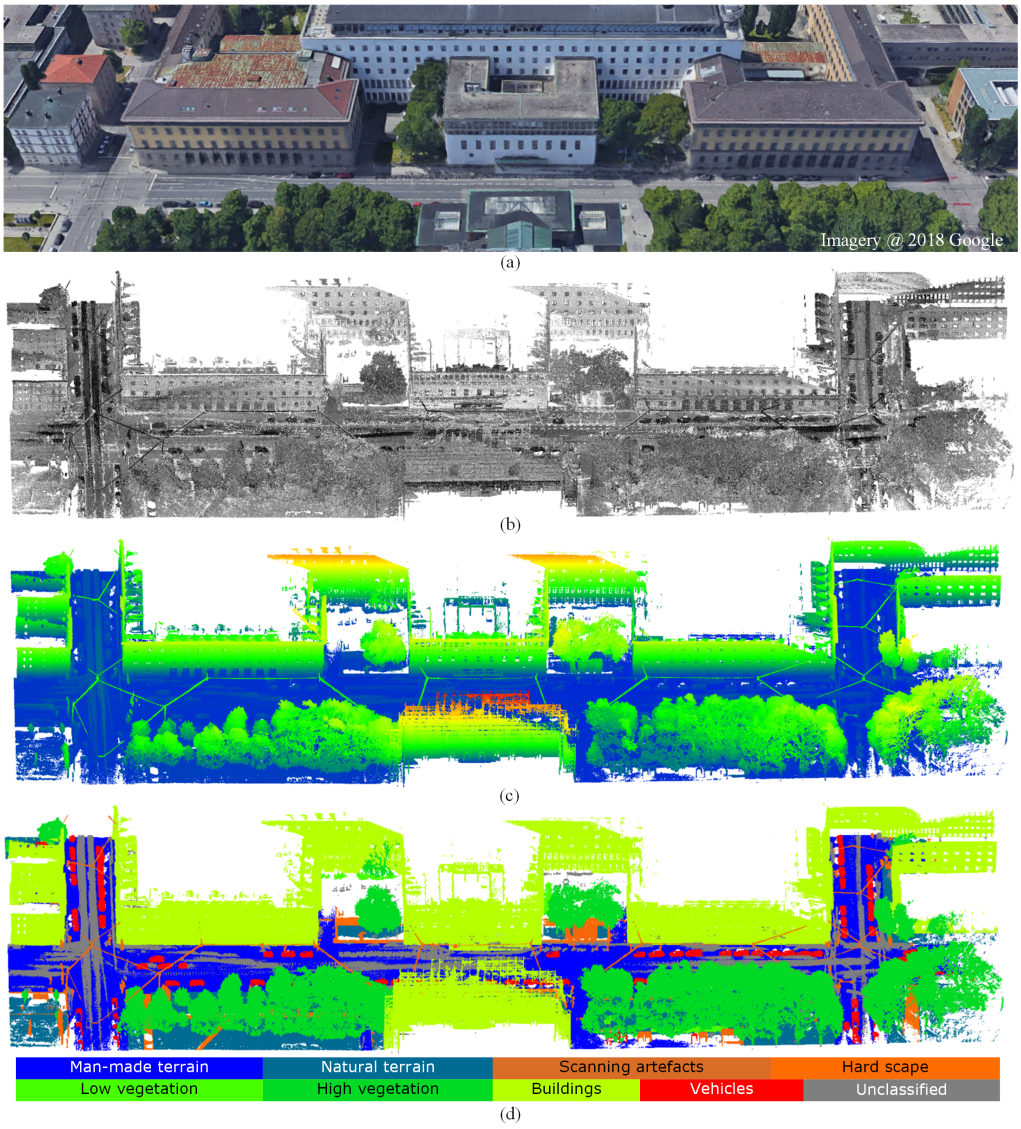

Figure 1. A small example area of the benchmark dataset for semantic labeling. (**a**) Real scene of the TUM main entrance from Google Maps, 2018. (**b**) MLS point cloud with intensities measured by the laser scanners. (**c**) MLS point cloud colored with respect to height. (**d**) Annotated point clouds with eight different semantic labels.

The innovative contributions of this paper are twofold: (1) We introduce and present three large-scale annotated point cloud datasets with point-wise labels and instance cases for semantic interpretation in urban areas. (2) We give an extensive comparison on the performance of semantic labeling methods on the proposed benchmark dataset.

The remainder of this paper is organized as follows: A brief literature view of Mobile LiDAR datasets is given in Section 2. The description of our proposed datasets is provided in Section 3. Subsequently, experimental evaluation using the proposed benchmark dataset and related discussions are given in Section 4. Finally, conclusions about the proposed datasets are drawn in Section 5.

2. Benchmark Datasets from MLS Point Clouds for Semantic Interpretation

With the rapid development of point cloud processing techniques, a wide range of benchmark datasets for various tasks have been presented. With respect to the semantic segmentation and semantic labeling, there are already plentiful benchmark datasets that have been presented, such as the Oakland outdoor MLS dataset [7], the Semantic3D.net TLS dataset (Semantic3D) [8], our own but unannotated TUM-City-Campus MLS (2016) dataset [4], the Paris-Lille-3D MLS dataset [9], the Toronto-3D MLS dataset [10], the Daimler urban segmentation dataset [11], and the A2D2 dataset [12]. However, for any of the benchmark point cloud datasets, there is always a delimitation for the platform used for measuring the 3D points. This means that the attributes, accuracy, density, and quality of different types of point clouds vary significantly due to different platforms used in the measuring [13]. Thus, for evaluating algorithms and methods designed for different applications, the different types of point clouds used for generating benchmark datasets should be considered. Moreover, the costs and difficulties of generating these benchmark datasets are totally different as well. Thus, there are only a few accessible benchmarks from MLS point clouds, and the representative ones include Oakland 3D [7], the Sydney Urban Objects Dataset [14], iQmulus [15], Paris-Lille-3D [9], SemanticKITTI [16], Toronto-3D [10], the Daimler urban segmentation dataset [11], and the A2D2 dataset [12]. The brief introduction of these datasets is as follows:

- The Oakland 3D dataset [7] is one of the earliest publicly accessible MLS datasets for semantic labeling. The dataset was acquired by a side-looking SICK LMS sensor in push-broom way around the campus of the Carnegie Mellon University Oakland, Pittsburgh, PA. This dataset has by default been separated into the training, validation, and test parts, with a total number of about 1.6 million points. All the points in this dataset were assigned with labels of 44 classes of objects, but only 5 classes among them can be used for the evaluation.
- The Sydney Urban Objects dataset [14] is a dataset containing a variety of common urban road objects. This dataset was collected in the CBD of Sydney, Australia, by a Velodyne HDL-64E LiDAR sensor. The entire dataset consists of 631 individual scans. All points were labeled with four classes of objects, including vehicles, pedestrians, traffic signs and trees. As an evaluation dataset, it was designed to test matching and classification algorithms, with a large variability in viewpoint and occlusion.
- The iQmulus dataset [15] is also an early published MLS dataset, which served the IQmulus and TerraMobilita Contest. This dataset was acquired in the 6th district of Paris by the Stereopolis II system with a Riegl LMS-Q120i LiDAR sensor. The entire dataset has collected more than 300 million points. All the points in this dataset were assigned with labels of 22 classes of objects. However, only a 200 m long subset, including 12 million points of 8 classes, is available for the public evaluation purpose.
- Paris-Lille-3D [9] is a recently published MLS dataset for both semantic labeling and instance segmentation. The dataset was acquired in the streets of Paris and Lille by an MLS system with a Velodyne HDL-32E LiDAR sensor. The entire dataset has collected more than 140 million points, covering approximately 2 km roadways. All the points in this dataset were assigned with labels of 50 classes of objects. For the public evaluation purpose, for benchmarks, labels of 9 classes are

provided. Moreover, not only point-wise labels, individual objects like cars and trees are also segmented as instances for evaluation use.

- The SemanticKITTI dataset [16] is one of the newest publicly accessible MLS datasets for semantic segmentation. This dataset was created by annotating the renowned KITTI dataset [17]. This dataset has collected about 4.5 billion points, covering a roadway of 40 km. This dataset is presented by a sequence of scans. The points of each sequential scan were labeled with 25 classes for the evaluation purpose.
- Toronto-3D [10] is a recent MLS dataset for semantic labeling. This dataset was acquired on Avenue Road in Toronto, Canada, via a vehicle-mounted MLS system with a 32-line LiDAR sensor. This dataset has collected approximately 78.3 million points, covering approximately 1 km of roadways. All the points in this dataset were assigned with labels of 7 classes of objects and 1 class of unclassified ones. This dataset has been separated into four parts in default, and each part covers a road length of about 250 m. For the evaluation purpose, theoretically, any part can be used as test data and the rest as training data, or vice versa.
- The Daimler urban segmentation dataset [11] is not an MLS dataset, but can still be considered related (3D): It consists of 5000 rectified stereo image pairs, and 500 frames come with pixel-level semantic class annotations into five classes. Dense disparity maps are provided as a reference computed using semi-global matching.
- Audi's recent A2D2 dataset [12] is provided for research in the context of autonomous driving. This dataset was acquired in three cities in the south of Germany, namely: Gaimersheim, Ingolstadt, and Munich. In total, six cameras and five Velodyne VLP-16 LiDAR sensors were used. 41,277 images have semantic and instance segmentation labels for 38 categories. The annotation of the point clouds is generated by projecting the points to the 38,481 semantically labeled images with calibrated relative position and orientation of the sensors.

Even though all the above-mentioned datasets contain semantically labeled 3D data, our MLS dataset differs from them in several aspects. Typically, related work focuses on either real-time computer vision tasks, such as autonomous driving, or on mobile mapping, e.g., for the generation of city models or other tasks related to geoinformatics. Typical representatives of the first type are the datasets provided by car companies. For example, Audi's A2D2 dataset and Daimler's urban segmentation dataset are designed for developments in view of autonomous driving and focus on traffic participants at the street level. A typical representative of the second type is the iQmulus dataset, where the focus is on high density data acquisition and offline scene analysis of large urban areas. Real-time aspects like the perception of current events during data acquisition are completely ignored in this case.

With our TUM-MLS-2016 dataset, we want to bridge the gap between different communities, e.g., computer vision, robotics, and geoinformatics. Our dataset covers the time course of the measuring run and the street scene with real-time events, but also a consistent and area-wide representation of the surveyed urban area including high-rise facades of buildings. A representation of a large urban area by a 3D point cloud and its semantic interpretation could leverage research on real-time applications like self-localization and traffic monitoring, for which we provide an all-in-one dataset. In addition, we provide different kinds of labels, including semantic labels and single instances of relevant objects. Compared with image-based 2D labels transferred to 3D points by projection, annotations of our point clouds have been made directly in 3D, which has been a labor-intensive operation but can be considered more reliable.

A special sensor configuration with two obliquely rotating laser scanners was chosen to provide a data basis as universal as possible under the two aspects mentioned above, real-time applications and mobile mapping. Although this configuration is special, the sensor data can represent or simulate that of state-of-the-art mapping sensor systems as well as sensors discussed and designed for autonomous driving (e.g., forward looking solid-state LiDAR sensors, flash LiDAR cameras with overlapping fields-of-view). We have put a lot of effort into correct georeferencing of the 3D data using an inertial

navigation system including RTK-GNSS. On the one hand, this is necessary to be able to use the dataset in the context of geoinformatics. On the other hand, and with the existence of loops in the trajectories, it can be a perfect testbed for sophisticated investigations like LiDAR-based Semantic SLAM.

To have a better impression of current benchmark datasets of MLS point clouds, we give a comparison of comprehensive indicators of the above-mentioned datasets in Table 1. As seen from Table 1, we can also find several remarkable limitations in current MLS benchmark datasets. Thus, different algorithms and methods usually suffer inconsistent performance on different datasets. Relatively, the algorithms or methods designed for certain tasks should be assessed by corresponding benchmark datasets. Otherwise, the evaluation would be biased.

Table 1. Representative MLS point cloud datasets for semantic interpretation.

Dataset	Year	Size (km)	# Points	# Classes	Sensor
Oakland 3D [7]	2009	1.5	1.6 M	44	Sick LMS
Sydney Urban Objects [14]	2013	-		4	Velodyne HDL-64E
iQmulus [15]	2015	0.20	12 M	22	Riegl LMS-Q120i
Paris-Lille-3D [9]	2018	1.94	143.1 M	50	Velodyne HDL-32E
SemanticKITTI [16]	2019	39.2	4.5 B	28	Velodyne HDL-64E
Toronto-3D [10]	2020	1.00	78.3 M	8	Velodyne HDL-32
Daimler urban segmentation [11]	2013	-		5	Stereo optical sensor
A2D2 [12]	2020	-		38	Five Velodyne VLP-16
TUM-City-Campus MLS	2016	0.97	41 M/1.7 B	8	Dual Velodyne HDL-64E

M: Million, B: Billion.

3. TUM-City-Campus MLS Dataset

Based on the analysis of the problems existing in the current benchmark datasets of MLS point clouds, we present our large-scale Mobile LiDAR datasets termed as TUM-City-Campus MLS (2016), which are designed for semantic interpretation of MLS point clouds in urban areas. Video clips illustrating our dataset and further information are available on the website (https://www.pf.bgu.tum.de/en/pub/tst.html) with supplementary material [18].

3.1. Data Acquisition, Preparation, and Annotation

The MLS data have been acquired in April 2016 by Fraunhofer IOSB with their MODISSA mobile sensor platform. At Fraunhofer IOSB, the experimental multi-sensor vehicle MODISSA (Mobile Distributed Situation Awareness) is used for hardware evaluation and software development in the contexts of automotive safety and security applications. At the time of the data acquisition in 2016, MODISSA was equipped with two Velodyne HDL-64E LiDAR sensors above the windshield, where each Velodyne HDL-64E was configured to have a rotational frequency of 10 Hz and acquired 130,000 range measurements (3D points) per rotation with distances up to 120 m. Each sensor consists of 64 laser rangefinders, which divide the vertical field of view of 26.8° into 64 scan lines. Both laser scanners were positioned on wedges at a 25° angle to the horizontal, rotated outwards at a 45° angle (see Figure 2a). Reasons to use two obliquely rotating laser scanners have already been given in Section 2. There are additional positive features of this configuration. To a large extend, it prevents measurements of the vehicle's roof and still guarantees a good coverage of the roadway in front and to the sides of the vehicle. At the same time, the facades of buildings are captured in their entire height, which is useful for mobile mapping purposes (see Figure 2b). With the data of both sensors being synchronized in time, the overlap between the two sensors can simulate the overlapping fields-of-view of directed LiDAR cameras to be used for autonomous driving. In addition, this configuration increases the overall point density. The LiDAR data were recorded synchronously with position and orientation data of an Applanix POS LV 520 inertial navigation system (INS), which was augmented by RTK correction data of the German SAPOS network. All lever arms and boresight directions of the system components had been thoroughly calibrated beforehand [19], such that it was possible to

perform direct georeferencing of the LiDAR data and aggregate all resulting 3D points in a common local ENU coordinate frame. Although the data acquisition is continuous, for convenience and by convention we split the stream of georeferenced 3D points to a sequence of scans of 1/10 second duration, corresponding to single 360° scans of the scanner heads rotating at 10 Hz. A comprehensive description of the sensor system can be found in [20].

The data acquisition took place in the area of the city campus of Technical University of Munich (TUM) in Munich, Germany. Figure 2b illustrates the data acquisition and shows the footprint of each laser scanner in different color [4].

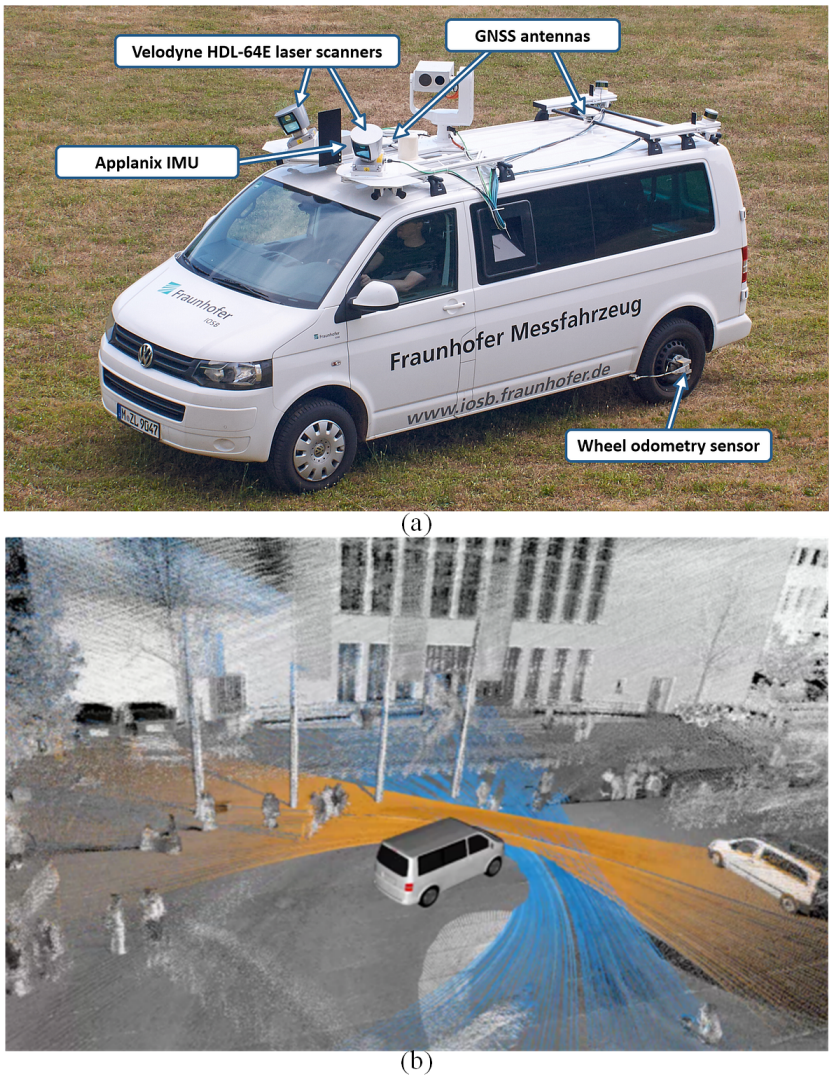

Figure 2. MLS system and its components. (**a**) MODISSA platform from Fraunhofer IOSB [20]. (**b**) Illustration of data acquisition with the two obliquely mounted laser scanners [4].

MLS data in more than 17 thousand 360° scans have been acquired by each of the two laser scanners and directly georeferenced while driving along the roads around the TUM city campus and the inner yard. This covers an urban scenario consisting of building facades, trees, bushes, parked vehicles, wedges, roads, grass and so on. Each point has 3D x-, y-, and z-coordinates and intensities of the laser reflectance. In Figure 3, we illustrate the aerial image of the measured areas and the acquired and georeferenced MLS point clouds.

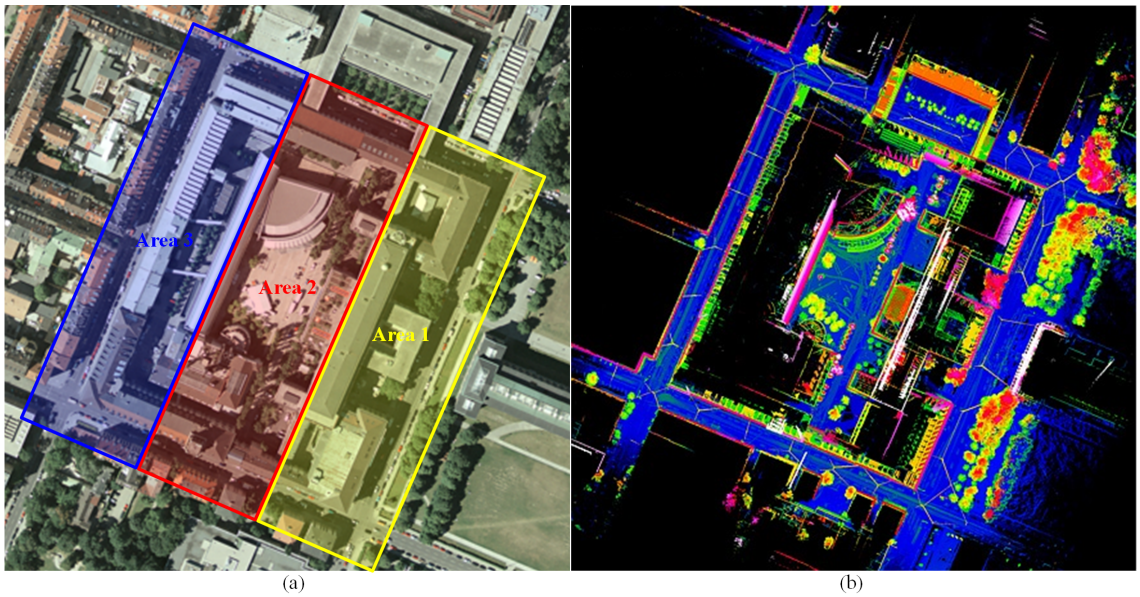

Figure 3. Measured MLS point clouds. (**a**) Aerial image of the measured TUM city campus. (**b**) Aggregated MLS point clouds colored with respect to height.

In the annotation, all the measured points in the scene were manually labeled with eight semantic classes following the ETH standard (Semantic3D.net benchmark) [21] and one *unclassified* class. The point-wise labels were assigned manually using CloudCompare 2.10 (https://www.danielgm.net/cc/). In Figure 1d, an illustration of points with these classes is given, with points of various labels rendered with different colors. To be specific, the details of these eight different classes are given in Table 2.

Table 2. Annotated classes of objects.

Classes	Label Index	Color Code	Content
Man-made terrain	1	#0000FF	Roads and impervious ground.
Natural terrain	2	#006B93	Grass and bare land.
High vegetation	3	#00DA24	Trees.
Low vegetation	4	#47FF00	Bushes and flower beds.
Buildings	5	#B6FF00	Building facades and roofs.
Hardscape	6	#FF6C00	Walls, fences, light poles.
Scanning artifacts	7	#D96D25	Power cables and artificial objects.
Vehicles	8	#FF0000	Parked cars and buses.
Unclassified	0	#7F7F7F	Noise, outliers, moving vehicles, pedestrians, and unidentified objects.

Based on the annotation of points, we created three datasets serving the evaluation of related methods and algorithms, including a benchmark dataset for semantic labeling, test data for instance segmentation, and test data for individual labeled 360° scans. These three datasets are related to the two core tasks of semantic interpretation, namely the semantic labeling and object segmentation. In Figure 4, we give an illustration of the creation of these three datasets with involved processing steps and a workflow.

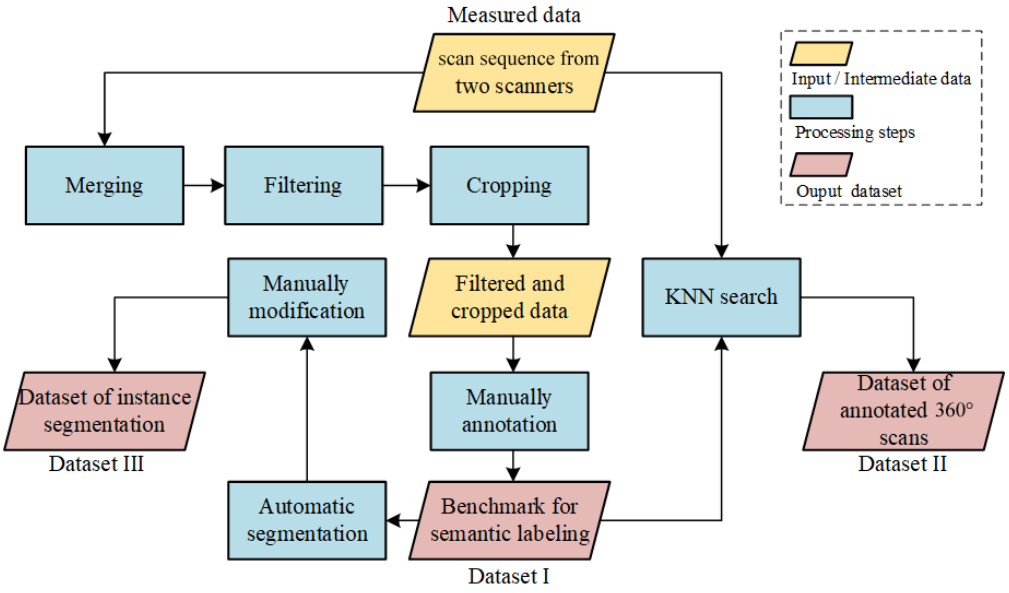

Figure 4. Workflow of the creation of three labeled datasets from the original MLS scans.

3.2. Benchmark for Semantic Labeling

As shown in the workflow of Figure 4, for generating the benchmark dataset for semantic labeling, we started with re-merging all points of all the georeferenced single scans into a large point cloud. Then, the merged point cloud was preprocessed by the statistical outlier removal (SOR) filter and downsampled, with duplicated points deleted. These duplicated points were mainly caused by repetitive scans when the vehicle was waiting for traffic lights. After these two steps, the number of points has been greatly reduced. Sequentially, the distant points in the scan with a sparse density were cropped and removed. Based on the filtered and cropped point cloud, we conducted the annotation of points manually according to the standard stated in the previous subsection. The total number of annotated points is more than 40 million. With these annotated points, we created a benchmark dataset for the evaluation of semantic labeling. Here, only annotated points of the eight semantic classes are kept, and those which belong to the *unclassified* class are removed. In Figure 5, we show the entire annotated benchmark dataset with eight object classes. For evaluation purposes, the entire labeled dataset of the TUM city campus has been evenly divided into three areas according to the covered area size, and the numbers of points in these three areas are around 20 million, 16 million, and 13 million, respectively. In Figure 3a, the separation of these three areas is displayed. Points of each area are saved in the same ply files.

In Figure 6, a statistic overview of labeled points in different areas is illustrated. As seen from the statistics, we can find that for all three areas, the percentages of points from different objects show different distributions and occur in an imbalanced way. The various distributions of points from different objects reveal different scenarios in these three areas. The scenario in Area 1 is a street scene hybridizing man-made objects and vegetation with a relatively balanced distribution. However, the numbers of points from hardscape and scanning artifacts are still considerably less than those of buildings and ground. The scenario in Area 2 is the inner yard of the campus, with all types of objects but very few points of vehicles. Actually, there were only one or two parked cars inside the yard. The scenario of Area 3 is a street scene with closely packed facades and almost no vegetation. Plenty of parked cars were also scanned in this scene. As a summary, we can comment that building points show a dominating tendency in all three areas. The ground points of man-made ground occupy the second-largest percentage for all areas. This would be beneficial for tasks like building reconstruction and vehicle detection. However, the points of artifacts and hardscape are less than 3% among all the annotated points in no matter which area, which makes the recognition of such objects a

challenge. For the supervised method, it is recommended to use the points of Area 1 as the training data, and points of Areas 2 and 3 as the test data. The imbalanced distributions of percentages of labeled points in each area should be considered when making the evaluation.

Figure 5. Benchmark dataset for semantic labeling with eight classes of objects.

3.3. Annotated Data for Instance Segmentation

Based on the annotated benchmark for semantic labeling, we also conducted an instance segmentation to the labeled points, so that points of the same instance can be separated and assigned with a unique label. For example, all points of a parked car are labeled as an individual object and rendered with a unique color. The instance segmentation is comparable to the labeled instances in the Paris-Lille-3D dataset. The creation of this dataset was achieved by two steps. The first step is the automatic segmentation using an unsupervised clustering [22]. Then, in the second step, a manual modification was carried out to correct errors and refine the boundaries of objects. In Figure 7, the instance segmentation of the labeled points is displayed. In total, there are 1002 objects of eight classes mentioned above labeled and segmented. The numbers of objects from different classes in the dataset for instance segmentation are given in Figure 8. As seen from the figure, we can find that there are a vast number of trees and vehicles that have been segmented and annotated. This could be a useful aspect for related tasks.

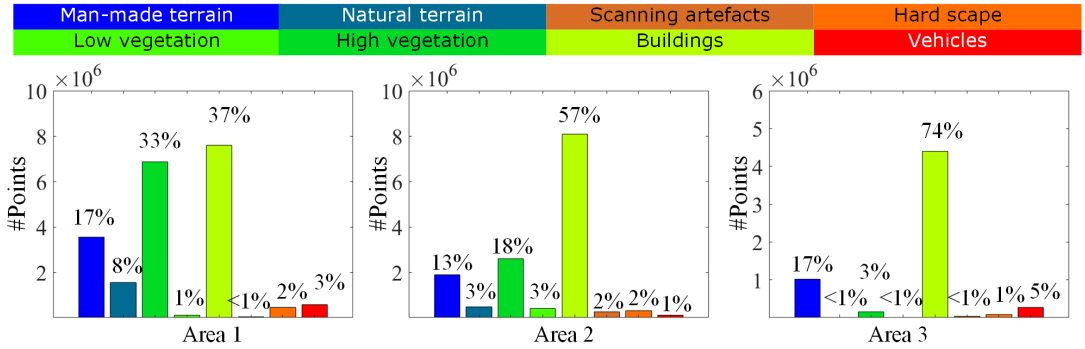

Figure 6. Percentage of labeled points in the three different areas.

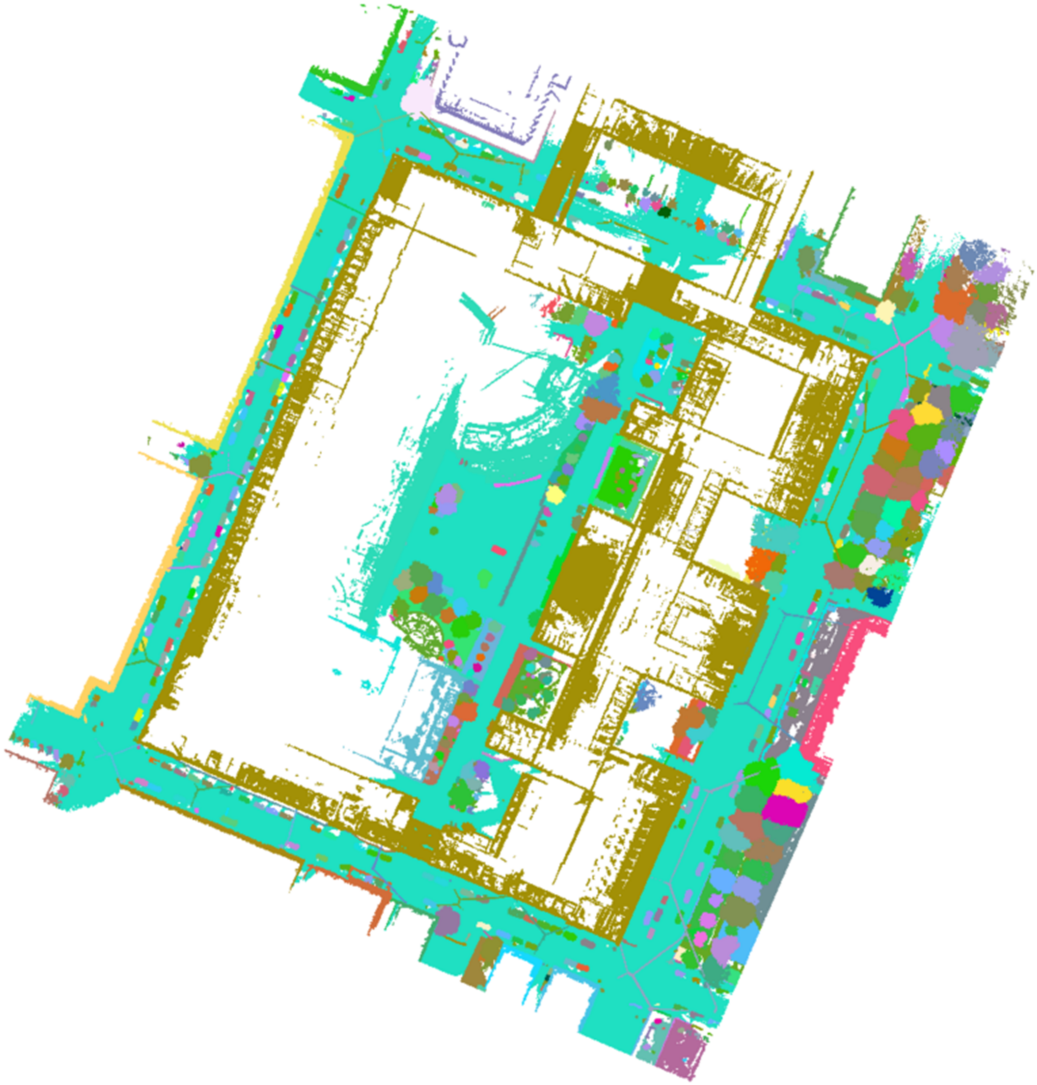

Figure 7. The instance segmentation of the labeled points. Points of the same objects are dyed into unique colors.

3.4. Annotated Data for Single 360° Laser Scans

For the single 360° laser scans from both scanners, we conducted a nearest neighbor search for assigning the points sequentially with a possible label, according to the annotated points in the large-scale benchmark dataset. In Figure 9, the trajectory of the MLS vehicle is illustrated by red dots (representing equal time steps). As can be seen from the trajectory, the scanning data

include two loops covering the inner yard and outer area of the campus, which can be used for the evaluation of SLAM methods like LiDAR-based Semantic SLAM. To do that, the georeferenced 360° scans can be transformed back to the sensor's coordinate frame, i.e., the motion compensation and direct georeferencing achieved by the INS can be undone, and the task for the SLAM method under consideration would be to replace the INS. The labeled sequence of scans is comparable to the labeled sequences of scans in the SemanticKITTI dataset. For the points in each scan, a point was given the same label as the nearest neighbor in the annotated scene within a given threshold (0.3 m). If there are no points found in the given radius, the point was labeled to belong to the *unlabeled* class. In Figure 10, annotated points of single 360° laser scans are displayed. Compared to the annotated points in the large benchmark dataset, the annotated single laser scans contain both the geometric characteristics of the scanners and the classification labels of the points.

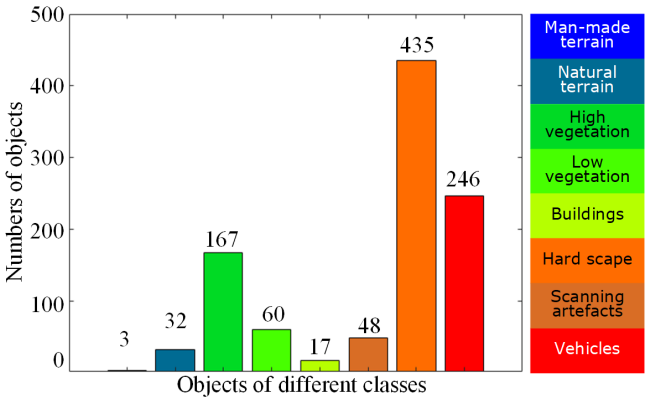

Figure 8. Numbers of objects from different classes in the dataset for instance segmentation.

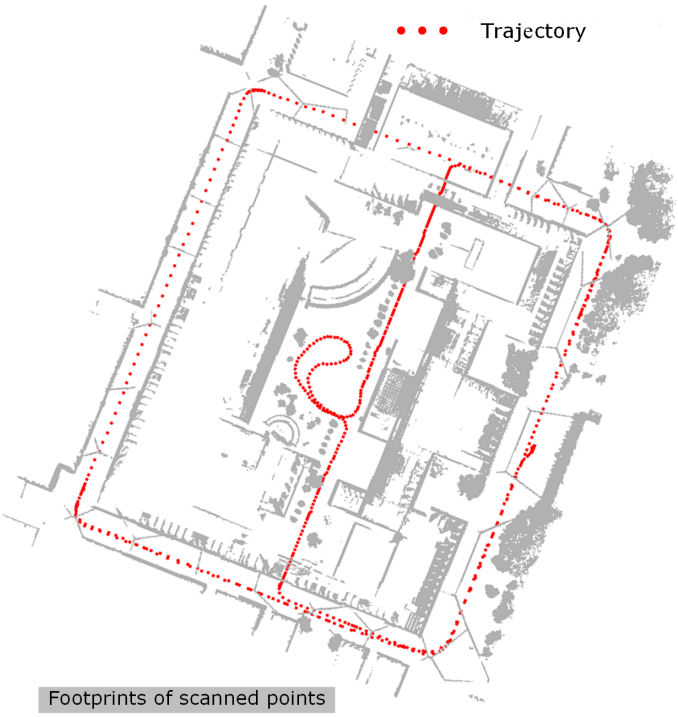

Figure 9. Vehicle trajectory. The red points correspond to equal time steps during the data acquisition, and the gray background indicates the footprint of scanned points.

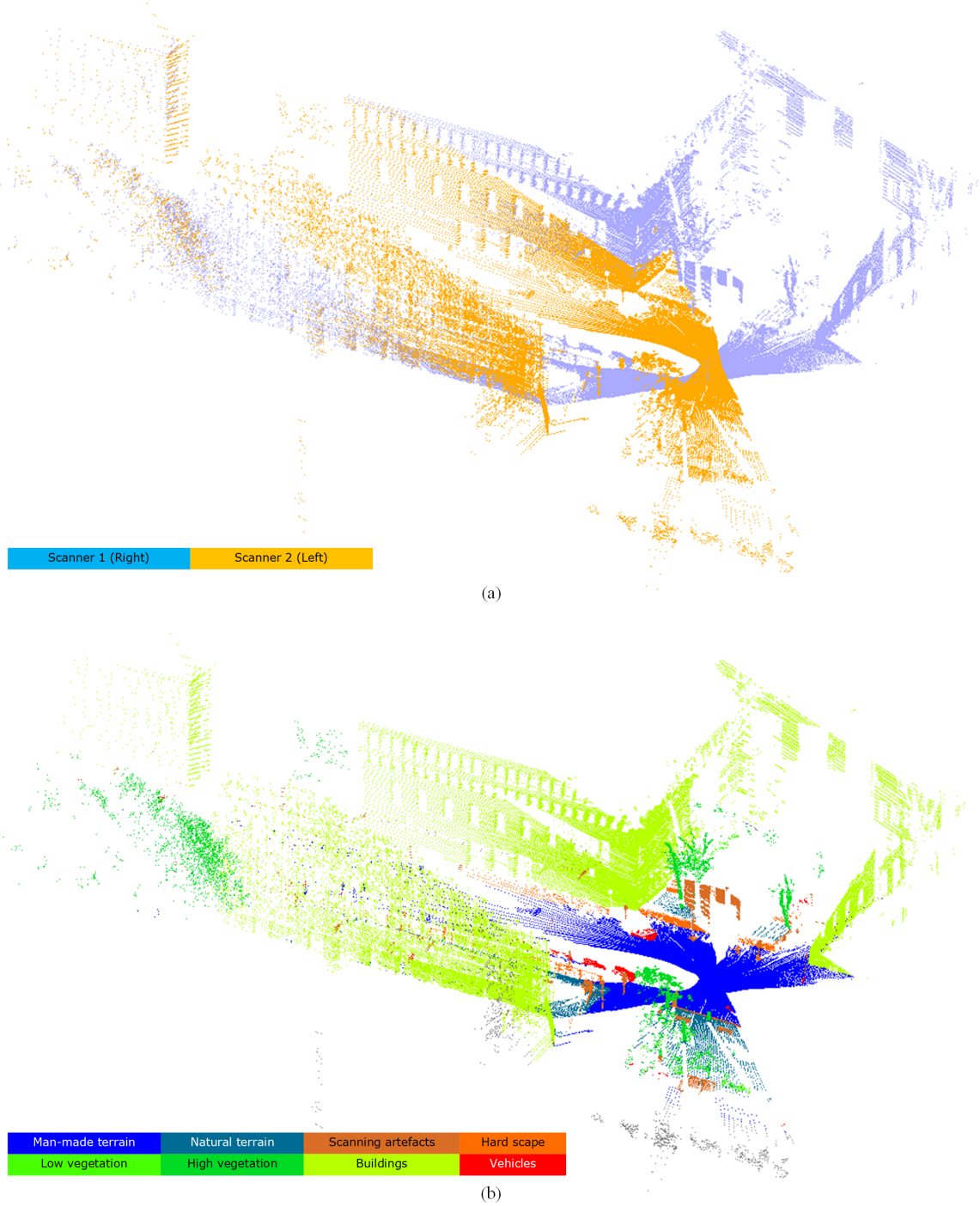

Figure 10. The annotated single 360° laser scans. (**a**) Points of two 360° scans from different scanners (i.e., left and right ones) at the same time of data acquisition. (**b**) Annotated points of two 360° scans at the same time of data acquisition.

In Figure 11, we also provide an illustration of the annotation result of a sequence of single scans with time index of 04068, 04108, 04148, respectively. These three 360° scans are not continuous but have a separation of 4 s.

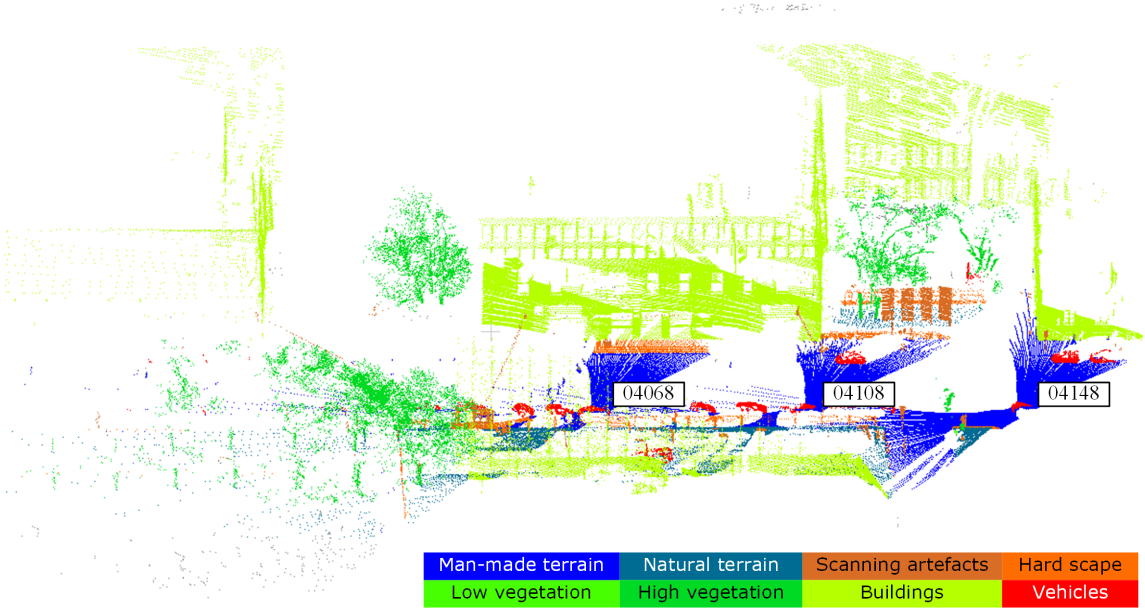

Figure 11. Three selected single 360° scans from scanner 1 with a separation of 4 s.

4. Evaluation

To give a brief evaluation of the dataset, we applied supervised classification to this dataset for labeling the points with five aforementioned classes of objects in the scene. In the experiments, we use points of Area 1 as the training data, while we use points of Areas 2 and 3 as the test data.

4.1. Baselines of Semantic Labeling

In the experiments, four point-based semantic labeling methods were tested on the proposed dataset as baselines, including:

- PointNet [23]: PointNet is a neural network that directly processes point clouds, which well respects the permutation invariance of points in the input. It provides a unified architecture for applications emerging from object classification.
- PointNet++ [24]: PointNet learns global features with MLPs for raw point clouds. PointNet++ applies PointNet to local neighborhoods of each point to capture local features, and a hierarchical approach is taken to capture both local and global features.
- Detrended geometric features and graph-based optimization (DEGO) [25]: This is a conventional handcrafted feature-based method using eigenvalue-based geometric features [5] with a detrended feature enhancement strategy and a random forest classifier. A post-processing with graph-structured optimization is applied for the refinement of initial labels.
- Hierarchical deep feature learning (HDL) [13]: This is a deep feature learning method based on the original PointNet++ [24], in which hierarchical data augmentation is used to create multi-scale pointsets as input. Pointsets subdivided with various scales will contain different levels of contextual information and be concatenated to a multi-scale deep feature vector, which is then classified by the random forest. The joint manifold-based embedding (JME) and global graph-based optimization (GGO) used in [13] are not included here.

4.2. Evaluation Metrics

For the evaluation of the classification results, we follow the Pascal VOC challenges [26] and use Intersection over Union (IoU) averaged over all classes. The evaluation measure for class i is defined as

$$IoU_i = \frac{TP_i}{TP_i + FP_i + FN_i} \quad (1)$$

The main evaluation measure is the $\overline{IoU}$, which is the averaged summation of IoU_i for each class i. Moreover, the overall accuracy is also calculated.

$$\overline{IoU} = \frac{1}{N}\sum_{i=1}^{N} IoU_i \tag{2}$$

Here, for the labeled result of each class, TP denotes the True positive, which is the number of points correctly labeled as this class, namely the points with correct labels. FP stands for the False positive, which means the number of points with incorrect labels. FN is the False negative, which is the number of points which should be labeled as other classes but incorrectly labeled as this class. Moreover, the precision (*Pre.*) and recall (*Rec.*) values are also given for assessing the performance,

$$Precision = \frac{TP}{TP + FP} \quad Recall = \frac{TP}{TP + FN} \tag{3}$$

and finally the overall accuracy (OA) is calculated as well.

$$OA = \sum_{i=1}^{N}\left(\frac{TP_i}{TP_i + TN_i + FP_i + FN_i}\right) \tag{4}$$

5. Experiments And Results

5.1. Training Settings

In this semantic labeling experiment, for the DEGO method, the feature extraction and initial classification have been implemented via C++, and these implementations ran on an Intel i7-6700 CPU @ 3.4GHz and with 32.0 GB RAM. The graph-based optimization is achieved via MATLAB 2018b on the same hardware. To set the key parameters, the size of the voxels (their edge length) is 0.3 m, while the seed resolution of supervoxels is 1.0 m. For the weight factors in the boundary refined process, w_n is set to one, while w_d is set to the reciprocal value of the size of voxels. The number of trees used in our RF classifier is 200. The threshold for the graph cut is set to 0.5. As input to the PointNet and PointNet++ methods, the entire point cloud is subdivided into thousands of sub-point chips, in which 10,000 points are contained. These chips are downsampled to 8192 points which represent the main structure of each chip, and the downsampled chips serve as the input for PointNet and PointNet++. Each point in the chip is represented by a 3D vector, containing the coordinates (x, y, z). Similarly, for the HDL method, we generated sub-point chips with different scales. The training of networks with these chips of points with different scales is carried out individually as well. Thus, we can acquire encoded features encapsulating different levels of contextual information from points. Considering the real scale of objects (e.g., buildings, trees, low vegetation), the sizes of the chips of points are empirically set to 10,000, 20,000, and 30,000 for different scales, respectively, since they showed satisfying performance in the experiments. For the training process of all the above three deep learning-based methods, each training batch contained 16 chips in total. The stochastic gradient descent algorithm with a learning rate η = 0.001 and a momentum value of p = 0.9 was used. For adjusting the learning rate, we decayed its value by the factor of 0.7 in every 40 training chips. The training process lasts for a total of 500 epochs. We monitored the progress of the validation loss and saved the weights if the loss improves. These three methods were implemented via Tensorflow and carried out on an NVIDIA TITAN X (Pascal) 12GB GPU.

5.2. Results And Discussion

In Table 3, we illustrate the comparison of the classification results. The results of PointNet and PointNet++, and DEGO methods have been reported in our previous work [25]. As seen from the results, the DEGO method has advantages over these three baseline methods, when checking the

overall accuracies. In particular, when compared with PointNet and PointNet++, the DEGO method outperforms them significantly in labeling buildings, man-made terrain, and high vegetation. However, we should be aware of the fact that DEGO includes graph-based optimization as post-processing. As we analyzed in [25], a possible reason is that these three kinds of objects have generally isotropic geometric characteristics, facilitating the graph-based optimization process considering the local contextual information. However, the deep learning-based methods also have a strong advantage for classifying points of vehicles and low vegetation. Rather than handcrafted features, the ones from deep learning can supervise and adaptively generate features for such irregular shaped objects, which better fits the reality. For PointNet and PointNet++ methods, the reason buildings and man-made terrain cannot be classified with satisfying results is due to the scale factor. In other words, PointNet and PointNet++ methods are originally designed for computer vision applications and mainly for indoor scenarios. However, when it comes to outdoor scenarios with large changes of object scales, the sampling of points for the input of the network should be further considered. In the HDL method, a hierarchical strategy for data augmentation is used for preparing the input, in which multi-scale pointsets are created. Consequently, the result of the HDL method reveals comparable performance like that of DEGO, even without any post-processing as refinement. Compared with PointNet++, HDL can get an improvement of around 7%. Since the HDL method is just an improved one based on the PointNet++, we can confirm that the modification of input scales can significantly enhance the power of networks for outdoor scenarios. In some recent publications like [27], the hierarchical structure has been used and achieved excellent performance.

Table 3. Comparision of semantic labeling results using the annotated TUM-City-Campus MLS dataset with different methods.

Method	**PointNet [23]**			**PointNet++ [24]**			DEGO [25]			**HDL [13]**		
Class	***Pre.***	***Rec.***	***IoU***	***Pre.***	***Rec.***	***IoU***	***Pre.***	***Rec.***	***IoU***	***Pre.***	***Rec.***	***IoU***
Buildings	0.871	0.915	0.764	0.922	0.910	0.792	0.937	0.970	0.911	0.891	0.872	0.788
Vehicles	0.464	0.569	0.761	0.708	0.926	0.788	0.697	0.544	0.440	0.263	0.374	0.183
Man-made terrain	0.878	0.591	0.546	0.923	0.459	0.442	0.896	0.709	0.655	0.753	0.733	0.591
Natural terrain	0.146	0.406	0.459	0.149	0.651	0.336	0.169	0.514	0.146	0.525	0.174	0.150
High vegetation	0.643	0.570	0.446	0.725	0.695	0.435	0.768	0.908	0.713	0.919	0.922	0.853
Low vegetation	0.155	0.015	0.443	0.458	0.151	0.432	0.000	0.000	0.000	0.370	0.195	0.146
Hardscape	0.035	0.020	0.764	0.364	0.163	0.792	0.833	0.011	0.011	0.296	0.443	0.215
Scanning artifacts	0.144	0.219	0.761	0.169	0.384	0.788	0.945	0.081	0.081	0.715	0.943	0.685
MEAN	0.417	0.413	0.618	0.552	0.542	0.601	0.656	0.467	0.370	0.592	0.582	0.452
OA		0.758			0.773			0.866			0.842	

By applying the trained models to the entire dataset, we can get a complete visualization of the results of every single method, cf. Figure 12. As seen from this figure, the visualization results can also support the quantitative evaluation: although all four methods can assign correct labels for the majority of points, the DEGO method, benefiting from the post-processing optimization, performs better when discriminating buildings and high vegetation. For the visualized results from the HDL method, we can observe that not only buildings and high vegetation but also scanning artifacts and man-made terrain can achieve good results even without post-refinement. From the aspect of the dataset, all these above-achieved results and feedback of various methods have supported the feasibility and effectiveness of the proposed dataset when using it for the performance evaluation.

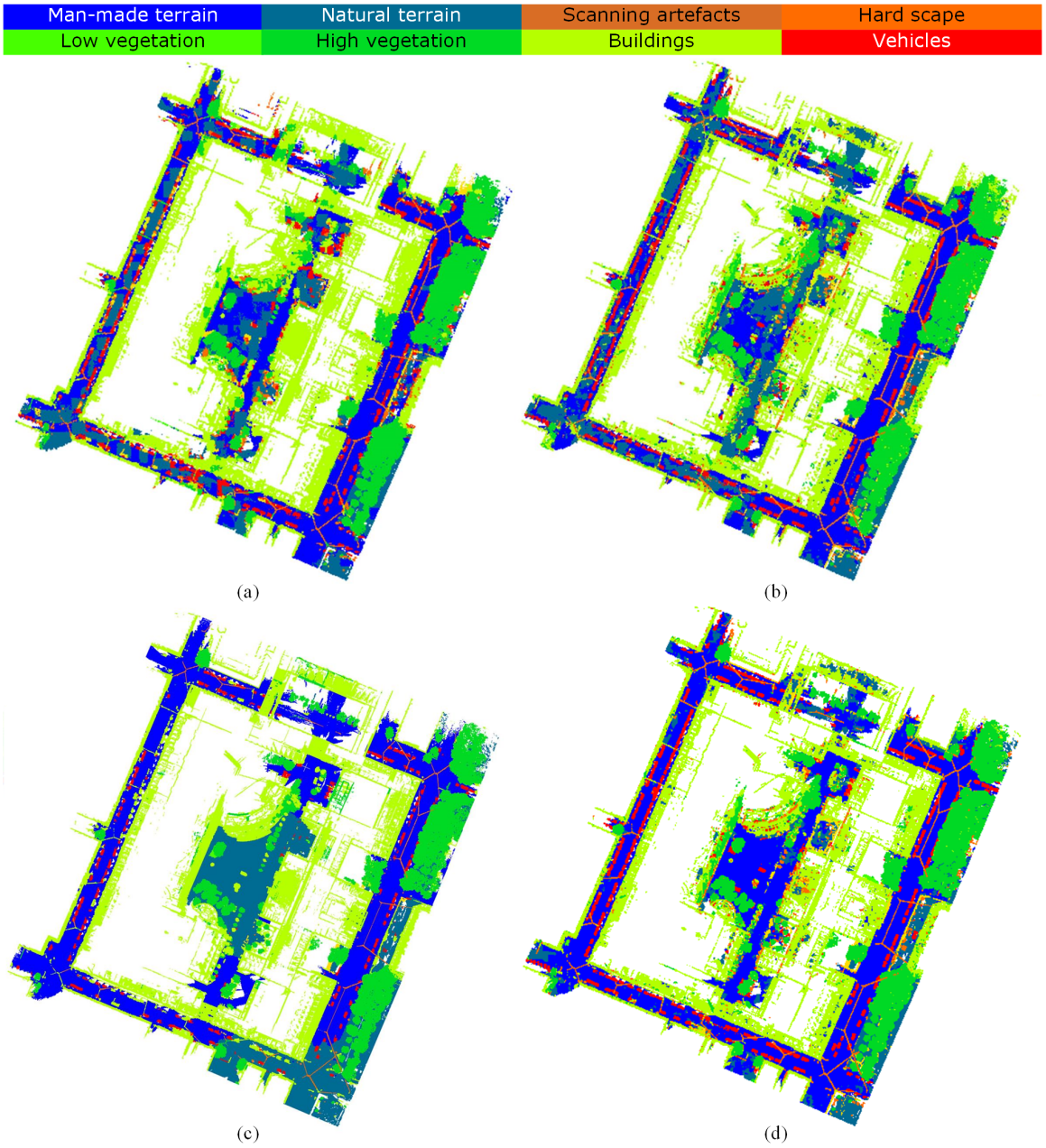

Figure 12. Comparison of classification results using different methods. (**a**) Classification result using PointNet [23], (**b**) classification result using PointNet++ [24], (**c**) classification result using DEGO [25], and (**d**) classification result using HDL [13].

6. Conclusions

In this work, we presented large-scale annotated MLS dataset using Mobile LiDAR point clouds acquired at the city campus of the Technical University of Munich for semantic interpretation evaluation. We presented three datasets including a benchmark dataset for semantic labeling, test data for instance segmentation, and test data for annotated single 360° laser scans. The benchmark dataset for semantic labeling covers an urban area of approximately 1 *km* long roadways, and it includes more than 40 million annotated points with eight classes of objects labeled. The dataset for instance segmentation provides annotated and segmented points of more than 1000 objects. The dataset for labeled single scans provides labeled points from more than 17,000 sequential 360° scans (rotations of the LiDAR scanner's head). Compared with other representative MLS benchmark datasets, the proposed MLS dataset provides not only a benchmark for point-wise semantic labeling

but also annotated instances of individual objects, as well as labeled points in a sequence of 360° scans. Moreover, we also reported evaluation experiments using the benchmark dataset with several reference methods. As a conclusion, we can have the following two remarks:

- The creation of benchmark datasets is essential to assess the performance of developed algorithms and methods. The analysis of our proposed large-scale annotated dataset has revealed its good potential for the evaluation of semantic interpretation in complex urban scenarios.
- Experiments have validated the feasibility and quality of the dataset for semantic labeling. The comparison with methods of different strategies also reveals the importance of considering the scale factors in deep learning-based feature descriptions.

In the future, the quality of these datasets could be further improved and updated according to the feedback from comprehensive experiments and evaluations.

Supplementary Materials: More information about our dataset can be found at: https://www.pf.lrg.tum.de/en/pub/tst.html.

Author Contributions: All authors contributed to this manuscript: Conceptualization, M.H., Y.X., and U.S.; experiment and analysis, R.H. and Y.X.; data acquisition and preprocessing, J.G., B.B., L.H. and M.H., data annotation, J.Z. and Z.S.; writing—original draft preparation, J.Z., M.H., and Y.X.; writing—review and editing, M.H. and U.S.; supervision and funding acquisition, M.H. and U.S. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the German Research Foundation (DFG) and the Technical University of Munich within the funding program Open Access Publishing.

Acknowledgments: This work was carried out within the frame of Leonhard Obermeyer Center (LOC) at Technische Universität München (TUM) [www.loc.tum.de], and the Fraunhofer Cluster of Excellence "Cognitive Internet Technologies".

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Weinmann, M.; Schmidt, A.; Mallet, C.; Hinz, S.; Rottensteiner, F.; Jutzi, B. Contextual classification of point cloud data by exploiting individual 3D neigbourhoods. *ISPRS Ann. Photogramm. Remote Sens. Spat. Inf. Sci.* **2015**, *2*, 271–278. [CrossRef]
2. Yang, B.; Dong, Z.; Zhao, G.; Dai, W. Hierarchical extraction of urban objects from mobile laser scanning data. *ISPRS J. Photogramm. Remote Sens.* **2015**, *99*, 45–57. [CrossRef]
3. Yu, Y.; Li, J.; Guan, H.; Jia, F.; Wang, C. Learning hierarchical features for automated extraction of road markings from 3-D mobile LiDAR point clouds. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2015**, *8*, 709–726. [CrossRef]
4. Gehrung, J.; Hebel, M.; Arens, M.; Stilla, U. An approach to extract moving objects from MLS data using a volumetric background representation. *ISPRS Ann. Photogramm. Remote Sens. Spat. Inf. Sci.* **2017**, *4*, 107. [CrossRef]
5. Weinmann, M.; Jutzi, B.; Hinz, S.; Mallet, C. Semantic point cloud interpretation based on optimal neighborhoods, relevant features and efficient classifiers. *ISPRS J. Photogramm. Remote Sens.* **2015**, *105*, 286–304. [CrossRef]
6. Xie, Y.; Tian, J.; Zhu, X. Linking Points With Labels in 3D: A Review of Point Cloud Semantic Segmentation. *IEEE Geosci. Remote Sens. Mag.* **2020**. [CrossRef]
7. Munoz, D.; Bagnell, J.A.; Vandapel, N.; Hebert, M. Contextual classification with functional Max-Margin Markov networks. In Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA , 20–25 June 2009; pp. 975–982.
8. Hackel, T.; Savinov, N.; Ladicky, L.; Wegner, J.D.; Schindler, K.; Pollefeys, M. Semantic3D.net: A new large-scale point cloud classification benchmark. *ISPRS Ann. Photogramm. Remote Sens. Spat. Inf. Sci.* **2017**, *IV-1-W1*, 91–98. [CrossRef]
9. Roynard, X.; Deschaud, J.E.; Goulette, F. Paris-Lille-3D: A large and high-quality ground-truth urban point cloud dataset for automatic segmentation and classification. *Int. J. Robot. Res.* **2018**, *37*, 545–557. [CrossRef]

10. Tan, W.; Qin, N.; Ma, L.; Li, Y.; Du, J.; Cai, G.; Yang, K.; Li, J. Toronto-3D: A Large-scale Mobile LiDAR Dataset for Semantic Segmentation of Urban Roadways. *arXiv* **2020**, arXiv:2003.08284.
11. Scharwächter, T.; Enzweiler, M.; Franke, U.; Roth, S. Efficient multi-cue scene segmentation. In *German Conference on Pattern Recognition*; Springer: Berlin, Germany, 2013; pp. 435–445.
12. Geyer, J.; Kassahun, Y.; Mahmudi, M.; Ricou, X.; Durgesh, R.; Chung, A.S.; Hauswald, L.; Pham, V.H.; Mühlegg, M.; Dorn, S.; et al. A2D2: Audi Autonomous Driving Dataset. *arXiv* **2020**, arXiv:2004.06320.
13. Huang, R.; Xu, Y.; Hong, D.; Yao, W.; Ghamisi, P.; Stilla, U. Deep point embedding for urban classification using ALS point clouds: A new perspective from local to global. *ISPRS J. Photogramm. Remote Sens.* **2020**, *163*, 62–81. [CrossRef]
14. De Deuge, M.; Quadros, A.; Hung, C.; Douillard, B. Unsupervised feature learning for classification of outdoor 3D scans. In *Australasian Conference on Robitics and Automation*; University of New South Wales: Kensington, Australia, 2013; Volume 2, p. 1.
15. Vallet, B.; Brédif, M.; Serna, A.; Marcotegui, B.; Paparoditis, N. TerraMobilita/iQmulus urban point cloud analysis benchmark. *Comput. Graph.* **2015**, *49*, 126–133. [CrossRef]
16. Behley, J.; Garbade, M.; Milioto, A.; Quenzel, J.; Behnke, S.; Stachniss, C.; Gall, J. SemanticKITTI: A dataset for semantic scene understanding of LiDAR sequences. In Proceedings of the IEEE International Conference on Computer Vision, Seoul, Korea, 27 October–2 November 2019; pp. 9297–9307.
17. Geiger, A.; Lenz, P.; Stiller, C.; Urtasun, R. Vision meets Robotics: The KITTI Dataset. *Int. J. Robot. Res. (IJRR)* **2013**, *32*, 1231–1237. [CrossRef]
18. TUM City Campus—MLS Test Dataset. Available online: https://www.pf.bgu.tum.de/en/pub/tst.html (accessed on 8 June 2020).
19. Diehm, A.L.; Gehrung, J.; Hebel, M.; Arens, M. Extrinsic self-calibration of an operational mobile LiDAR system. In *Laser Radar Technology and Applications XXV*; Turner, M.D., Kamerman, G.W., Eds.; International Society for Optics and Photonics, SPIE: Paris, France, 2020; Volume 11410, pp. 46–61. [CrossRef]
20. Borgmann, B.; Schatz, V.; Kieritz, H.; Scherer-Klöckling, C.; Hebel, M.; Arens, M. Data Processing and Recording Using a Versatile Multi-sensor Vehicle. *ISPRS Ann. Photogramm. Remote Sens. Spat. Inf. Sci.* **2018**, *4*, 21–28. [CrossRef]
21. Hackel, T.; Wegner, J.D.; Schindler, K. Fast semantic segmentation of 3D point clouds with strongly varying density. *ISPRS Ann. Photogramm. Remote Sens. Spat. Inf. Sci.* **2016**, *3*, 177–184. [CrossRef]
22. Xu, Y.; Heogner, L.; Tuttas, S.; Stilla, U. A voxel- and graph-based strategy for segmenting man-made infrastructures using perceptual grouping laws: Comparison and evaluation. *Photogramm. Eng. Remote Sens.* **2018**. [CrossRef]
23. Qi, C.R.; Su, H.; Mo, K.; Guibas, L.J. Pointnet: Deep learning on point sets for 3D classification and segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 652–660.
24. Qi, C.R.; Yi, L.; Su, H.; Guibas, L.J. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. In Proceedings of the Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; pp. 5099–5108.
25. Xu, Y.; Ye, Z.; Yao, W.; Huang, R.; Tong, X.; Hoegner, L.; Stilla, U. Classification of LiDAR Point Clouds Using Supervoxel-Based Detrended Feature and Perception-Weighted Graphical Model. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2019**, 1–17. [CrossRef]
26. Everingham, M.; Van Gool, L.; Williams, C.K.; Winn, J.; Zisserman, A. The pascal visual object classes (VOC) challenge. *Int. J. Comput. Vis.* **2010**, *88*, 303–338. [CrossRef]
27. Li, W.; Wang, F.D.; Xia, G.S. A geometry-attentional network for ALS point cloud classification. *ISPRS J. Photogramm. Remote Sens.* **2020**, *164*, 26–40. [CrossRef]