remote sensing

MDPI

Article

CLOUDSPAM: Contrastive Learning On Unlabeled Data for Segmentation and Pre-Training Using Aggregated Point Clouds and MoCo

Reza Mahmoudi Kouhi [1,*], Olivier Stocker [1], Philippe Giguère [2] and Sylvie Daniel [1]

[1] Department of Geomatics Sciences, Université Laval, Québec City, QC G1V 0A6, Canada; olivier.stocker.1@ulaval.ca (O.S.); sylvie.daniel@scg.ulaval.ca (S.D.)

[2] Department of Computer Science and Software Engineering, Université Laval, Québec City, QC G1V 0A6, Canada; philippe.giguere@ift.ulaval.ca

* Correspondence: reza.mahmoudi-kouhi.1@ulaval.ca

Abstract: SegContrast first paved the way for contrastive learning on outdoor point clouds. Its original formulation targeted individual scans in applications like autonomous driving and object detection. However, mobile mapping purposes such as digital twin cities and urban planning require large-scale dense datasets to capture the full complexity and diversity present in outdoor environments. In this paper, the SegContrast method is revisited and adapted to overcome its limitations associated with mobile mapping datasets, namely the scarcity of contrastive pairs and memory constraints. To overcome the scarcity of contrastive pairs, we propose the merging of heterogeneous datasets. However, this merging is not a straightforward procedure due to the variety of size and number of points in the point clouds of these datasets. Therefore, a data augmentation approach is designed to create a vast number of segments while optimizing the size of the point cloud samples to the allocated memory. This methodology, called CLOUDSPAM, guarantees the performance of the self-supervised model for both small- and large-scale mobile mapping point clouds. Overall, the results demonstrate the benefits of utilizing datasets with a wide range of densities and class diversity. CLOUDSPAM matched the state of the art on the KITTI-360 dataset, with a 63.6% mIoU, and came in second place on the Toronto-3D dataset. Finally, CLOUDSPAM achieved competitive results against its fully supervised counterpart with only 10% of labeled data.

Keywords: self-supervised; contrastive learning; label-efficient learning; mobile mapping; LiDAR point cloud

check for updates

Citation: Mahmoudi Kouhi, R.; Stocker, O.; Giguère, P.; Daniel, S. CLOUDSPAM: Contrastive Learning On Unlabeled Data for Segmentation and Pre-Training Using Aggregated Point Clouds and MoCo. *Remote Sens.* **2024**, *16*, 3984. https://doi.org/10.3390/rs16213984

Academic Editors: Dong Chen, Jiaming Na, Jiju Poovvancheri and Norbert Pfeifer

Received: 19 September 2024
Revised: 23 October 2024
Accepted: 24 October 2024
Published: 26 October 2024

1. Introduction

In recent years, LiDAR technology has emerged as a valuable tool for capturing detailed 3D environmental geometry [1–3]. However, labeling outdoor 3D LiDAR point clouds is a time-consuming and labor-intensive process, which significantly limits their practical utility in various applications, such as urban planning, smart cities, autonomous driving, and environmental monitoring [4–6]. Consequently, the task of semantic labeling for outdoor LiDAR point clouds holds great importance in the fields of computer vision and 3D data analysis. Previous advancements in this area have relied on supervised learning techniques based on labeled datasets such as SemanticKITTI [7] and KITTI-360 [8]. These datasets contain over 6 billion manually labeled points, providing valuable training data. However, the manual labeling process not only hinders scalability but also introduces inconsistencies in the labeled data, challenging the overall effectiveness of supervised learning approaches [4].

To address these challenges, recent research has focused on developing self-supervised learning methods that leverage unlabeled LiDAR point clouds for semantic segmentation [4]. Self-supervised learning is a machine learning technique that allows systems to learn data representations without explicit human supervision. To this end, a pretext task needs to be designed to guide the networks to capture the inherent structure and relationships within the data [4,9]. Among these self-supervised methods, contrastive learning has demonstrated promising results in various downstream tasks, such as semantic segmentation, classification, and object detection [9–16]. Their pretext task aims to pull together, in latent space, representation of similar samples (positive pairs) while pushing them apart from representations of different samples (negative pairs) [11].

While contrastive learning has shown potential for 3D LiDAR point clouds, most existing works have primarily focused on indoor datasets [10,12,17]. Some studies have attempted to apply contrastive learning to unlabeled outdoor point clouds, such as SegContrast [13]. These approaches were only designed for individual scans in the context of autonomous driving and object detection [13,15]. The data structure of individual scan datasets is fundamentally different from that of mobile mapping datasets. The former contains multiple point clouds, spanning a few square kilometers, with an average point density of 10 pts/m^2, while the latter is composed of a few point clouds spanning more than 100 square kilometers, with an average range of 200 pts/m^2. As such, individual scans fail to capture the full complexity and diversity present in outdoor environments due to their limited density, range, and coverage. Class distribution and class diversity are also wildly different and further emphasize the need for tailored approaches in contrastive learning [7,8].

Two constraints arise from this disparity in the data structures of mobile mapping point clouds. The first is the scarcity of contrastive pairs. With fewer point clouds, the number of positive and negative pairs is reduced, hindering the richness brought by the contrastive learning pre-training phase. The second concerns memory limitations, owing to which point density is drastically increased compared to scan point clouds. Thus, fitting the same number of points as the scan point clouds in memory means sampling smaller areas and smaller contexts, resulting in weaker features describing the neighborhood. A naive solution to this constraint could be downsampling. However, contrastive learning, as a self-supervised method, relies on the capture of strong features from the scene to learn data representations without labels. Downsampling is inadequate in this context, as it would eliminate important details and local features, preventing the network from fully learning the intricate aspects of the scene.

In summary, existing contrastive learning methods cannot be applied directly to large-scale mobile mapping point clouds and need to be re-thought specifically for mobile mapping applications.

In this paper, we present a solution to bridge the gap in contrastive learning for real-world applications by addressing the challenges posed by mobile mapping datasets. To overcome the above-mentioned constraints, we propose the merging of heterogeneous datasets, specifically KITTI-360 [8], Toronto-3D [18], and Paris-Lille-3D [19]. This approach enriches the pool of positive and negative pairs, improving the model's versatility and generalization during pre-training. However, directly merging these datasets is not straightforward is expected to worsen the second constraint. The point clouds from these datasets vary significantly in size, and directly merging them would result in an excessively large number of points, exceeding memory capacity. To address these issues, we developed a novel data augmentation strategy that increases the number of point clouds, standardizes their size to fulfill the memory constraints, and provides enough positive and negative pairs for contrastive loss.

To assess the benefit of the proposed approach and strategy, we used it to adapt a contrastive learning method. SegContrast was selected because it was the state of the art at the time of this research. By applying this data augmentation approach, the SegContrast framework was successfully adapted to mobile mapping datasets, enhancing its effectiveness for real-world applications. This methodology guarantees the performance of self-supervised learning for small-scale datasets (with fewer than 100 million points, such as the Toronto3D dataset). This learning strategy is named "CLOUDSPAM: Contrastive

Learning On Unlabeled Data for Segmentation and Pre-training using Aggregated point clouds and MoCo", as depicted in Figure 1. CLOUDSPAM was evaluated against a classical supervised approach and with different labeled data ratios and matched the state of the art on the KITTI-360 dataset. In short, the key contributions of this work are outlined as follows:

- We adapt a contrastive learning approach, namely SegContrast, to address the challenges of large-scale, mobile mapping with 3D LiDAR point cloud;
- We design a data augmentation approach for mobile mapping point clouds;
- We leverage merged heterogeneous mobile mapping datasets during the pre-training phase of self-supervised learning to provide enough positive and negative pairs for contrastive learning, thereby improving accuracy and generalizability.

A review of existing works related to contrastive learning applied to large-scale mobile mapping LiDAR point clouds is provided below.

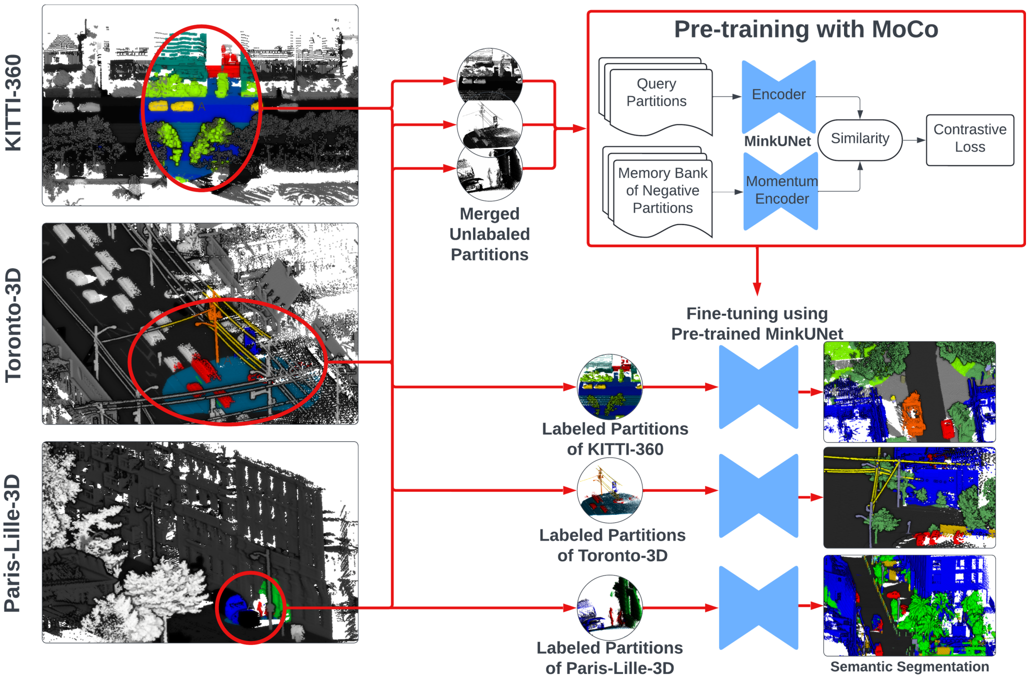

KITTI-360
Toronto-3D
Paris-Lille-3D
Merged Unlabaled Partitions
Pre-training with MoCo
Query Partitions
Encoder
MinkUNet
Similarity
Contrastive Loss
Memory Bank of Negative Partitions
Momentum Encoder
Fine-tuning using Pre-trained MinkUNet
Labeled Partitions of KITTI-360
Labeled Partitions of Toronto-3D
Labeled Partitions of Paris-Lille-3D
Semantic Segmentation

Figure 1. An overview of CLOUDSPAM. Leveraging the proposed data augmentation method, heterogeneous mobile mapping point clouds are merged for pre-training with MoCo (Momentum Contrast). During the pre-training phase, the "query partitions" represent the positive pairs processed by the encoder, while the "memory Bank" contains the negative pairs input into the momentum encoder. Subsequently, fine-tuning is conducted separately for each dataset using the labeled partitions generated by the proposed data augmentation method.

Literature Review

Recently, researchers have been exploring the performance of contrastive learning for 3D LiDAR point clouds [4]. In a contrastive learning approach, the representation learning procedure starts with the selection of positive and negative pairs. This selection process involves augmenting the original data to create two different augmented versions. These augmented versions are referred to as "*query*" and "*key*" datasets. For each data point in the query dataset, the positive pair is defined as another data point that shares the same semantic label. On the other hand, the negative pairs are selected by randomly sampling data points from the key dataset that does not have the same semantic label as the corresponding query data point [11]. By contrasting positive and negative pairs during the training process, the model learns to map similar data points close together in the feature space while pushing them away from dissimilar data points [4]. This contrast helps the model learn to distinguish data points with different semantic meanings.

PointContrast [12], a pioneering method, introduces PointInfoNCE loss, inspired by InfoNCE loss [20], to learn effective representations from unlabeled point clouds. PointContrast employs a strategy to select positive pairs for the query point within a certain radius in the key point cloud. This encourages the model to learn representations that capture the local geometry and semantic information of the point cloud. While PointContrast has shown promising results for unsupervised pre-training of 3D point clouds, there are some limitations associated with this method. One of them is its disregard for spatial contexts during pre-training. This can be problematic for tasks that require an understanding of the spatial relationships and dependencies between points, such as semantic segmentation. Moreover, the method may face scalability challenges when applied to large-scale point clouds. As the size of the point cloud increases, the number of negative samples also increases, making contrastive loss more computationally expensive and memory-intensive [10].

To address such limitations, contrastive scene contrast [10] was proposed, offering a novel approach to positive and negative pair selection inspired by ShapeContext [21–23]. This approach effectively utilizes both point-level correspondences and spatial contexts within a scene. By dividing the space into different cells based on the relative distances and angles between points, contrastive scene contexts enable contrastive learning to be performed independently within each spatial cell. To incorporate spatial information, negative samples are sampled within these spatial cells. The performance of this method is highly influenced by the selection of hyperparameters, particularly the radius used to create the spatial cells for the selection of negative pairs. The choice of the radius should be context-dependent, taking into consideration the characteristics of the point cloud data. Contrastive scene contrast [10] considers a relatively small radius based on the specific context of indoor point clouds utilized in the experiments. However, applying contrastive learning to mobile mapping LiDAR point clouds presents its own challenges. The diversity of mobile mapping LiDAR point clouds in terms of point density poses difficulties in defining meaningful negative and positive samples for contrastive loss. Additionally, a small radius is insufficient to cover complex and large objects in outdoor point clouds, while increasing the radius is impractical in terms of memory and time efficiency. Furthermore, uneven density and incomplete coverage of outdoor LiDAR data, stemming from differences in angles and distances of data collection, makes it challenging to develop a robust and generalizable model [17].

Most existing research on self-supervised contrastive learning for semantic labeling of 3D LiDAR point clouds has predominantly focused on indoor datasets [4,17]. However, the challenge of large-scale, outdoor 3D LiDAR point clouds has received limited attention until recently. Pair selection poses a significant challenge in contrastive learning, particularly in the context of outdoor point clouds. The inherent diversity and sparsity of such point clouds in terms of density and class make it even more demanding to identify meaningful and true positive and negative pairs. While sphere-based methods like those employed in [10,12] fall short in producing comprehensive object representations due to the limited size of receptive fields and excessive focus on fine-grained features, region-based approaches such as SegContrast prove more suitable for semantic segmentation tasks in outdoor environments. SegContrast [13] stands out as a state-of-the-art method utilizing self-supervised contrastive learning to acquire representations from 3D LiDAR data. SegContrast introduces a unique approach for selecting positive and negative pairs. It begins by extracting class-agnostic segments from the point cloud and utilizes segment-wise contrastive loss on augmented pairs of these segments. This approach allows SegContrast to exploit contextual information, resulting in superior performance compared to other contrastive learning methods. Notably, SegContrast demonstrates remarkable advantages, even when provided with limited labeled data, such as only 1% of the total data. It generates robust and detailed feature representations and exhibits strong transferability across diverse datasets [17]. Nevertheless, its application to mobile mapping 3D

LiDAR point clouds has not yet been explored. Hence, this paper specifically addresses this particular challenge by investigating the adaptability of SegContrast to mobile mapping LiDAR point clouds.

The rest of this paper is organized as follows. Section 2 presents the methodology and the datasets used in the experiments. The results are presented in Section 3, and a thorough analysis and discussion of the results are presented in Section 4.

2. Materials and Methods

In this section, the methodology employed in this study is described, starting with revisiting of the pre-training methodology proposed in the SegContrast paper and outlining its limits regarding its application to mobile mapping LiDAR point clouds. Subsequently, the specific adaptations and approaches developed to overcome the challenges posed by such datasets and enhance the performance of SegContrast within this application are presented.

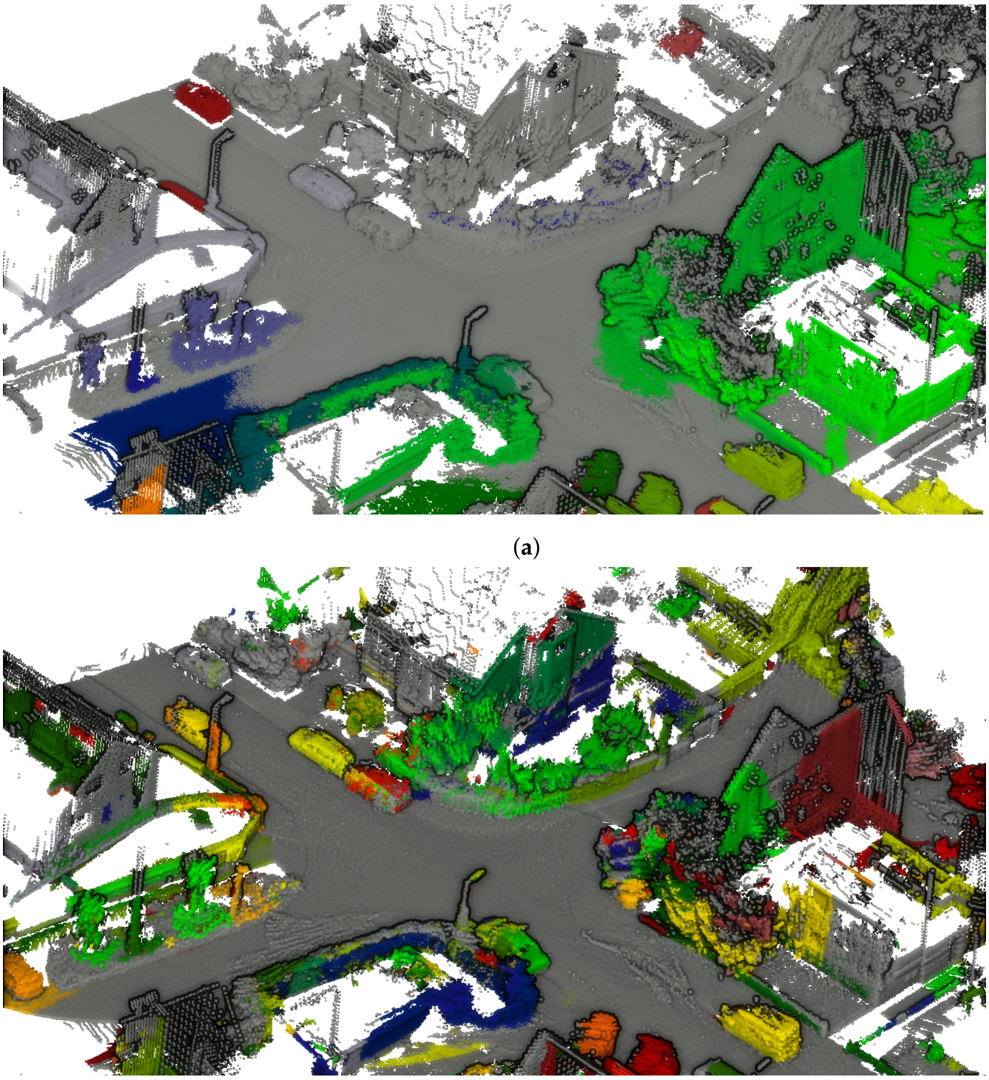
2.1. Revisiting SegContrast the Pre-Training Pipeline

SegContrast uses Momentum Contrast (MoCo) [11] as the pre-training pipeline. MoCo offers a highly effective and scalable approach to unsupervised representation learning. By contrasting positive and negative pairs and leveraging a large memory bank for negative sampling, MoCo efficiently learns high-quality representations from large-scale unlabeled datasets. Its robustness to data augmentation, memory-efficient training, and strong transfer learning performance make it a versatile and powerful framework for pre-training, facilitating effective generalization to downstream tasks without the need for labeled data, thereby significantly reducing annotation costs and data dependencies in various computer vision applications [4].

To employ MoCo to 3D point cloud learning, the SegContrast method implements a pair selection strategy by extracting segments from unlabeled point clouds and utilizing them as positive and negative pairs. The process begins with ground removal using RANSAC [24] and clustering of the remaining points with DBSCAN [25] to obtain segments. The segmented point indices are preserved through the pre-training process. Two augmented views are generated through random transformations of the point cloud. These augmented views are then processed by a backbone network to compute point-wise features. Then, augmented segments are extracted from these point-wise features based on the point segment indices. Finally, the contrastive loss is calculated to differentiate between positive and negative pairs, enabling effective pre-training for segmentation tasks.

2.2. Adapting the SegContrast Pre-Training Pipeline for Mobile Mapping Point Clouds

To adapt the SegContrast pre-training pipeline to the context of mobile mapping LiDAR point clouds, it is imperative to acknowledge and address the shortcomings encountered when applying SegContrast to aggregated point clouds. Using the KITTI-360 dataset as an example, an aggregated cloud is formed by combining hundreds of significantly overlapping scans, creating a large-scale mobile mapping point cloud. Consequently, the number of segments extracted from this single aggregated cloud is significantly lower than the total number of segments extracted in all the individual scans. Thus, as mentioned in the introduction, the contrastive approach applied to aggregated point clouds faces a scarcity of contrastive pairs. Moreover, during the backbone's forward pass, the model needs to consider the relationship between the segments and the scene, leveraging the contextual information from the entire point cloud [13]. However, due to memory limitations, processing the complete aggregated point cloud in the forward pass can lead to memory overflow and hinder the training process. Finally, the reliability of RANSAC for ground segmentation is limited when dealing with sloped terrain or complex surfaces [26]. While individual scans typically cover small areas with minimal slope complexity, aggregated point clouds can span larger areas with diverse terrain and topographies. To illustrate this issue, Figure 2a shows the segmentation of an aggregated point cloud using RANSAC and DBSCAN. As can be seen, RANSAC could not separate the ground properly from other classes, such as cars, buildings, and vegetation. Thus, the majority of the points were segmented as ground.



(**a**)

(**b**)

Figure 2. Segmentation of the KITTI-360 dataset (**a**) without the proposed data augmentation and (**b**) with the proposed data augmentation. The ground segment, computed using RANSAC, is displayed in gray. All the other segments, computed using the DBSCAN algorithm, are shown in colors other than gray.

2.2.1. Dedicated Data Augmentation Approach

Addressing these limitations and adapting SegContrast to the context of mobile mapping LiDAR point clouds requires strategies carefully designed to partition the aggregated point clouds, augment the number of segments, decrease the number of input points, and reduce the complexity of the ground topography to be able to successfully apply RANSAC and DBSCAN.

To this end, a dedicated data augmentation approach tailored for efficient segmentation is introduced, drawing from a two-step guideline outlined in [27]. The first step involves selecting seed points as initial representatives for partitioning, followed by the identification of neighboring points around each seed point to form partitions. Building upon these two steps, a partitioning method is devised that is customized to the characteristics of the dataset. For seed point selection, the enhanced Furthest Point Sampling

(FPS) method proposed in [27] is leveraged, which involves splitting the point cloud into smaller blocks before applying FPS. This optimization significantly reduces computing time while ensuring robust seed point selection, which is a crucial aspect, especially for large-scale mobile mapping point clouds, where computational efficiency is paramount. Figure 3a illustrates the distribution of selected seed points (purple squares) in an aggregated point cloud, showcasing coverage across both low- and high-density areas, leading to diverse partitions in terms of point density and class distribution, which are essential for the pre-training process.

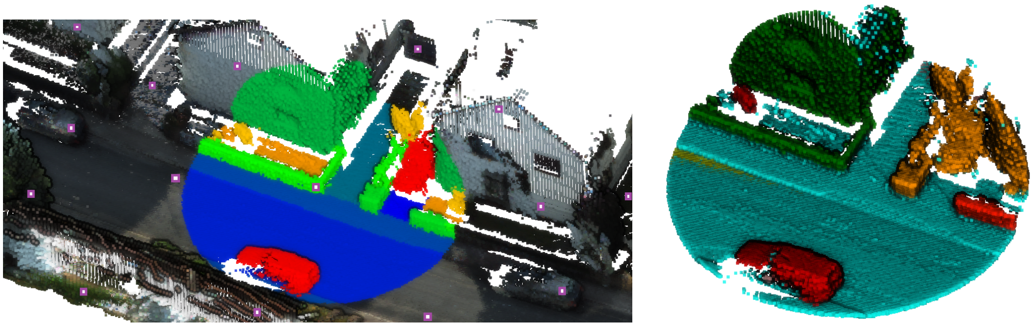


Figure 3. Visualization of (**a**) one partition extracted from the aggregated KITTI-360 dataset using the proposed partitioning approach and (**b**) its associated segments. White and purple squares represent the seed points selected with the FPS approach over this area. Colors in (**a**) represent true labels, while those in (**b**) represent different segments.

Subsequently, neighboring points are grouped by selecting the K nearest neighbors (KNN) of each seed point. This approach offers several advantages, particularly for datasets with uneven point density, the like Toronto-3D and KITTI-360 datasets. By selecting the K nearest neighbors of each seed point, irrespective of their distances, we ensure comprehensive coverage of the point cloud, capturing both local and global contextual information. This comprehensive grouping strategy contributes to the creation of partitions that accurately represent the underlying structure of the point cloud, which is essential for subsequent pre-training and fine-tuning tasks. Using these partitions at the network input addresses the memory limitation, as each partition contains a fixed number of points (K points), preventing memory overflow. Additionally, employing the proposed partitioning approach allows us to create sets of both labeled and unlabeled point cloud partitions for self-supervised pre-training and supervised fine tuning, respectively. Figure 3b depicts a partition and its associated segments, showcasing the effectiveness of KNN in selecting points to obtain a relevant representation of the scene contents while simplifying its complexity for RANSAC and DBSCAN segmentation, yielding high-quality segments for pre-training. The impact of the dedicated data augmentation approach on the efficiency of RANSAC and DBSCAN segmentation is shown in Figure 2a. We named this approach "data augmentation" because it involves cropping the mobile mapping point clouds and generating multiple partitions from a scene, thereby augmenting the dataset with additional points for training. Additionally, unlike conventional data augmentation techniques that occur during each training iteration, this process is conducted offline before training begins.

2.2.2. Heterogeneous Dataset Merging

As underlined in the previous paragraph, the benefit of the dedicated data augmentation approach is the relevance of the segments for pre-training. However, the scarcity of segments remains an issue for small-scale datasets, such as Toronto-3D, for contrastive learning. To overcome this problem, we propose the merging of heterogeneous mobile mapping datasets. Dataset merging presents an opportunity to combine the strengths of multiple sources, enhancing the diversity and richness of the pre-training data. By combining datasets, the pool of available segments is augmented, thereby enhancing the generalization and performance of pre-trained models. It is important to stress that this merging of heterogeneous datasets is not possible with SegContrast, as its architecture is unable to absorb such a massive volume of points. The point cloud partitioning approach proposed in our adaptation of SegContrast allows us to overcome this obstacle.

A comparative study was conducted to demonstrate the feasibility of self-supervised contrastive learning on mobile mapping datasets. We utilized the proposed data augmentation method and examined the benefits of leveraging merged heterogeneous datasets. This study involved three learning strategies. The first one, called Baseline, is a supervised baseline with MinkUNet [28] trained with aggregated point clouds. MinkUnet was chosen because it is the network used in the SegContrast pipeline. The second, called DA-supervised, is the same supervised network as the Baseline but trained with partitions from the aggregated point clouds computed using the proposed data augmentation method. The third, called CLOUDSPAM, is our adaptation of SegContrast. Thus, it consists of self-supervised pre-training with MoCo using unlabeled partitions from the merged heterogeneous datasets, followed by supervised fine-tuning using labeled partitions from the aggregated point clouds of the targeted dataset using the proposed data augmentation method. The three learning strategies are shown in Figure 4. The first aim of this comparative study is to highlight the improvement brought about by diversity-rich labeled data using the proposed data augmentation approach in the context of supervised learning. Secondly, it aims to demonstrate the ability of the contrastive approach conditioned by our adaptations in comparison with that of the supervised approach. The three learning strategies are outlined as follows:

1. Baseline: a classical supervised baseline with MinkUNet [28] trained with original data.
2. DA-supervised: A classical supervised baseline with MinkUNet [28] trained with augmented partitions.
3. CLOUDSPAM: A self-supervised pre-training algorithm with MoCo [11] using unlabeled partitions from merged heterogeneous datasets, followed by supervised fine-tuning with labeled partitions from the targeted dataset.

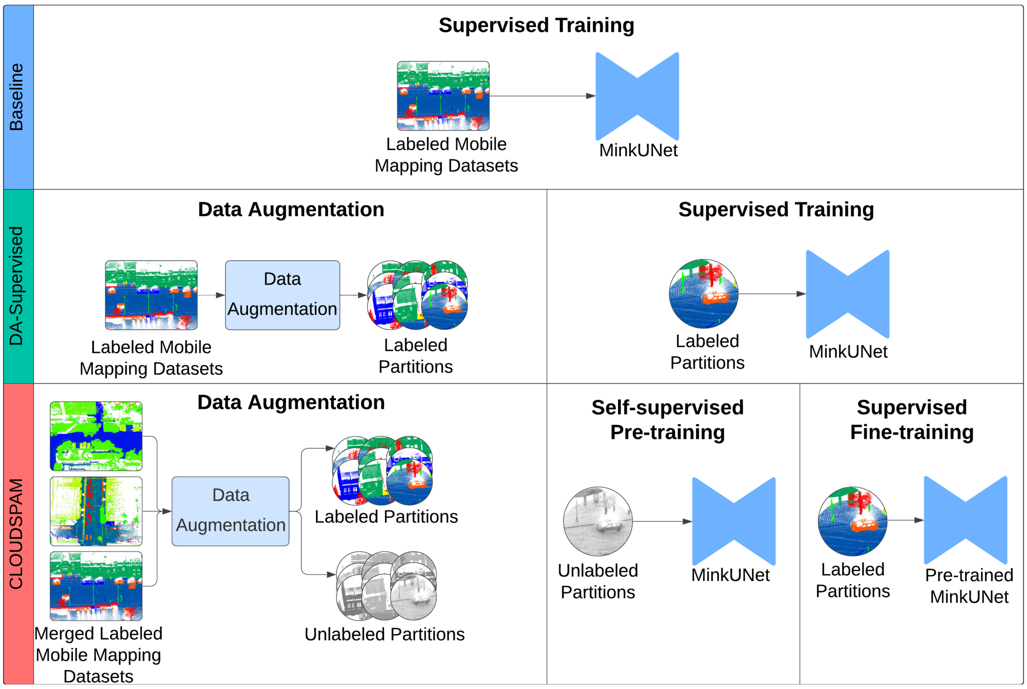


Figure 4. Overview of three learning strategies used in the comparative study. "Baseline" strategy refers to a supervised training. "DA supervised" is equivalent to the "Baseline" but using labeled partitions generated with the proposed data augmentation approach. The "CLOUDSPAM" strategy

refers to self-supervised pre-training with MoCo using unlabeled partitions, followed by supervised fine-tuning using labeled partitions, with both labeled and unlabeled partitions provided by the proposed data augmentation approach.

2.3. Data and Experimental Configuration

For the comparative study, three datasets were chosen that were collected in three cities in Europe and Canada, namely the KITTI-360, Paris-Lille 3D, and Toronto 3D datasets. The following paragraphs describe these datasets and their main differences:

- **KITTI-360:** This dataset covers a 73.7 km of medium-population-density streets in Karlsruhe, Germany. It consists of 9 labeled sequences containing over 1.2 billion points for training and more than 340 million points for validation. This dataset encompasses 46 classes grouped into 7 categories. The point clouds were post-processed, and point density was uniformized. A single aggregated point cloud of this dataset has an average of 2.5 million points, with a density of around 500 pts/m^2.
- **Paris-Lille3D:** This dataset covers 1.9 km of urban streets of Paris and Lille in France and contains 119.8 million points. The dataset encompasses 50 classes grouped into 9 categories. It is split into 4 point clouds, with point density ranging from 1000 pts/m^2 to 2000 pts/m^2. The test datasets were published as an add-on, and have different locations; 1 point cloud was acquired in Dijon, and 2 were acquired in Ajaccio, France. Each of them consists of exactly 10 million points.
- **Toronto-3D:** This dataset covers a 1 km road in a dense suburban area of Toronto, Canada. It contains 78.3 million points split into 8 classes. The dataset is divided into 4 sections within a driving distance of 250 m. In addition, there is overlap among the sections. The second section is kept as a test and contains 6.7 million points. The Toronto-3D dataset differs from the two other datasets due to its significant disparity in point density. In Toronto-3D, every point is detected within a 100 m LiDAR range, while a 20 m range is used in the other datasets. Moreover, there no post-processing trimming or downsampling was applied to this dataset.

MinkUNet is used was the backbone architecture for all experiments. MinkUNet works with sparse voxels thanks to its Minkowski engine [28] and requires voxelization of the point space. A voxel size of 5 centimeters was selected for all experiments. In order to input the point clouds into MinkUNet for the classical supervised baseline, we did not apply any data pre-processing other than the default MinkUnet voxelization and random selection. A value of 600K was chosen for the random point selection of MinkUNet, with a batch size of 2 in all of the baseline experiments. For the Paris-Lille-3D and Toronto-3D datasets, each point cloud was divided into sub-clouds of around 2 million points. For each of the learning strategies, the supervised phases were implemented using the following six labeled data regimes: 1%, 2%, 10%, 20%, 50%, and 100% availability of labeled data. As a result, 6 experiments were carried out for each of the three datasets and each of the three learning strategies, for a total of 54 experiments. In KNN point selection in the data augmentation approach, a K value equal to 131,072 (2^{17}) was chosen to generate partitions representing scene with diverse classes. This value is also close to the average number of points per scan in datasets such as SemanticKITTI [7]. DBSCAN segmentation uses two parameters—namely, Epsilon, a distance measure used to locate the points in the neighborhood of any point, and the minimum number of points clustered together for a region to be considered dense. Different combinations of values were tested to maximize the purity of segments, as well as the total number of segmented points. Here, purity refers to the property of a segment only containing points from one class. Selecting 0.25 for Epsilon and 10 for the minimum number of points yielded the best results. For self-supervised pre-training, a learning rate of 0.12 was used, and for all of the supervised training and fine-tuning experiments, a learning rate of 0.24 was chosen. All experiments were performed using an NVIDIA V100 Volta graphics card.

3. Results

Table 1 presents the semantic segmentation results for each of the three datasets, namely KITTI-360, Toronto-3D, and Paris-Lille-3D, and for three learning strategies, namely Baseline, DA-supervised, and CLOUDSPAM. For each dataset and strategy, the table reports segmentation performance for the following percentages of labeled data: 1%, 2%, 10%, 20%, 50%, and 100%. The segmentation performance is measured in terms of Intersection over Union (IoU). The results of the table highlight improvements achieved by the DA-supervised and CLOUDSPAM methods over the baseline approach across all datasets, confirming the benefit of the proposed data augmentation approach and contrastive learning adaptation in enhancing the performance of semantic segmentation models in urban scene understanding tasks, especially when labeled data are limited or expensive to obtain.

Table 1. Semantic segmentation results (% mIoU) on the validation set of KITTI-360 and the test sets of Toronto-3D and Paris-Lille-3D. "Baseline" denotes supervised training using MinkUNet without any preprocessing. "DA-supervised" is the proposed supervised training approach with augmented data and MinkUNe. "CLOUDSPAM" is our self-supervised pre-training approach using the three merged datasets, followed by supervised fine tuning using only the targeted dataset.

Labeled Dataset	Method	1%	2%	10%	20%	50%	100%
KITTI-360	Baseline	23.1%	29.1%	37.9%	39.1%	41.3%	51.0%
	DA-supervised	38.1%	42.4%	52.4%	58.3%	**61.9%**	**64.1%**
	CLOUDSPAM	**41.3%**	**46.3%**	**53.3%**	**59.0%**	**61.9%**	63.6%
Toronto-3D	Baseline	27.7%	29.8%	38.4%	39.9%	41.9%	57.0%
	DA-supervised	47.8%	54.4%	59.2%	66.0%	69.7%	69.3%
	CLOUDSPAM	**49.3%**	**65.1%**	**62.7%**	**70.4%**	**71.3%**	**71.8%**
Paris-Lille-3D	Baseline	32.7%	45.9%	52.1%	57.2%	69.1%	68.9%
	DA-supervised	33.4%	44.6%	52.9%	55.2%	66.5%	63.8%
	CLOUDSPAM	**44.1%**	**55.5%**	**60.1%**	**66.7%**	**70.8%**	**73.8%**

4. Discussion

In the subsequent subsections, we offer a comprehensive analysis of the results, focusing on the outcomes of each learning strategy. This detailed examination aims to provide deeper insights into the performance variations and effectiveness of the investigated approaches.

4.1. DA-Supervised

The effectiveness of the proposed data augmentation approach is underscored by the notable improvements observed across the KITTI-360 and Toronto-3D datasets. As illustrated in Table 1, substantial increases in mIoU scores of 15% and 20% at the lowest data regime can be observed on the KITTI-360 and Toronto-3D dataset, respectively. While these performance boosts slightly diminish with full data training, they still contributes significant gains of 13% and 12% for the KITTI-360 and Toronto-3D datasets, respectively. For the Paris-Lille-3D dataset, a different pattern is observed. Indeed, data augmentation appears to impede learning outcomes across all data regimes. A closer examination of the cloud statistics presented in Table 2 reveals a notable distinction between Paris-Lille-3D and the other two datasets; the average radius of point clouds is smaller than the radius for the KITTI-360 and Toronto-3D datasets (12 m, 20 m, and 25 m, respectively). This discrepancy suggests that the proposed data augmentation method struggles with high-density point clouds, resulting in the extraction of smaller segments. Nevertheless, for the KITTI-360 and Toronto-3D datasets, the proposed data augmentation approach proves highly effective, particularly in scenarios with limited data availability. These findings underscore the critical role of meticulous data preparation and selection in the context of deep learning-based point cloud semantic segmentation.

4.2. CLOUDSPAM

CLOUDSPAM showcases significant performance improvements over the baseline, even surpassing the DA-supervised consistently. Notably, in the case of the Paris-Lille-3D dataset, where the proposed data augmentation approach hindered the learning quality, the contrastive learning process achieved superior results, yielding enhancements ranging from 5% to 12% compared to the baseline across various data regimes.

Regarding label-efficient learning, the findings in this research demonstrate the advantages of dataset merging in self-supervised pre-training. This merging strategy enables the network to start the supervised phase (fine tuning) from a more favorable initialization state compared to random initialization. Consequently, the network achieves notable gains in mIoU scores, ranging from 11% to 18%, with only 1% of labeled data available across all three datasets.

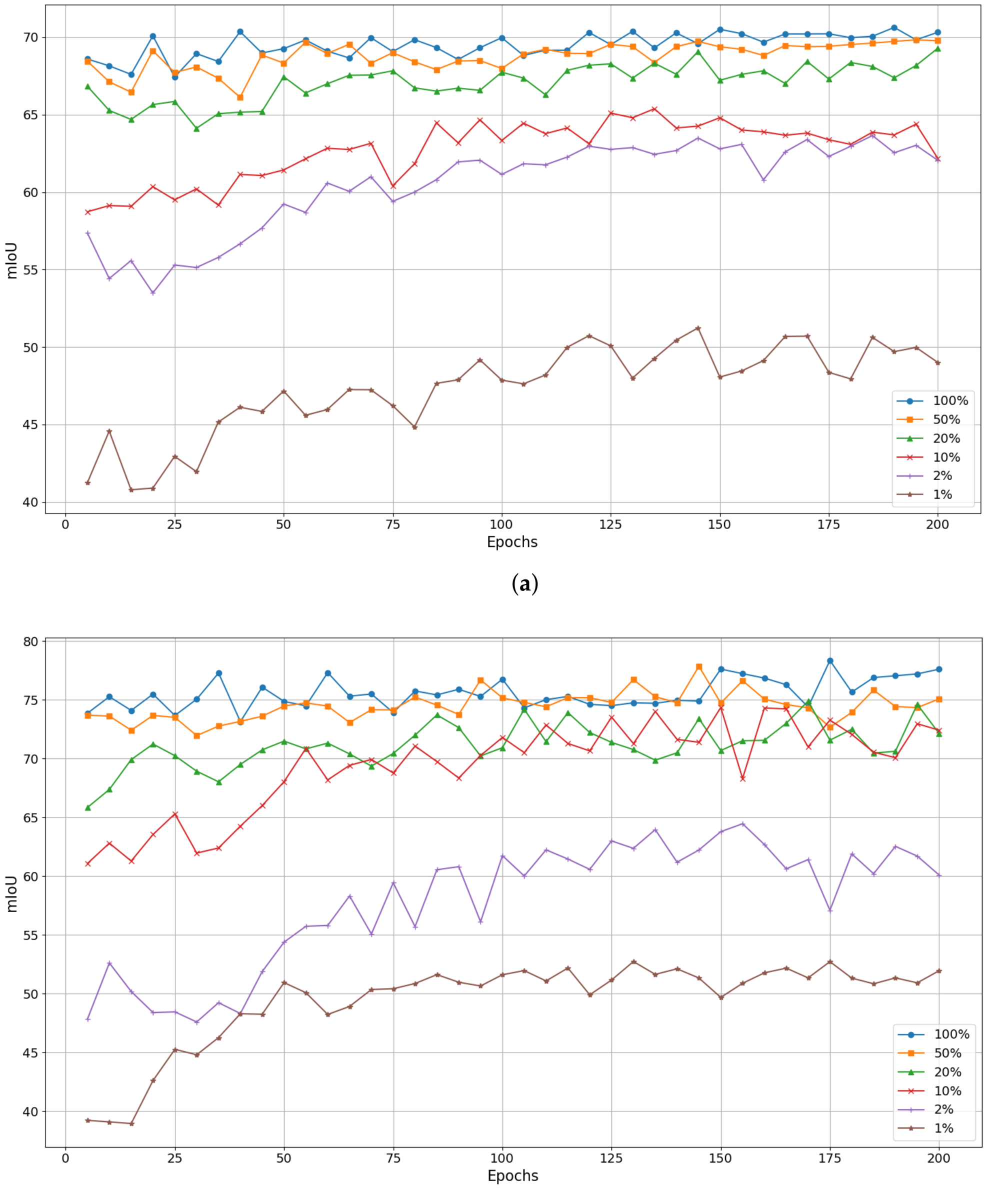
CLOUDSPAM has achieves a lower mIoU score than DA-supervised only when training with 100% of the KITTI-360 dataset. Upon closer examination of Table 2, it becomes evident that, post data augmentation, the KITTI-360 dataset contains approximately 1.9 billion points, which is four times more than the Paris-Lille-3D dataset and eleven times more than the Toronto-3D dataset. This substantial amount of labeled data appears sufficient for efficient training under the 100% data regime. Additionally, pre-training of CLOUDSPAM was conducted using the merged datasets, potentially resulting in a less specialized pre-trained network state for the KITTI-360 dataset. The architecture of MinkUNet might not have been sufficiently deep to guarantee the learned features during the pre-training stage from being overwritten during fine tuning using 100% of labeled data. Consequently, employing a deeper network architecture could potentially yield better results for the CLOUDSPAM process, particularly in this specific scenario.

Table 2. Statistical analysis of the data augmentation approach applied to the three datasets.

	KITTI-360		**Toronto-3D**		**Paris-Lille-3D**	
	Original	**Ours**	**Original**	**Ours**	**Original**	**Ours**
# of segments	11,950	552,012	1348	46,426	2405	48,532
# of clouds	239	14,340	3	1231	4	1580
Avg. # of pts per cloud	2,689,600	131,072	12,866,207	131,072	29,945,846	131,072
Avg. radius of clouds (m)	113	20	125	25	300	12
Total # of points (in millions)	1200.0	1,879.5	78.3	161.3	119.8	205.9

4.3. Impact of Pre-Training

To delve deeper into the implications of the pre-training step, fine tuning iterations were conducted every five epochs, evaluating the mIoU score without voting. This analysis was performed across various data regimes with the Paris-Lille-3D validation set and the Toronto-3D test set. Due to computational constraints, similar experiments were not conducted with the KITTI-360 dataset. The results, as depicted in Figure 5, highlight the effectiveness of pre-training, particularly on data regimes with limited labeled data. Notably, the initial 100 epochs are pivotal for lower-data regimes, whereas prolonged pre-training demonstrates greater efficiency for higher-data regimes. In summary, these findings underscore the following two key points: Firstly, the contrastive self-supervised pre-training facilitated by the proposed data augmentation method proves highly effective; secondly, dataset merging serves to create a versatile network suitable for label-efficient learning tasks.



(**a**)

(**b**)

Figure 5. Comparison of mIoU (%) scores of CLOUDSPAM per epoch of pre-training for each of 6 data regimes on (**a**) the test set of the Toronto-3D dataset and (**b**) the validation set of the Paris-Lille-3D dataset.

4.4. Impact of Data Augmentation

As previously emphasized, the availability of numerous positive and negative pairs is pivotal for meaningful pre-training via contrastive learning, requiring both segment purity and quantity.

In terms of segment quantity, Table 2 provides a comparison of statistics before and after implementing the proposed data augmentation method. This approach substantially boosts the number of segments across all three datasets. Specifically, for the KITTI-360 dataset, a significant increase can be witnessed, multiplying the number of segments by 45. Similarly, for Toronto-3D and Paris-Lille-3D datasets, there are 34 and 20 times as many segments, respectively. Visual assessment, as depicted in Figure 2, confirms this increase in the number of segments.

Given that point cloud partitions overlap, the total number of points per dataset is effectively increased. A uniform distribution of seed points is achieved for partitioning using FPS, ensuring that overlapping points only pertain to the external zone of partitions.

Additionally, Figure 6 demonstrates how segments at the same location can differ from one partition to another. Consequently, duplicated points provide new information to facilitate the refinement of feature representation for negative and positive pairs.

Regarding segment purity, a purity analysis was conducted for the KITTI-360 dataset. It involved comparing true labels against segments generated by the proposed data augmentation method. Impressively, 94.28% of the points were segmented while creating only 18.72% mixed segments. It is statistically logical that mixed segments tend to be larger than pure ones, as the probability of containing points from different classes correlates with the segment's size.

In summary, by enhancing both segment quantity and purity, the execution of contrastive learning on mobile mapping datasets was enabled.

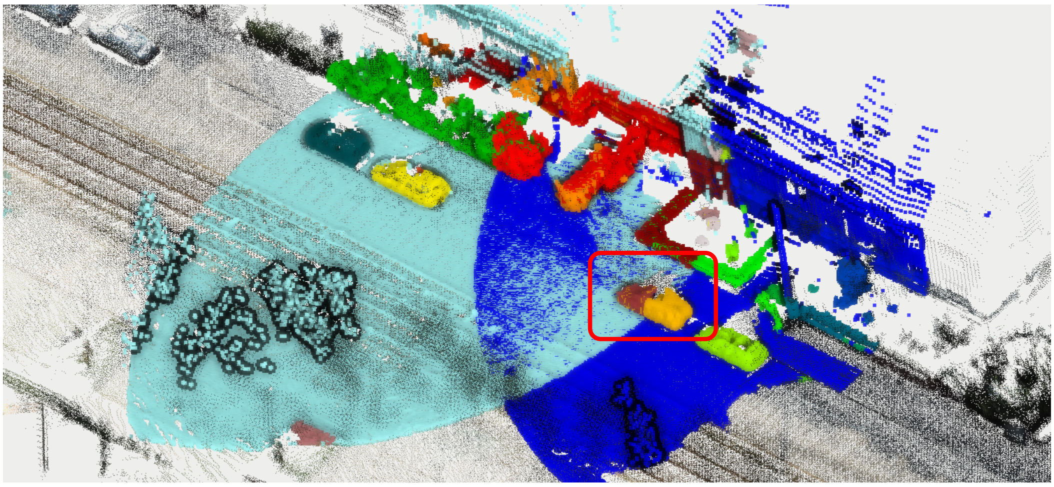


Figure 6. Two overlapping partitions generated by the proposed data augmentation approach. Each color represents a different segment. The same objects can appear in two different segments in two partitions, such as the car outlined by a red square.

4.5. Comparison Against the State of the Art

For each dataset, a comparative analysis of the two learning strategies, namely DA-supervised and CLOUDSPAM, is provided against current state-of-the-art and formerly highly ranked architectures. Both mIoU and class-wise IoU scores are presented for the validation set of the KITTI-360 dataset and the test sets of the Toronto-3D and Paris-Lille-3D datasets in Tables 3–5, respectively. For visual analysis purposes, images of the inference results of the CLOUDSPAM strategy on the test sets of all three datasets are provided in Figure 7. It can be assessed that scene coherence is achieved, even for the smallest data regimes. If some obvious errors still remain, most of them are corrected when a level of 10% labeled data is reached. This can be seen in the Toronto-3D dataset, where poles are correctly segmented, and in the KITTI-360 dataset, where ground segmentation mistakes disappear when 10% labeled data is available. Nevertheless, some errors still remain, especially for the Paris-Lille-3D dataset, such the faulty segmentation of a central car in the middle of the street. In the following paragraphs, the results obtained for each dataset are analyzed one after the other.

***KITTI-360*:** When utilizing 100% of the available labeled data, CLOUDSPAM achieved state-of-the-art performance for KITTI-360, surpassing the results of SPT by 0.1%. Furthermore, DA-supervised achieved a higher mIoU score and outperformed SPT by 0.6%. Even for a dataset as extensive as KITTI-360, where transformer architectures might be expected to excel due to their ability to extract more robust features, strategic statistical data selection can be equally effective. Examining class-wise IoU scores, a significant difference can be observed relative to SPT scores only for the traffic light and bicycle classes. For all other classes, CLOUDSPAM either achieved superior IoU scores or came close to matching them.

***Toronto-3D*:** Both DA-supervised and CLOUDSPAM reached achieved second position in the state-of-the-art ranking, just behind RandLA-Net, with mIoU scores of 67.9% and 71.8%, respectively. This illustrates that the proposed methodology enabled the use of self-supervised learning for a relatively small-scale dataset. However, both strategies failed to detect road marks, as depicted in Figure 7, row 2. This could be related to the network architecture that CLOUDSPAM uses. MinkUnet voxelizes the point clouds and creates a coarser representation of the point space. As such, the network is unable to pick up very fine geometrical details like road marks (less than 2.5 cm).

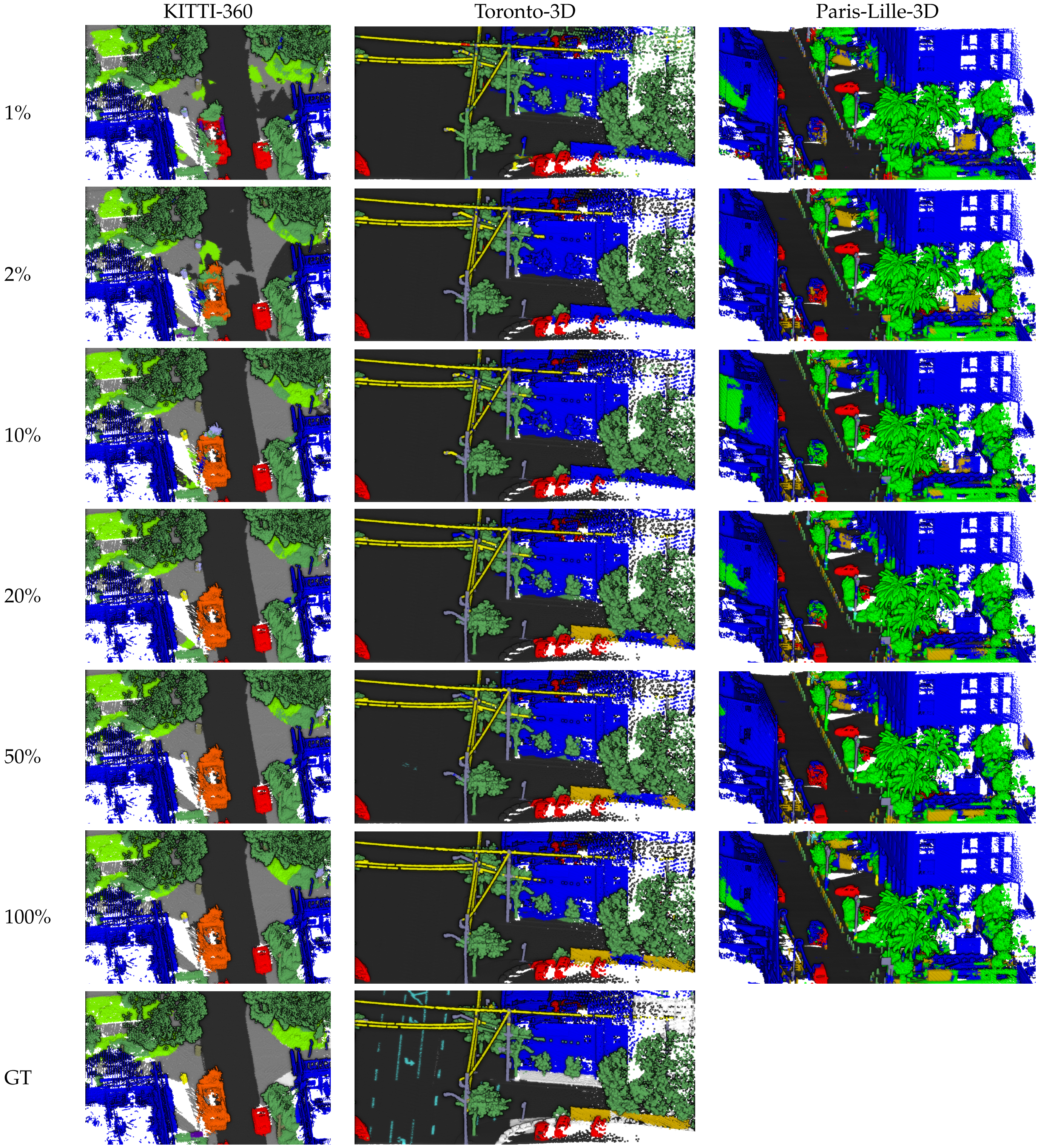


Figure 7. Inference results of the CLOUDSPAM strategy on the KITTI-360 (KIT-360), Toronto-3D (T3D) and Paris-Lille-3D (PL3D) test sets for every investigated data regime compared to the ground truth (GT). The ground truth of the Paris-Lille-3D test set was not provided by the authors.

Table 3. Comparison of mIoU (%) and class-wise IoU (%) scores for the validation set of KITTI-360. Results for MinkUnet come from the DeepViewAggregation paper [29].

Method	**mIoU**	**Road**	**Sidewalk**	**Building**	**Wall**	**Fence**	**Pole**	**Traffic lig.**	**Traffic sig.**	**Vegetation**	**Terrain**	**Person**	**Car**	**Truck**	**Motorcycle**	**Bicycle**
MinkUNet*	54.2	90.6	74.4	84.5	45.3	42.9	52.7	0.5	38.6	87.6	70.3	26.9	87.3	66.0	28.2	17.2
DeepViewAgg	57.8	93.5	77.5	89.3	53.5	47.1	55.6	18.0	44.5	**91.8**	71.8	40.2	87.8	30.8	39.6	**26.1**
SPT	63.5	93.3	79.3	**90.8**	**56.2**	45.7	52.8	**20.4**	51.4	89.8	73.6	61.6	95.1	79.0	**53.1**	10.9
DA-supervised	**64.1**	**95.6**	83.3	90.4	**56.2**	**50.2**	**60.9**	0.0	**53.7**	90.7	**75.7**	**73.4**	**96.4**	**82.5**	47.5	4.5
CLOUDSPAM	63.6	**95.6**	**83.4**	90.4	**56.2**	48.7	60.6	10.4	52.7	90.7	75.5	62.0	96.3	75.6	49.3	6.9

Table 4. Comparison of mIoU (%) and class-wise IoU (%) scores for the test set of Toronto-3D.

Method	**mIoU**	**Road**	**Road Mark**	**Natural**	**Building**	**Utility Line**	**Pole**	**Car**	**Fence**
PointNet++ [30]	56.5	91.4	7.6	89.8	74.0	68.6	59.5	54.0	7.5
PointNet++(MSG) [30]	53.1	90.7	0.0	86.7	75.8	56.2	60.9	44.5	10.2
DGCNN [31]	49.6	90.6	0.4	81.2	**93.9**	47.0	56.9	49.3	7.3
KPConv [32]	60.3	90.2	0.0	86.8	86.8	81.1	73.1	42.8	21.6
MS-PCNN [33]	58.0	91.2	3.5	90.5	77.3	62.3	68.5	53.6	17.1
TGNet [34]	58.3	91.4	10.6	91.0	76.9	68.3	66.2	54.1	8.2
MS-TGNet [18]	61.0	90.9	18.8	92.2	80.6	69.4	71.2	51.0	13.6
RandLA-Net [35]	**77.7**.	94.6	**42.6**	**96.9**	93.0	**86.5**	**78.1**	**92.8**	37.1
DA-supervised	69.3	94.9	0.0	94.9	90.0	84.4	73.8	89.7	26.5
CLOUDSPAM	71.8	**95.0**	0.0	95.7	90.5	85.7	77.1	91.7	**38.7**

Table 5. Comparison of mIoU (%) and class-wise IoU (%) scores for the test set of Paris-Lille-3D.

Method	**mIoU**	**Ground**	**Building**	**Pole**	**Bollard**	**Trash Can**	**Barrier**	**Pedestrian**	**Car**	**Nature**
RF_MSSF [36]	56.3	99.3	88.6	47.8	67.3	2.3	27.1	20.6	74.8	78.8
MS3_DVS [37]	66.9	99.0	94.8	52.4	38.1	36.0	49.3	52.6	91.3	88.6
HDGCN [38]	68.3	99.4	93.0	67.7	75.7	25.7	44.7	37.1	81.9	89.6
MS-RRFSegNet [39]	79.2	98.6	98.0	**79.7**	74.3	75.1	57.9	55.9	82.0	91.4
ConvPoint [40]	75.9	99.5	95.1	71.6	88.7	46.7	52.9	53.5	89.4	85.4
KPConv [32]	82.0	99.5	94.0	71.3	83.1	**78.7**	47.7	**78.2**	94.4	91.4
FKACon [41]	**82.7**	**99.6**	**98.1**	77.2	**91.1**	64.7	**66.5**	58.1	**95.6**	**93.9**
RandLA-Net [35]	78.5	99.5	97.0	71.0	86.7	50.5	65.5	49.1	95.3	91.7
DA-supervised	63.8	99.1	95.8	55.8	48.6	35.4	37.9	23.7	86.3	91.8
CLOUDSPAM	73.8	99.4	95.7	56.7	66.4	64.4	58.0	39.8	92.5	91.0

Paris-Lille-3D: This dataset stands out as the only one where neither DA-supervised nor CLOUDSPAM achieved state-of-the-art performance. The point density in this dataset significantly influences the efficiency of the semantic segmentation strategies. As observed by Mahmoudi Kouhi et al. [27], radius search pre-processing is more optimal for high-density point clouds than KNN search or random sampling. Table 2 reveals that the average partition radius for the Paris-Lille-3D dataset is 12 m, compared to 20 m and 25 m for the KITTI-360 and Toronto-3D datasets, respectively. The smaller partitions hindered training quality by reducing the receptive field. As shown in Figure 7, the car was incorrectly segmented in the Paris-Lille-3D dataset. This mistake can be linked to the small receptive field, as the middle of the road is the densest part of the scans. Similarly, RandLA-Net, which uses random sampling, outperformed KPConv by more than 17% in mIoU on the

Toronto-3D dataset. However, KPConv, utilizing radius search, nearly matched the state-of-the-art performance on the Paris-Lille-3D dataset and surpassed RandLA-Net by 3.5%.

These comparative results highlight the competitiveness of CLOUDSPAM with state-of-the-art approaches, especially in limited-label scenarios, as well as the superior performance of DA-supervised compared to the state of the art on the KITTI-360 dataset. Two limitations emerge—one from the KNN search, restricting the capabilities of the learning strategies in areas with very high point density (more than 1000 points/m^2), and the other from voxelization, which creates a coarser representation of the point space and renders the networks unable to capture very fine geometrical details such as road marks (less than 2.5 cm). To address these limitations, future work could explore alternatives to KNN search to handle areas with extremely high point density, such as radius search techniques, which could be adjusted to the neighborhood size based on point density. This would allow the network to better capture local context. Additionally, to overcome the limitations of voxelization and improve the network's ability to capture fine geometrical details, networks such as transformers could be implemented as the backbone. These approaches would preserve finer details, such as road marks, while maintaining computational efficiency.

5. Conclusions

In conclusion, this paper addresses the challenges of semantic segmentation for mobile mapping LiDAR point cloud datasets. Through the implementation of innovative methodologies and adaptations of existing techniques, we have demonstrated significant advancements in self-supervised pre-training and label-efficient learning strategies. The proposed data augmentation approach, leveraging merged heterogeneous datasets and contrastive self-supervised pre-training (CLOUDSPAM), shows notable effectiveness in enhancing semantic segmentation performance across various datasets. By augmenting segment quantity and purity, we successfully unlocked the potential for contrastive learning on mobile mapping datasets, even in scenarios with limited labeled data. Furthermore, the experiments showcased the importance of careful data selection and preparation in deep learning-based point cloud segmentation. Thanks to such a data preparation approach, we were able to merge heterogeneous mobile mapping datasets to enhance the versatility and generalizability of the networks. This led to the achievement of performance on par with that of state-of-the-art transformer architectures. While adapted contrastive learning demonstrated competitive performance across different datasets, there remain avenues for future exploration. Deepening our understanding of pre-training initialization and investigating the effectiveness of deeper networks could further enhance segmentation performance, particularly for datasets with uneven densities and characteristics. Overall, this study contributes valuable insights and methodologies to the field of 3D LiDAR point cloud segmentation, paving the way for improved understanding and utilization of large-scale outdoor datasets in various applications, such as urban planning and environmental monitoring.

Author Contributions: Conceptualization, R.M.K., O.S., S.D. and P.G.; methodology, R.M.K., S.D. and P.G.; software, R.M.K.; validation, R.M.K., O.S., S.D. and P.G.; formal analysis, R.M.K., O.S., S.D. and P.G.; investigation, R.M.K.; resources, R.M.K., S.D. and P.G.; data curation, R.M.K.; writing—original draft preparation, R.M.K. and O.S.; writing—review and editing, R.M.K., S.D. and P.G.; visualization, R.M.K. and O.S.; supervision, S.D. and P.G.; project administration, S.D.; funding acquisition, S.D. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Natural Sciences and Engineering Research Council (NSERC) of Canada (grant number RGPIN-2018-04046).

Data Availability Statement: The KITTI-360 dataset's official website is https://www.cvlibs.net/datasets/kitti-360/ (accessed on 13 September 2024). The Toronto-3D dataset's official website is https://github.com/WeikaiTan/Toronto-3D?tab=readme-ov-file (accessed on 13 September 2024). The Paris-Lille-3D dataset's official website is https://npm3d.fr/paris-lille-3d (accessed on 13 September 2024).

Acknowledgments: The authors gratefully acknowledge the support of the NSERC, the Research Center pf Geospatial Data and Intelligence of Université Laval, the Institute of Intelligence and Data of Université Laval, and NVIDIA for providing a high-performance GPU for our research project through their hardware acceleration program, as well as the Digital Alliance of Canada for providing us the opportunity to run our experiments on their clusters.

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

References

1. Griffiths, D.; Boehm, J. A Review on Deep Learning Techniques for 3D Sensed Data Classification. *Remote Sens.* **2019**, *11*, 1499. [CrossRef]
2. Guo, Y.; Wang, H.; Hu, Q.; Liu, H.; Liu, L.; Bennamoun, M. Deep learning for 3d point clouds: A survey. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *43*, 4338–4364. [CrossRef] [PubMed]
3. Xie, Y.; Tian, J.; Zhu, X.X. Linking points with labels in 3D: A review of point cloud semantic segmentation. *IEEE Geosci. Remote Sens. Mag.* **2020**, *8*, 38–59. [CrossRef]
4. Xiao, A.; Huang, J.; Guan, D.; Zhang, X.; Lu, S.; Shao, L. Unsupervised point cloud representation learning with deep neural networks: A survey. *IEEE Trans. Pattern Anal. Mach. Intell.* **2023**, *45*, 11321–11339. [CrossRef]
5. Arora, S.; Khandeparkar, H.; Khodak, M.; Plevrakis, O.; Saunshi, N. A theoretical analysis of contrastive unsupervised representation learning. *arXiv* **2019**, arXiv:1902.09229.
6. Xiao, A.; Zhang, X.; Shao, L.; Lu, S. A Survey of Label-Efficient Deep Learning for 3D Point Clouds. *arXiv* **2023**, arXiv:2305.19812. [CrossRef]
7. 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/CVF International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 27 October–2 November 2019.
8. Liao, Y.; Xie, J.; Geiger, A. KITTI-360: A Novel Dataset and Benchmarks for Urban Scene Understanding in 2D and 3D. *Pattern Anal. Mach. Intell. (PAMI)* **2022**, *45*, 3292–3310. [CrossRef] [PubMed]
9. Gui, J.; Chen, T.; Cao, Q.; Sun, Z.; Luo, H.; Tao, D. A Survey of Self-Supervised Learning from Multiple Perspectives: Algorithms, Theory, Applications and Future Trends. *arXiv* **2023**, arXiv:2301.05712.
10. Hou, J.; Graham, B.; Nießner, M.; Xie, S. Exploring Data-Efficient 3D Scene Understanding with Contrastive Scene Contexts. *arXiv* **2021**, arXiv:2012.09165.
11. He, K.; Fan, H.; Wu, Y.; Xie, S.; Girshick, R. Momentum Contrast for Unsupervised Visual Representation Learning. *arXiv* **2020**, arXiv:1911.05722.
12. Xie, S.; Gu, J.; Guo, D.; Qi, C.R.; Guibas, L.J.; Litany, O. PointContrast: Unsupervised Pre-training for 3D Point Cloud Understanding. In Proceedings of the IEEE/CVF Europian Conference Computer Vision (ECCV), Glasgow, UK, 23–28 August 2020.
13. Nunes, L.; Marcuzzi, R.; Chen, X.; Behley, J.; Stachniss, C. SegContrast: 3D Point Cloud Feature Representation Learning through Self-supervised Segment Discrimination. *IEEE Robot. Autom. Lett. (RA-L)* **2022**, *7*, 2116–2123. [CrossRef]
14. Grill, J.B.; Strub, F.; Altché, F.; Tallec, C.; Richemond, P.; Buchatskaya, E.; Doersch, C.; Avila Pires, B.; Guo, Z.; Gheshlaghi Azar, M.; et al. Bootstrap your own latent-a new approach to self-supervised learning. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 21271–21284.
15. Jiang, L.; Shi, S.; Tian, Z.; Lai, X.; Liu, S.; Fu, C.W.; Jia, J. Guided point contrastive learning for semi-supervised point cloud semantic segmentation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 6423–6432.
16. Li, L.; Shum, H.P.; Breckon, T.P. Less is more: Reducing task and model complexity for 3d point cloud semantic segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 9361–9371.
17. Fei, B.; Yang, W.; Liu, L.; Luo, T.; Zhang, R.; Li, Y.; He, Y. Self-supervised Learning for Pre-Training 3D Point Clouds: A Survey. *arXiv* **2023**, arXiv:2305.04691.
18. 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. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, Seattle, WA, USA, 14–19 June 2020; pp. 202–203.
19. 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]
20. van den Oord, A.; Li, Y.; Vinyals, O. Representation Learning with Contrastive Predictive Coding. *arXiv* **2018**, arXiv:1807.03748.
21. Belongie, S.; Malik, J.; Puzicha, J. Shape Matching and Object Recognition Using Shape Contexts. *IEEE Trans. Pattern Anal. Mach. Intell.* **2002**, *24*, 509–522. [CrossRef]
22. Körtgen, M.; Park, G.J.; Novotni, M.; Klein, R. 3D Shape Matching with 3D Shape Contexts. In Proceedigs of the 7th Central European Seminar on Computer Graphics, Vienna, Austria, 7–9 May 2003.
23. Xie, S.; Liu, S.; Chen, Z.; Tu, Z. Attentional ShapeContextNet for Point Cloud Recognition. In Proceedigs of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018. [CrossRef]

24. Fischler, M.A.; Bolles, R.C. Random Sample Consensus: A Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography. *Assoc. Comput. Mach. Commun. ACM* **1981**, *24*, 381–395. [CrossRef]
25. Ester, M.; Kriegel, H.P.; Sander, J.; Xu, X. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. In Proceedings of the Second International Conference on Knowledge Discovery and Data Mining, Portland, OR, USA, 2–4 August 1996; pp. 226–231.
26. Narksri, P.; Takeuchi, E.; Ninomiya, Y.; Morales, Y.; Akai, N.; Kawaguchi, N. A Slope-robust Cascaded Ground Segmentation in 3D Point Cloud for Autonomous Vehicles. In Proceedings of the 2018 21st International Conference on Intelligent Transportation Systems (ITSC), Maui, HI, USA, 4–7 November 2018; pp. 497–504. [CrossRef]
27. Mahmoudi Kouhi, R.; Daniel, S.; Giguère, P. Data Preparation Impact on Semantic Segmentation of 3D Mobile LiDAR Point Clouds Using Deep Neural Networks. *Remote Sens.* **2023**, *15*, 74. [CrossRef]
28. Choy, C.; Gwak, J.; Savarese, S. 4d Spatio-Temporal Convnets: Minkowski Convolutional Neural Networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 3075–3084.
29. Robert, D.; Vallet, B.; Landrieu, L. Learning Multi-View Aggregation In the Wild for Large-Scale 3D Semantic Segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LO, USA, 19–20 June 2022; pp. 5575–5584.
30. 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 31st International Conference on Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; Volume 30.
31. Wang, Y.; Sun, Y.; Liu, Z.; Sarma, S.E.; Bronstein, M.M.; Solomon, J.M. Dynamic graph cnn for learning on point clouds. *ACM Trans. Graph. (TOG)* **2019**, *38*, 1–12. [CrossRef]
32. Thomas, H.; Qi, C.R.; Deschaud, J.E.; Marcotegui, B.; Goulette, F.; Guibas, L.J. Kpconv: Flexible and deformable convolution for point clouds. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 Ocotber–2 November 2019; pp. 6411–6420.
33. Ma, L.; Li, Y.; Li, J.; Tan, W.; Yu, Y.; Chapman, M.A. Multi-scale point-wise convolutional neural networks for 3D object segmentation from LiDAR point clouds in large-scale environments. *IEEE Trans. Intell. Transp. Syst.* **2019**, *22*, 821–836. [CrossRef]
34. Li, Y.; Ma, L.; Zhong, Z.; Cao, D.; Li, J. TGNet: Geometric graph CNN on 3-D point cloud segmentation. *IEEE Trans. Geosci. Remote Sens.* **2019**, *58*, 3588–3600. [CrossRef]
35. Hu, Q.; Yang, B.; Xie, L.; Rosa, S.; Guo, Y.; Wang, Z.; Trigoni, N.; Markham, A. Learning semantic segmentation of large-scale point clouds with random sampling. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, *44*, 8338–8354. [CrossRef] [PubMed]
36. Thomas, H.; Goulette, F.; Deschaud, J.E.; Marcotegui, B.; LeGall, Y. Semantic classification of 3D point clouds with multiscale spherical neighborhoods. In Proceedings of the 2018 International conference on 3D vision (3DV), Verona, Italy, 5–8 September 2018; pp. 390–398.
37. Roynard, X.; Deschaud, J.E.; Goulette, F. Classification of point cloud for road scene understanding with multiscale voxel deep network. In Proceedings of the 10th Workshop on Planning, Perceptionand Navigation for Intelligent Vehicules PPNIV'2018, Madrid, Spain, 1–5 October 2018.
38. Liang, Z.; Yang, M.; Deng, L.; Wang, C.; Wang, B. Hierarchical depthwise graph convolutional neural network for 3D semantic segmentation of point clouds. In Proceedings of the 2019 International Conference on Robotics and Automation (ICRA), Montreal, QC, Canada, 20–24 May 2019; pp. 8152–8158.
39. Luo, H.; Chen, C.; Fang, L.; Khoshelham, K.; Shen, G. MS-RRFSegNet: Multiscale regional relation feature segmentation network for semantic segmentation of urban scene point clouds. *IEEE Trans. Geosci. Remote Sens.* **2020**, *58*, 8301–8315. [CrossRef]
40. Boulch, A.; Le Saux, B.; Audebert, N. Unstructured point cloud semantic labeling using deep segmentation networks. *3dor@ Eurographics* **2017**, *3*, 1–8.
41. Boulch, A.; Puy, G.; Marlet, R. FKAConv: Feature-kernel alignment for point cloud convolution. In Proceedings of the Asian Conference on Computer Vision, Kyoto, Japan, 30 November–4 December 2020; pp. 381–399.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.