PCRMLP : A Two-Stage Network for Point Cloud Registration in Urban Scenes

Jingyang Liu , Yucheng Xu , Lu Zhou , Lei Sun *

Posted Date: 23 April 2023

doi: 10.20944/preprints202304.0804.v1

Keywords: registration; point clouds; urban scene; deep learning

Article

PCRMLP: A Two-Stage Network for Point Cloud Registration in Urban Scenes

Jingyang Liu [1], Yucheng Xu [2], Lu Zhou [1] and Lei Sun [1,*]

[1] College of Artificial Intelligence, Nankai University, Tianjin, P.R.China; 2120210419@mail.nankai.edu.cn
[2] School of informatics, University of Edinburgh, Edinburgh EH89YL, Scotland; s2038119@ed.ac.uk
* Correspondence: sunl@nankai.edu.cn

Abstract: Urban scene point cloud pose significant challenges for registration due to its large data volume, similar scenarios and dynamic objects. In this paper, we propose PCRMLP, a model for urban scene point cloud registration that achieves comparable registration performance to prior learning-based methods. Compared to previous works which focus on extracting features and estimating correspondence, the model estimates the transformation implicitly from concrete instances. An instance-level urban scene representation method is introduced to extract instance descriptors via semantic segmentation and DBSCAN, which enable the model to obtain robust instance features, filter dynamic objects and estimate transformation in a more logical manner. Then a lightweight network consisting of MLPs is employed to obtain transformation in an encoder-decoder manner. We validate the approach on KITTI dataset. Experimental results demonstrate that PCRMLP can obtain a satisfactory coarse transformation from instance descriptors just in 0.0028s. With a subsequent ICP refinement module, the proposed method achieves higher registration accuracy and computational efficiency than prior learning-based works.

Keywords: registration; point clouds; urban scene; deep learning

1. Introduction

Point cloud registration is the task of estimating the rigid transformation that aligns a pair of overlapping point clouds. It is important for autonomous driving [1,2], pose estimation [3,4], 3D reconstruction [5,6], simultaneous localization and mapping(SLAM) [7,8]. In the field of autonomous driving, urban point clouds have the characteristics of sparsity, multiple dynamic objects, and susceptibility to environmental influences during collection, which make feature extraction and registration challenging.

The most common registration method is Iterative Closest Point(ICP) [9] and some relative algorithms [10,11], which solve the problem by alternating between finding the closest points and computing the optimal transformation. However, ICP relies on initial values, usually converges to local optima when dealing with non-convexity problems. Recently, Learning-based methods have become more and more popular. Methods based on deep learning can extract more robust features and correspondences [12–14] or solve transformations in an end-to-end manner [15,16]. However, when dealing with urban point clouds, prior works need to downsample the initial data, which is time consuming and makes the algorithm sensitive to sampling density and susceptible to the influence of point cloud scale. Besides, the performance suffers a lot from geometrically-similar scenarios [17] and dynamic objects. Registration on urban scene point clouds remains a challenge.

Considering that the relative position of static instances in the surrounding environment remains constant during vehicle movement, which makes instance-level registration a more logical approach. Moreover, instance features are more robust and computationally efficient. Prior works usually extract features or superpoints [18] with DNNs, while it is more intuitive to obtain the concrete instances directly. We propose to utilize abundant semantic features in urban scenes by mapping point clouds to a high-dimensional feature space, namely instance level. The instance-level scene representation provides instance descriptors for registration. The proposed network first uses an

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 23 April 2023 doi:10.20944/preprints202304.0804.v1

efficient point cloud semantic segmentation model to extract point-wise semantic label, then obtains instance bounding box information through the DBSCAN [19] clustering algorithm. Unlike previous feature-pair-based methods, the instance descriptors of input point clouds are fed into the network to obtain the transformation.

Overall, the main contribution of this paper is proposing an instance-level representation method for urban point clouds and an efficient registration network with simple MLP s, named PCRMLP, based on it. the secondary contribution is demonstrating that the proposed method can provide more accurate coarse registration results in urban scenes, providing a satisfactory initial value for ICP. Moreover, it achieves average 0.0028 s per frame in registration stage.

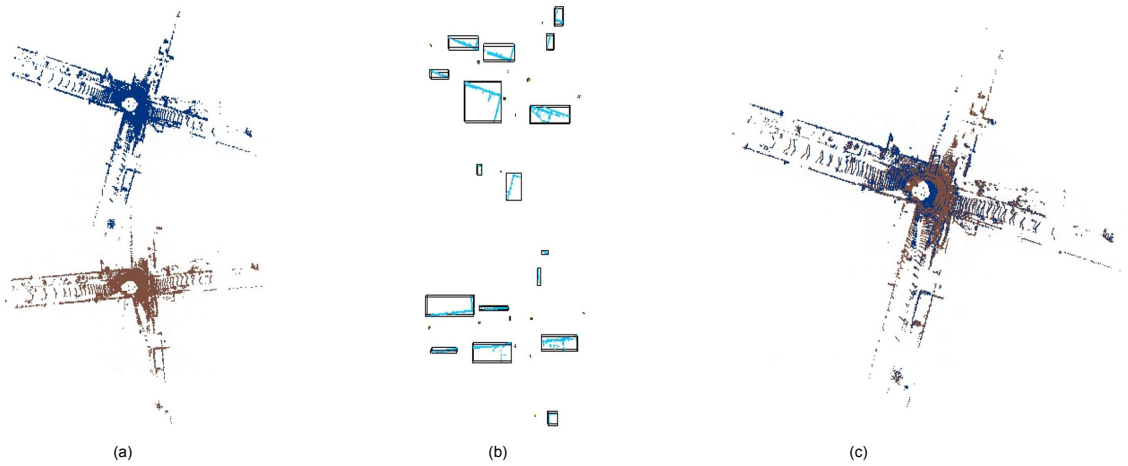

Figure 1. The process of PCRMLP. (**a**) raw point clouds (**b**) Instance descriptors based on semantic segmentation and DBSCAN clustering (**c**) point clouds registered by MLP network.

2. Related Works

In this section, the related 3D point clouds processing methods and point cloud registration methods will be briefly introduced.

2.1. 3D Point Clouds Processing

Existing point cloud processing algorithms can be roughly divided into point-based methods and voxel-based methods according to the input data structure.

Point-based methods take advantage of some inherent symmetric functions(e.g., shared MLP) to handle unordered 3D point clouds. PointNet [20] is one of the pioneer attempt of point-based methods. Vanilla PointNet shows remarkable efficiency in processing irregular 3D point clouds. However it can not aggregate local feature. Subsequent studies focus on modeling local contexts. [21–23] hierarchically stack PointNet module and apply local neighborhood query module to extract local context.

Voxel-based methods utilize volumes to represent point clouds. [24] initially introduces 3D CNN to process voxelized 3D point clouds. VoxelNet [25] discards empty voxels to generate a sparse tensor to reduce the memory usage and computation cost significantly. In most cases, the higher the voxel resolution is, the better the performance is, while the more computation is required.

Several following researches take both the advantages of point-based methods and voxel-based methods. Point-Voxel Convolution [26] is composed of point-branch, which uses PointNet to extract point-wise fine-grained features and voxel-branch that obtain coarse-grained features using sparse 3D CNN, then the fused features will be applied to different tasks.

2.2. Optimization-Based Registration

The most well-known registration method Iterative Closest Point (ICP) conduct two stages iteratively: correspondence obtaining and transformation estimation by solving a least squares

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 23 April 2023 doi:10.20944/preprints202304.0804.v1

equation. Implementations of ICP [10,27] have been proposed to accelerate or improve accuracy by introducing extra features. Optimization-based methods are mathematically rigorous and can recover closed-form solution.

2.3. Learning-Based Registration

Learning-based method introduces DNNs to extract more robust local or global features. These works are mainly divided into two kinds: learning based feature extraction and matching method and end-to-end method. The first one mainly uses DNNs as a feature extractor to extract local features and corresponding relationships in point cloud scenes for following estimation via classical methods (e.g., RANSAC [28]). FCGF [12] proposes a fully convolutional geometric feature network, which efficiently learns more compact geometric features. DCP [29] makes a hard assumption about the distribution of points and corresponding points and is not suitable for partially overlapping scenes. End-to-end methods use end-to-end networks to solve the registration problem. Two frame point clouds are input into the network, and the transformed predicted value is output directly. In this process, the traditional optimization ideas are integrated into the training process of the network [30], and the loss function is used as the constraint solution. Thanks to the strong neighborhood coding ability of PointNet [20] for point clouds, methods such as deepVCP [16] and PCRNet [15] use PointNet to extract point cloud features and estimate pose transformations by MLPs. In [31], reliable line features from poles and buildings are extracted to perform registration. Learning-based methods can extract more robust features and more accurate corresponding relationships from the scene for transformation estimation. However, in urban scene point cloud, these methods often require downsampling to reduce computation, which will lead to loss of information and degradation of model performance.

3. Method

In this section, the proposed two-stage registration framework will be presented, which estimates the transformation between a pair of point cloud from urban scene from raw and irregular point clouds. The whole structure of PCRMLP framework shown in Figure 2, which is composed of descriptor generation stage and registration stage with MLPs.

3.1. Problem Statement

Given two point clouds $P = \{p_i \in R^3 \mid i = 1, \cdots, m\}$ and $Q = \{q_i \in R^3 \mid i = 1, \cdots, n\}$,the goal is to recover a rigid transformation $T = \{R, t\}$, the rotation matrix $R \in SO(3)$ and the translation vector $t \in R^3$, which aligns P and Q. The transformation can be estimated by solving

$$\arg\min_{R,t} \sum_{(p_i,q_j)\in C} \left\| Rp_i + t - q_j \right\| \tag{1}$$

where p_i, q_j are the corresponding points in source and target point cloud.

3.2. Instance Descriptor Generation

The stage 1 of our previous work PointTrans [32] is introduced to generate instance-wise masks. Point-Voxel Convolution [26] is adopted to extract features from raw points. Then a 3D U-Net [33] is leveraged as the semantic segmentation branch due to its strong capacity of learning and segmentation on voxel-based representation. Following is a segmentation head uses simple fully connected to project the features to semantic labels. Given the semantic label of each point, the object points can directly be selected out according to its label. 3D point clouds show apparent separability in original 3D space because of the natural depth information, which means individual point cloud object instances can be segmented from the object points using simple cluster methods in 3D space. However, number of object instances various in different scene, in short, the exact number of object instances of each point cloud frame is not known. So those cluster methods requiring known instance number can not satisfy

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 23 April 2023 doi:10.20944/preprints202304.0804.v1

out requirement. Therefore, We consider applying a density-based cluster method, DBSCAN [19], to address this problem. As it can be seen from the Figure 2, individual instances can be segmented by DBSCAN directly [32].

Only static instances of buildings, poles, and traffic lights are kept for following registration and generate a bounding box for each instance just based on filtered semantic points. Noticing that the relationship among instances is more important for vehicle to recognize the scene, for each descriptor, a vector that contains coordinates, box size and semantic label is generated for neural network to implicitly learn associations between instances. The descriptors are defined as $F = \{coor, l, w, h, L\}$, $coor \in R^3$ is the coordinate of the center point of the generated bounding box, l, w, h are the length, width, height of the box and $L \in R^3$ is the semantic label of the instance which is embedded to an one-hot vector.

In conclusion, the 3D semantic segmentation method is adopted in the proposed network as an accurate region proposal module. We ingeniously take advantages of the separability of 3D point clouds and combination 3D semantic segmentation method with a density-based cluster method to directly generate masks for every object instance. However, this also means that the detection result of our algorithm will, to some extent, rely on the segmentation result of the semantic segmentation stage [32].

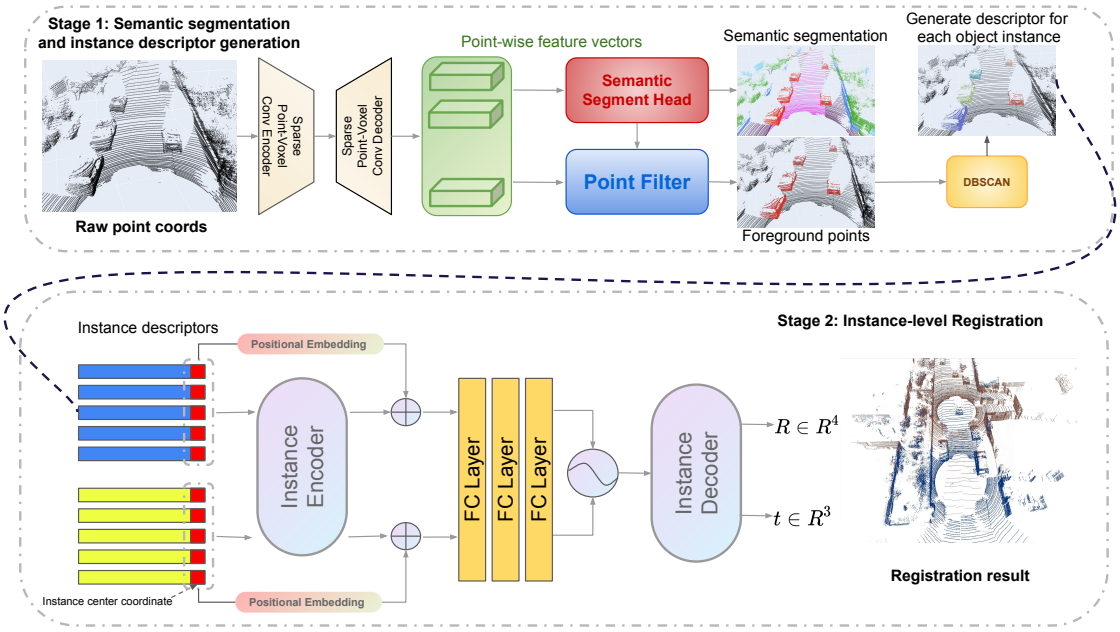

Figure 2. PCRMLP Architecture: the model contains two stages: (1) Semantic segmentation and instance descriptor generation. We utilize our proposed object mask generator [32], which is consist of PVConv feature extract module, 3D U-Net segmentation module, a simple task head and DBSCAN clustering, to extract instance-wise masks of specified semantic labels. Then the axis aligned instance bounding boxes are obtained just via open3D (2) Instance-level registration. We use shared-MLPs as encoder-decoder manner, estimate the transformation from input instance descriptors.

3.3. Instance-Level Registration

Consider that instance-level representation already contains rich geometric and semantic information, simple MLPs are applied to estimate the transformation from two instance-level frames. Besides, the instance center coordinates are embedded with MLPs as positional embedding and feed the embeddings to the instance feature vectors. The positional embedding(PE) module can be defined as

$$PE(Ins_i) = MLP(x_i, y_i, z_i) \tag{2}$$

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 23 April 2023 doi:10.20944/preprints202304.0804.v1

Similar to Siamese architecture, shared-MLP is introduced as encoder to map the descriptors to high dimensional space. After concatenating two feature tensors, a similar MLP decoder maps the features back to the output estimation. Figure 3 shows the decoder module. To satisfy the quaternion limitation, a normalization operation is applied in the rotation prediction branch. The output rotation is represented by quaternion since it is continuous.

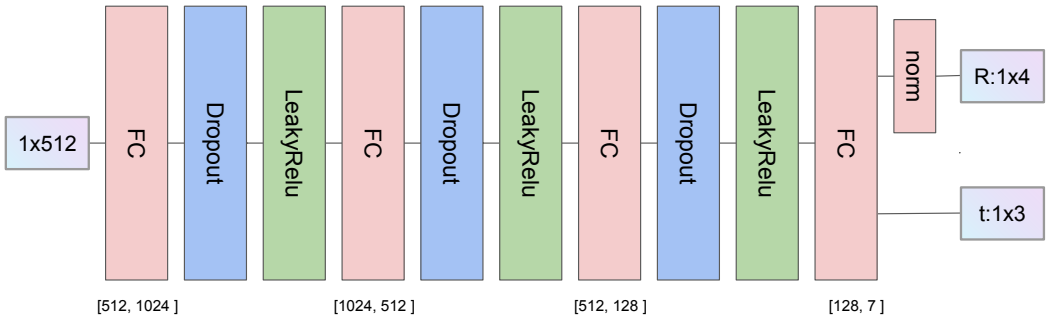

Figure 3. Decoder Architecture: we use 4 full connection layers to map high-dimension features to the output transformation. The rotation is represented by quaternions which is limited by the norm layer.

3.4. Loss Function

As the point-wise semantic labels are only required for point filter, the semantic segmentation stage and the subsequent registration stage are trained separately.

Semantic segmentation loss Following the prior point cloud semantic segmentation algorithm, we use a simple cross entropy loss function for this stage.

$$L_{sem_cls} = \sum CE(\hat{sem}_{cls}, sem_{cls}) \tag{3}$$

Registration loss The loss function in the registration stage restricts the predicted transformation to be as close to the ground truth. We sample n points from the source point cloud and calculated the average distance between the virtual points projected by the estimate value and the ground truth. The loss function can be defined as

$$L_{reg} = \frac{1}{n} \sum_{x \in P_s} \left\| T(x) - \hat{T}(x) \right\| \tag{4}$$

where P_s is the source point cloud, x is the sampled source point, T and $\hat{T}$ are the ground truth and estimation of the transformation.

4. Experiments and Results

In this section the implement details of PCRMLP will be introduced first. Then evaluation of PCRMLP on Semantic KITTI dataset [34] will be illustrated. Afterwards, we demonstrate a comparison with other point cloud registration methods on accuracy and computational efficiency.

4.1. Dataset and Training

The proposed network is trained and evaluated on the dataset generated from Semantic Kitti dataset, which contains point cloud data collected by Velodyne HDL64 LiDAR, ground truth poses provided by GPS, and point-wise semantic labels. In the KITTI point cloud dataset, only four sequences (0, 5, 7, 8) are collected in urban road scenes. For each frame in these sequences, we take every third frame as its corresponding frame, with a maximum interval of 30 frames, resulting in 111,060 pairs of point clouds. We divided all point cloud pairs into a training set of 100 K pairs, a validation set of 1150 pairs, and a test set of 9910 pairs.

For descriptor generation stage, the raw points are fed to the model. In stage 1, we follow the structure of [33], but replace conventional convolution operations with sparse point-voxel

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 23 April 2023 doi:10.20944/preprints202304.0804.v1

convolution [26]. Then, a semantic segmentation head is adopted to project features to point-wise semantic labels. Next, the static object points are filtered by selecting the specified semantic labels. The DBSCAN algorithm does not require the exact number of object instances [32]. Different parameters of DBSCAN cluster are set for different classes of object instances. For building instance, *eps*, the maximum distance between two points, is set as 2.3, and the minimum number of points of each neighborhood is set as 80. For pole instance, we set *eps* as 2, and set the minimum number of points of each neighborhood as 1. For traffic sign instance, the parameters are 3 and 1.

For registration stage, the feature dimension is set as 256, the encoder and the positional embedding module MLPs are [512, 256], the decoder MLPs are [1024, 512, 128, 7]. LeakyReLU is used as the activation function.

The models are trained and evaluated on a single RTX-TiTAN GPU and Intel Xeon Gold 5218 CPU. During the training period, random rotation and translation are applied to each pair of point clouds as data augmentation. We use the Adam optimizer with an initial learning rate of 1e-3, and the learning rate begin to exponentially decay with gamma of 0.99 after a warm-up period of 20 epochs.

4.2. Evaluation Metrics

The metrics of [14] is used to measure the performance of the proposed method on the test split. The formulas are $TE = \left\|\hat{t} - t_{gt}\right\|$, and $RE = arccos[Tr(R_{gt}^{T}\hat{R} - 1)/2]$ for translation error(TE) and rotation error(RE). In addition, the recall is calculated according to the pre-set threshold, which is the success rate of registration.

4.3. Comparison with Existing Methods

Models are evaluated on the test split. The performance of the proposed model is compared to other methods in Table 1. We compare PCRMLP with classical ICP [9], RANSAC [28] and learning-based FCGF [12], PCAM [35]. Figure 4 shows the coarse registration result from PCRMLP and fine registration result from PCRMLP combined with ICP. The proposed method with ICP outperforms on mean RE, mean TE and Recall. As discussed in [17], we notice that, as Figure 5 shows, FCGF, PCAM tend to degenerate when the scene contains lots of geometrically-similar objects (e.g., cars, buildings), while the proposed method performs better in urban scenes. PCRMLP can provide a satisfactory coarse registration initial value for ICP.

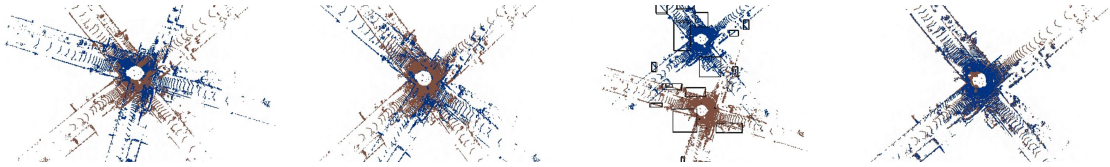

Figure 4. Qualitative visualization on test split. Left to right: a pair of point clouds, ICP result, PCRMLP stage.1 result, PCRMLP registration result.

Table 1. Performance comparison of rigid registration with previous methods on the *test* split set. Recall is defined as RE < 5°.

Method	RE(deg)↓		TE(m)↓		Recall↑
	Mean	Median	Mean	Median	
ICP [9]	7.55	1.67	10.31	8.29	71.50%
RANSAC [28]	6.65	1.13	2.85	0.24	83.08%
FCGF [12]	7.88	1.55	5.97	1.56	90.35%
PCAM [35]	2.81	**0.29**	3.68	**0.14**	92.08%
PCRMLP(ours)	3.42	1.56	3.56	2.77	81.82%
PCRMLP+ICP(ours)	**2.01**	0.79	**1.58**	1.03	**93.24%**

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 23 April 2023 doi:10.20944/preprints202304.0804.v1

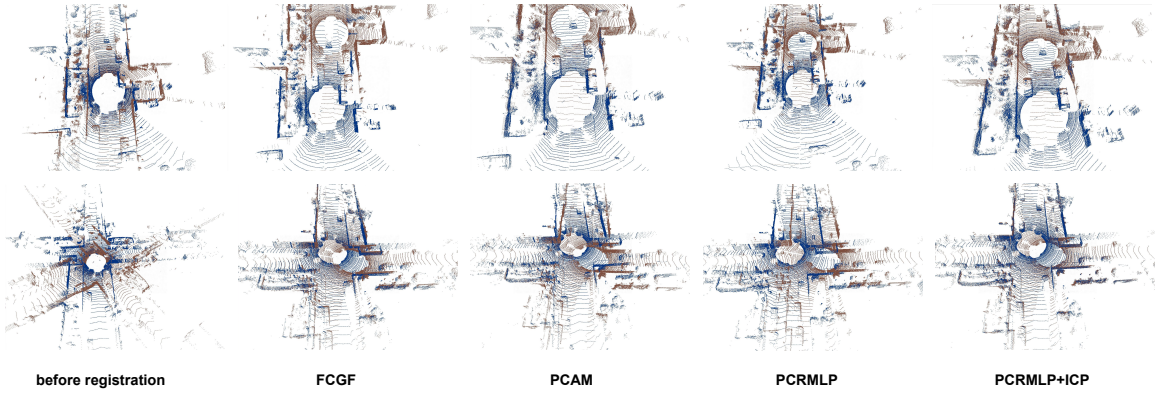

Figure 5. Failure cases of FCGF (**above**) and PCAM (**bottom**).

4.4. Run-Time Analysis

The running time of PCRMLP and other methods are tested on the point cloud pairs containing 35 K points. Table 2 shows that the proposed method achieves better computational efficiency. For FCGF, the running time of feature extraction (feat) module and registration (reg) module are tested separately, It is remarkable that the registration stage cost only 0.0028 s per pair of point clouds since we disaggregate mass point cloud data into a small amount of instance data.

Table 2. Running time of ICP, RANSAC, FCGF, PCAM and PCRMLP on point cloud pairs containing about 35K points. We also tested the running time of the stage.2 of PCRMLP, which indicates the potential of the introduced instance-level representation method.

	ICP	RANSAC	PCAM	FCGF		Ours		
				Feat	Reg	PCRMLP	PCRMLP+ICP	Stage 2
time (s)	10.31	7.64	**0.42**	0.05	12.05	0.65	4.01	**0.0028**

5. Discussion

In this section, we will analyze the advantages and disadvantages of the introduced method in conjunction with prior works and discuss about future research direction. Most previous learning-based registration methods focus on extracting local features and correspondences to constrain the rigid transformation, while it is more logical to recognize the scene by the concrete instances such as buildings, traffic signs and their relative relationship. Besides, when generalize the algorithm to urban scenes, these methods are usually limited by repeated scenarios and dynamic objects. Therefore, we propose the instance-level urban scene representation method to provide a novelty scene recognization paradigm for autonomous vehicles. This process also effectively reduces the data volume from tens of thousands points to dozens of instances. Then we design a simple registration network with MLPs to implicitly extract the relationship between instances. The proposed method can estimate a coarse registration from the instance-level scene with just 0.0028 s.

However, as the rotation increases, the performance of PCRMLP decreases, which means more iterations for following fine registration algorithm. We think there are two factors: (1) We only roughly estimate the bounding box of the instances based on the segmented points, which introduces error. (2) The rotation invariance of MLPs is poor, so that the network is hard to predict a large rotation angle. Another disadvantage is the generalization. When the trained model is applied to our own data of city point clouds, the performance decreases. In spite of the degraded performance, the result is a valid initial value for ICP.

Noticing the strong ability of reducing data volume and generating robust scene representation, we plan to apply the proposed method to the SLAM system for global localization and relocalization.

6. Conclusions

In this work, we have proposed a two-stage urban scene point cloud registration network PCRMLP. In the first stage, instance descriptors are generated by semantic segmentation and DBSCAN clustering. In the second stage, simple shared-MLPs are introduced to realize coarse registration based on instance-level representation. Experiment results show that the proposed algorithm achieves satisfactory and real-time performance on urban point clouds. In the future, we will try to apply PCRMLP to lidar mapping and localization.

Author Contributions: Conceptualization, J.L., Y.X., L.Z. and L.S.; methodology, J.L., Y.X. and L.S.; software, J.L.; validation, J.L.; formal analysis, J.L.; investigation, J.L.; resources, J.L.; data curation, J.L. and Y.X.; writing—original draft preparation, J.L.; writing—review and editing, J.L., Y.X., L.Z. and L.S.; visualization, J.L.; supervision, L.S.; project administration, L.Z. and L.S.; funding acquisition, L.S. and L.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China (No. 62173192) and the Shenzhen Natural Science Foundation (No. JCYJ20220530162202005).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data generation code will be available on request.

Acknowledgments: The authors extend their appreciation to reviewers.

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

References

1. Yan, G.; Luo, Z.; Liu, Z.; Li, Y. SensorX2car: Sensors-to-car calibration for autonomous driving in road scenarios. *arXiv* **2023**, arXiv:2301.07279.
2. Cattaneo, D.; Vaghi, M.; Valada, A. Lcdnet: Deep loop closure detection and point cloud registration for lidar slam. *IEEE Trans. Robot.* **2022**, *38*, 2074–2093.
3. Jiang, B.; Shen, S. Contour Context: Abstract Structural Distribution for 3D LiDAR Loop Detection and Metric Pose Estimation. *arXiv* **2023**, arXiv:2302.06149.
4. Huang, J.K.; Clark, W.; Grizzle, J.W. Optimal target shape for LiDAR pose estimation. *IEEE Robot. Autom. Lett.* **2021**, *7*, 1238–1245.
5. Wu, C.Y.; Johnson, J.; Malik, J.; Feichtenhofer, C.; Gkioxari, G. Multiview Compressive Coding for 3D Reconstruction. *arXiv* **2023**, arXiv:2301.08247.
6. Geiger, A.; Ziegler, J.; Stiller, C. Stereoscan: Dense 3d reconstruction in real-time. In Proceedings of the 2011 IEEE Intelligent Vehicles Symposium (IV). IEEE, 2011; pp. 963–968.
7. Durrant-Whyte, H.; Bailey, T. Simultaneous localization and mapping: Part I. *IEEE Robot. Autom. Mag.* **2006**, *13*, 99–110.
8. Shan, T.; Englot, B. Lego-loam: Lightweight and ground-optimized lidar odometry and mapping on variable terrain. In Proceedings of the 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). IEEE, 2018; pp. 4758–4765.
9. Besl, P.J.; McKay, N.D. Method for registration of 3-D shapes. In *Sensor Fusion IV: Control Paradigms and Data Structures*; SPIE, 1992; Volume 1611, pp. 586–606.
10. Segal, A.; Haehnel, D.; Thrun, S. Generalized-icp. In Proceedings of the Robotics: Science and Systems, Seattle, WA, USA, 2009; Volume 2, p. 435.
11. Biber, P.; Straßer, W. The normal distributions transform: A new approach to laser scan matching. In Proceedings of the 2003 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2003) (Cat. No. 03CH37453). IEEE, 2003; Volume 3, pp. 2743–2748.

12. Choy, C.; Park, J.; Koltun, V. Fully convolutional geometric features. In Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019; pp. 8958–8966.
13. Bai, X.; Luo, Z.; Zhou, L.; Chen, H.; Li, L.; Hu, Z.; Fu, H.; Tai, C.L. Pointdsc: Robust point cloud registration using deep spatial consistency. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021; pp. 15859–15869.
14. Choy, C.; Dong, W.; Koltun, V. Deep global registration. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020; pp. 2514–2523.
15. Sarode, V.; Li, X.; Goforth, H.; Aoki, Y.; Srivatsan, R.A.; Lucey, S.; Choset, H. Pcrnet: Point cloud registration network using pointnet encoding. *arXiv* **2019**, arXiv:1908.07906.
16. Lu, W.; Wan, G.; Zhou, Y.; Fu, X.; Yuan, P.; Song, S. Deepvcp: An end-to-end deep neural network for point cloud registration. In Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019; pp. 12–21.
17. Ao, S.; Hu, Q.; Yang, B.; Markham, A.; Guo, Y. Spinnet: Learning a general surface descriptor for 3d point cloud registration. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021; pp. 11753–11762.
18. DeTone, D.; Malisiewicz, T.; Rabinovich, A. Superpoint: Self-supervised interest point detection and description. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2018; pp. 224–236.
19. Ester, M.; Kriegel, H.P.; Sander, J.; Xu, X. Density-based spatial clustering of applications with noise. In Proceedings of the International Conference Knowledge Discovery and Data Mining, 1996; Volume 240.
20. 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, 2017; pp. 652–660.
21. Klokov, R.; Lempitsky, V. Escape from cells: Deep kd-networks for the recognition of 3d point cloud models. In Proceedings of the IEEE International Conference on Computer Vision, 2017; pp. 863–872.
22. Qi, C.R.; Yi, L.; Su, H.; Guibas, L.J. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *Adv. Neural Inf. Process. Syst.* **2017**, *30*.
23. Yan, Y.; Mao, Y.; Li, B. Second: Sparsely embedded convolutional detection. *Sensors* **2018**, *18*, 3337.
24. Maturana, D.; Scherer, S. Voxnet: A 3d convolutional neural network for real-time object recognition. In Proceedings of the 2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). IEEE, 2015; pp. 922–928.
25. Zhou, Y.; Tuzel, O. VoxelNet: End-to-End Learning for Point Cloud Based 3D Object Detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018.
26. Liu, Z.; Tang, H.; Lin, Y.; Han, S. Point-voxel cnn for efficient 3d deep learning. *Adv. Neural Inf. Process. Syst.* **2019**, *32*, 965–975.
27. Low, K.L. *Linear Least-Squares Optimization for Point-to-Plane Icp Surface Registration*; University of North Carolina: Chapel Hill, NC, USA, 2004; Volume 4, pp. 1–3.
28. Fischler, M.A.; Bolles, R.C. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Commun. ACM* **1981**, *24*, 381–395.
29. Wang, Y.; Solomon, J.M. Deep closest point: Learning representations for point cloud registration. In Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019; pp. 3523–3532.
30. Huang, X.; Mei, G.; Zhang, J.; Abbas, R. A comprehensive survey on point cloud registration. *arXiv* **2021**, arXiv:2103.02690.
31. Zhao, X.; Yang, S.; Huang, T.; Chen, J.; Ma, T.; Li, M.; Liu, Y. SuperLine3D: Self-supervised Line Segmentation and Description for LiDAR Point Cloud. In Proceedings of the Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, 23–27 October 2022; Proceedings, Part IX; Springer, 2022; pp. 263–279.
32. Liu, J.; Xu, Y.; Lin, W.; Sun, L. PointTrans: Rethinking 3D Object Detection from a Translation Perspective with Transformer. In Proceedings of the Chinese Control Conference, 2023; submitted.
33. Çiçek, Ö.; Abdulkadir, A.; Lienkamp, S.S.; Brox, T.; Ronneberger, O. 3D U-Net: Learning dense volumetric segmentation from sparse annotation. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2016; pp. 424–432.

34. 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, 2019; pp. 9297–9307.
35. Cao, A.Q.; Puy, G.; Boulch, A.; Marlet, R. PCAM: Product of cross-attention matrices for rigid registration of point clouds. In Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021; pp. 13229–13238.