symmetry

MDPI

Article

SCRnet: A Spatial Consistency Guided Network Using Contrastive Learning for Point Cloud Registration

Huixiang Shao, Zhijiang Zhang, Xiaoyu Feng and Dan Zeng *

School of Communication and Information Engineering, Shanghai University, Shanghai 200444, China; adamshao@shu.edu.cn (H.S.); zjzhang@shu.edu.cn (Z.Z.); xiaoyufeng@shu.edu.cn (X.F.)
* Correspondence: dzeng@shu.edu.cn

Abstract: Point cloud registration is used to find a rigid transformation from the source point cloud to the target point cloud. The main challenge in the point cloud registration is in finding correct correspondences in complex scenes that may contain many noise and repetitive structures. At present, many existing methods use outlier rejections to help the network obtain more accurate correspondences, but they often ignore the spatial consistency between keypoints. Therefore, to address this issue, we propose a spatial consistency guided network using contrastive learning for point cloud registration (SCRnet), in which its overall stage is symmetrical. SCRnet consists of four blocks, namely feature extraction block, confidence estimation block, contrastive learning block and registration block. Firstly, we use *mini-PointNet* to extract coarse local and global features. Secondly, we propose confidence estimation block, which formulate outlier rejection as confidence estimation problem of keypoint correspondences. In addition, the local spatial features are encoded into the confidence estimation block, which makes the correspondence possess local spatial consistency. Moreover, we propose contrastive learning block by constructing positive point pairs and hard negative point pairs and using Point-Pair-INfoNCE contrastive loss, which can further remove hard outliers through global spatial consistency. Finally, the proposed registration block selects a set of matching points with high spatial consistency and uses these matching sets to calculate multiple transformations, then the best transformation can be identified by initial alignment and Iterative Closest Point (ICP) algorithm. Extensive experiments are conducted on KITTI and nuScenes dataset, which demonstrate the high accuracy and strong robustness of SCRnet on point cloud registration task.

Keywords: point cloud registration; contrastive learning; spatial consistency; deep learning

Citation: Shao, H.; Zhang, Z.; Feng, X.; Zeng, D. SCRnet: A Spatial Consistency Guided Network Using Contrastive Learning for Point Cloud Registration. *Symmetry* **2022**, *14*, 140. https://doi.org/10.3390/sym14010140

Academic Editor: Theodore E. Simos

Received: 13 December 2021
Accepted: 5 January 2022
Published: 12 January 2022

1. Introduction

Point cloud registration is an important and fundamental field in 3D computer vision and graphics. It has many applications, such as 3D reconstruction [1], 3D image fusion [2], simultaneous localization and mapping (SLAM), [3–5], among others. In recent years, remarkable progress has been made in the point cloud registration, which aims to align the source to the target point cloud, so as to unify the two into the agreed coordinate system.

There have been several traditional efforts exploring on point cloud registration. One is based on the iterative closest point (ICP) algorithm [6,7], which iteratively estimates and finds the rigid transformation in a coarse-to-fine manner. However, the ICP algorithm easily falls into the local optimum due to the need to solve non-convex problems and the high dependence on initial values. The other is based on manual designed features [8]. The existing relatively good hand-crafted features, such as local feature statistic histogram (LFSH) [9], fast point feature histogram (FPFH) [10], and signature of histograms of orientations (SHOT) [11] have achieved remarkable results for feature extraction of point clouds in special scenes, but they often ignore the geometric relation and lack the semantic information of point clouds, and are often of low robustness.

Recently, following the success of deep convolutional neural networks, many learning-based point cloud feature extraction networks have emerged [12–16]. They often involve three steps. Firstly, the keypoint and feature descriptors are extracted from input point cloud by neural networks. Secondly, the outlier rejection method and the nearest neighbor algorithm are used to select the inner keypoints and obtain more accurate keypoint correspondences. Finally, the transformation from the source to target point cloud is estimated. Despite their favorable performance, there exists the correspondence generated by feature matching is still prone to outliers [17]. Therefore, removing outliers is obviously very crucial in the point cloud registration [18], which affects the accuracy of registration. Learning-based outlier rejection methods often rely on geometric operations to capture semantic information, such as sparse convolution [19] and pointwise MLP [20,21], but it ignores spatial information and may restrict the effect of outlier rejection. With the in-depth study of point cloud registration methods based on deep learning, a series of point cloud registration algorithms have been proposed to consider the spatial consistency between point clouds [22,23], but they are still hard to remove hard outliers.

In this paper, we propose a learning-based point cloud registration network named SCRnet, which comprehensively considers local spatial geometric consistency and global spatial consistency to remove outliers. The overall framework is displayed in Figure 1. Firstly, we use the farthest point sample (FPS) algorithm [24] to subsample the point cloud. Secondly, we perform a mini-PointNet for local and global feature embedding; it has a simple residual structure and outputs the corresponding high-dimensional features of each point cloud. Then, in order to obtain more accurate inliers, we design a confidence estimation block, which takes the outlier rejection problem as the confidence estimation problem of keypoint correspondences. The higher the confidence of two corresponding keypoints, the more likely they are to be corresponding inner points. Moreover, to consider the global spatial consistency of corresponding points, the contrastive learning [25,26] is introduced by constructing positive and hard negative keypoint pairs to remove global spatial inconsistent keypoints. Finally, we design a registration block, which selectively matches the keypoints filtered out by local and global spatial consistency, and estimates the optimal transformation through ICP algorithm. Therefore, the best rigid transformation can be determined by the above steps.

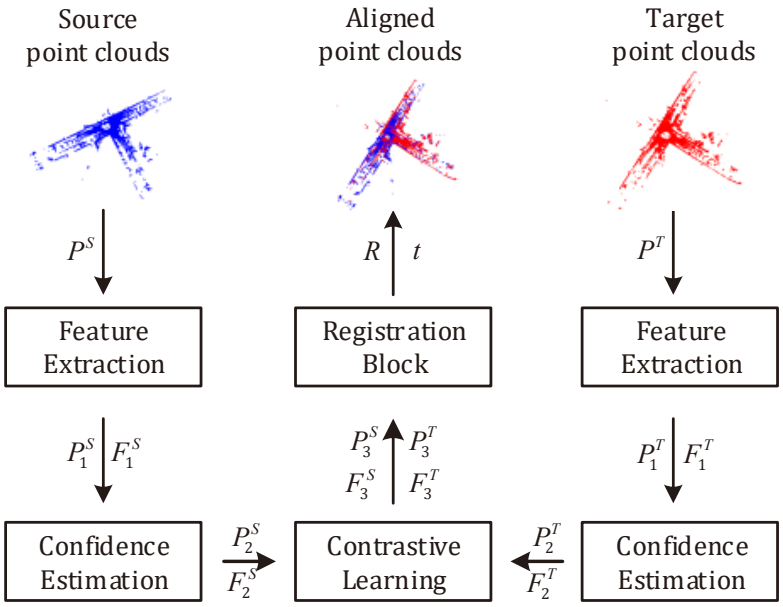


Figure 1. The overall procedure of SCRnet architecture. The point cloud is expressed as keypoints and features by the feature extraction block. Then we perform the confidence estimation block and the contrastive learning block to gradually remove outliers, and the registration block is followed to estimate the fine transformation by ICP algorithm.

In summary, our contributions are as follows:

- To obtain accurate inliers and remove hard outliers, we design a confidence estimation block, which considers the local spatial consistency and regards the outlier rejection problem as the confidence estimation problem of whether the keypoints correspond or not, and a contrastive learning block that embeds the global spatial consistency into the outlier rejection problem.
- To get tighter alignment, we design a registration block based on ICP algorithm; it calculates the initial transformation from the corresponding keypoints filtered by spatial consistency guidance, then refines and determines the optimal transformation using the ICP algorithm.
- Extensive experiments on the KITTI and nuScenes dataset demonstrate the superiority of our method, which achieves the state-of-the-art registration performance.

The rest of the paper is organized as follows. Section 2 reviews related works, to provide a better understanding of SCRnet. Section 3 introduces the details of the proposed methods. Section 4 provides experiment results and analyses. Section 5 discusses the role of contrastive learning and the limitations of the proposed network, followed by the conclusions in Section 6.

2. Related Works

In many point cloud registration pipelines, the first step is feature extraction. Currently, there are two ways of feature extraction. One is the traditional hand-crafted features [8–11]. These features mainly rely on geometry and mathematics understanding, and often model a special area to get the description of point cloud [27]. The other is the feature learned from the convolution neural network [28], they contain useful information distilled through multiple convolution and pooling operations [21]. In our work, we use a structure mini-PointNet similar to PointNet [20] to extract features. After extracting the keypoints and their features, it is important to filter out the outliers for more robust matching. At present, there are two techniques to remove outliers. One is the traditional outliers rejection technology based on RANSAC [29] and its variants, which uses the internal constraint relationship of point set to remove abnormal points. However, the main disadvantages are slow convergence speed and low accuracy when the outlier ratio is large. The other is the learning-based outlier rejection method [17,18,30,31], these methods regard the outlier rejection problem as an inlier and outlier classification problem, in which they embed the features corresponding to the keypoints, and predict the possibility that each pair of matching is an inline. For example, DGR [13] designs a six-dimensional convolution for point cloud feature extraction and internal and external point detection, and uses the differentiable weighted procrustes [32] algorithm for non rigid registration, and then obtains the optimal transformation through gradient optimization. The 3DRegNet [15] uses the deep learning network to directly regress the transformation, which performs registration by minimizing non-corresponding feature-metric projection errors rather than common geometric errors. However, the learning-based method often ignores the spatial consistency. Recently, many outlier removal methods combined with spatial consistency have been proposed, for example, PointDSC [33] uses a nonlocal feature aggregation module and a differentiable spectrum matching module, which explicitly combines spatial consistency to trim outlier correspondence, and achieves high performance and wide applicability. HRegNet [16] extracts keypoints and features of point cloud hierarchically, and designs a novel similarity feature that integrates bilateral consistency and neighborhood consistency into the registration pipeline, which significantly improves the registration performance. Unlike these methods, we designed the confidence estimation block and contrastive learning block to fully consider local and global spatial consistency during removing the outer point in the proposed SCRnet, then the estimated transformation is continuously optimized by using the ICP algorithm.

PointNet: PointNet [20] is the first network to classify and segment the original point cloud using deep neural networks (DNN) [28]. The network provides a simple and efficient

architecture for point cloud processing, and has verified its advanced nature on many tasks. Due to carrying out feature learning on the global point cloud, PointNet cannot obtain local features, which makes it difficult to analyze complex scenes. Therefore, PointNet++ [21] was proposed as an improvement of PointNet, which extracts the local features of the point set through hierarchical feature extraction structure.

Contrastive learning: contrastive learning [25,26] is a method that learns representation by comparing similar and dissimilar pairs. Among them, Deep InfoMax [34] proposes a contrastive learning framework for the first time, which maximizes the mutual information between the local patches and the global context. SimCLR [25] proposes a simple contrastive learning framework, which makes contrastive learning a hot research topic; it needs a big batch size to obtain superior performance. In the field of 3D point cloud processing, PointContrast [35] first proposes a framework of joint contrastive learning for point cloud representation. In addition, many studies show that hard negative samples are beneficial to contrastive learning. In our framework, by constructing positive samples and hard negative samples, we skillfully embed contrastive learning into the outlier rejection problem, so that our model can pay attention to the global spatial consistency and remove hard outliers.

3. Methods

When the source point cloud $P^S \in \mathbb{R}^{N\times 3}$ and the target point cloud $P^T \in \mathbb{R}^{N\times 3}$ are obtained, SCRnet aims to predict the optimal rigid transformation (including rotation matrix $\mathbf{R} \in G(3)$ and the translation vector $\mathbf{t} \in \mathbb{R}^3$) through removing the outliers step-by-step, where $G(3)$ is the three-dimensional rotation group of the Euclidean space. As shown in Figure 1, we propose and design multiple blocks for point cloud registration. Given a sparse point cloud, it is fed into the feature extraction block, outputs the keypoints $P \in \mathbb{R}^{M\times 3}$, and their features $F \in \mathbb{R}^{M\times C}$, M is number of the keypoints, C is dimension of the feature. In order to obtain better inliers and registration accuracy, we use the confidence estimation block for removing generous easy outliers, and contrastive learning is performed by constructing positive and hard negative point pairs for removing hard outliers. Finally, the remaining interior points after filtering are further fed into the registration block, it can calculate the optimal initial transformation matrix, and obtain the best registration result R, t through the ICP algorithm.

3.1. Feature Extraction

The input of the feature extraction block is the sampled point cloud; the farthest point sample (FPS) is often used to subsample the original point cloud. After that, as shown in Figure 2, we input the sampled point cloud (expressed as an $N \times c$ tensor) into the mini-PointNet. The output is the learned feature (expressed as an $N \times c2$ tensor), which contains information of the local feature embedding ($N \times c$) and global feature embedding ($c1$). The mini-PointNet structure we used here is implemented as a nonlinear function $h : \mathbb{R}^c \to \mathbb{R}^{c'}$ with a skip connection, realized as two cascaded shared-MLPs, followed by a max-pooling layer:

$$f' = h(MAX(h(f)) \oplus f), \tag{1}$$

where f is the coordinates of the input point cloud, f' is the output feature, $\oplus$ is the concatenation operation, which splices the input point cloud coordinates with the global embedding, and the $MAX(\cdot)$ operation is max-pooling.

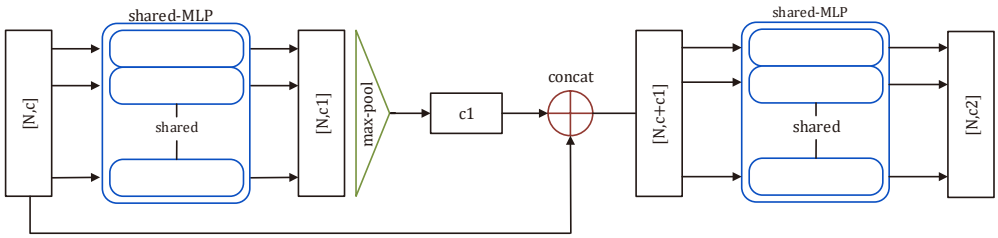


Figure 2. Mini-PointNet architecture. The block applies two MLPs with shared weights and performs max-pooling over point dimension to obtain a single feature embedding. The residual structure is used in this block, which connects the input coordinates and coarse global spatial embedding so that the block can obtain the point cloud feature with local and global properties.

3.2. Confidence Estimation Block

The confidence estimation block absorbs the essence of internal and external point classification, but adopts different methods, which uses the correspondence confidence of keypoint to judge the internal and external points. Unlike directly inputting the correspondence of two keypoints, our confidence estimation block inputs the cluster of keypoints and their features, which is beneficial to ensure local spatial consistency. As shown in Figure 3b, for a keypoint, taking the keypoint as the center of the sphere and r as the radius of the sphere, the other keypoints in the sphere area and the central keypoint form a sphere cluster. We use average distance coding (ADC) and angle coding (AC) to encode local spatial features of sphere clusters, the local spatial coding is shown in Figure 3a, which is a geometric feature inspired by the idea of symmetry. Then, these features are fed into a shared-MLP and a max-pooling layer, followed by a *sigmoid* operation. Thus, the confidence estimation block can learn the "local spatial correspondence possibility" between the source and target point cloud. Importantly, it can suppress the matching range and the matching noise by removing a large number of external points, which is also beneficial to the accuracy and robustness of the model.

Average distance coding: each point in the point cloud consists of coordinates (x, y, z), for the central keypoint p_i, the relative distance in a sphere cluster can be given:

$$ADC_i = \sum_{j=1}^{k} \{(p_i, p_j) | p_i \oplus \frac{1}{k} \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2 + (z_i - z_j)^2}\}, \quad (2)$$

where ADC_i is the average distance coding between the central keypoint and k the neighborhood keypoints in the sphere area, $\oplus$ is the concatenation operation.

Angle coding: similarly, to form the angle-based feature embedding, we calculate angles between the points. For the central keypoint p_i, the angle feature of a sphere cluster can be realized as $\{\alpha_1, \alpha_2, ..., \alpha_k\}$, where the angle between the central keypoint and the neighborhood keypoints can be obtained by Triangular Cosine Formula.

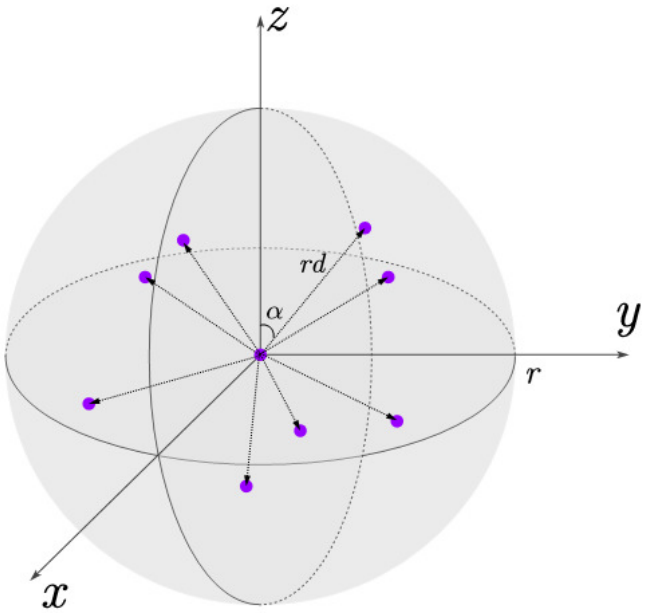
3.3. Contrastive Learning

Positive and hard negative points: the key of point cloud contrastive learning is the definition of positive keypoints and negative keypoints. Thanks to the above block design, we can easily build positive and hard negative keypoints. The above confidence estimation block first roughly outputs corresponding inliers and their confidence, as shown in Figure 4, for a inlier $p_i^S, i = 1, 2, .., K$ from the previous block in source point cloud domain. Firstly, we perform k-nearest-neighbor (k-NN) to search k neighboring inliers $p_j^T, j = 1, 2, ..., k$ (k is different from K) in the target point cloud domain. Then the inlier p_i^S and k neighboring keypoints form a cross-domain cluster and total of K cross-domain clusters can be formed. Last, we calculate the average confidence of each cross-domain cluster through the confidence of correspondence of each keypoint output by the confidence estimation block. The level of "hardness" for negative clusters is dependent on the average

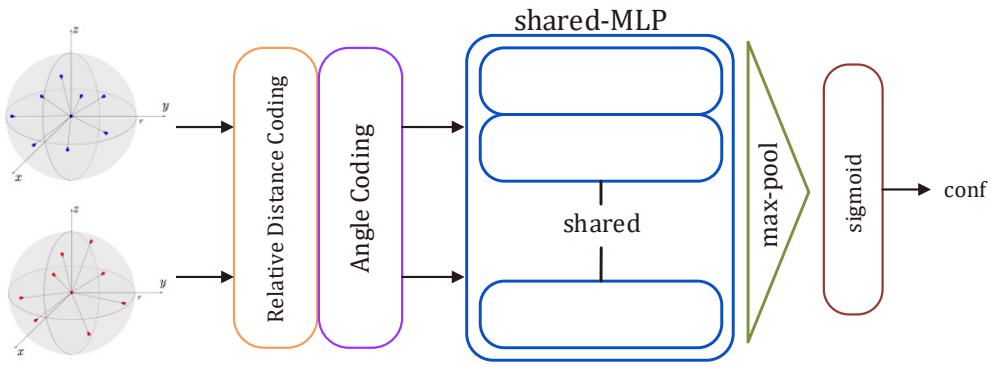
confidence, i.e., the higher the confidence of cluster, the more likely it is to be a hard negative cluster. After constructing the positive and hard negative point pairs, inspired by information supervision between point pairs, we design a point-pairs INfoNCE contrastive loss for outlier rejection, which can help the confidence estimation network remove hard outliers and output more accurate confidence scores:

$$\mathcal{L}_{cl} = -\sum_{i=0}^{K} \log\left(\frac{\exp((g_i \cdot g_+)/\tau)}{\exp((g_i \cdot g_-)/\tau)}\right), \tag{3}$$

where g_i is the feature of cluster formed by the inlier and its neighboring keypoints, g_+ and g_- are the characteristics of positive clusters and hard negative clusters, respectively. τ is the temperature factor, K is the number of cross-domain clusters.



(**a**)



(**b**)

Figure 3. (**a**) Local spatial coding: local spatial features are encoded by the relative distance and angle of the neighborhood points, where *rd* represents a relative distance and α represents the angle in a point cloud domain. (**b**) Confidence estimation block: the input is the sphere cluster feature encoded by local spatial coding, and the output is the confidence score of the corresponding keypoints, where the blue round is the source point cloud and the red round is the target point cloud.

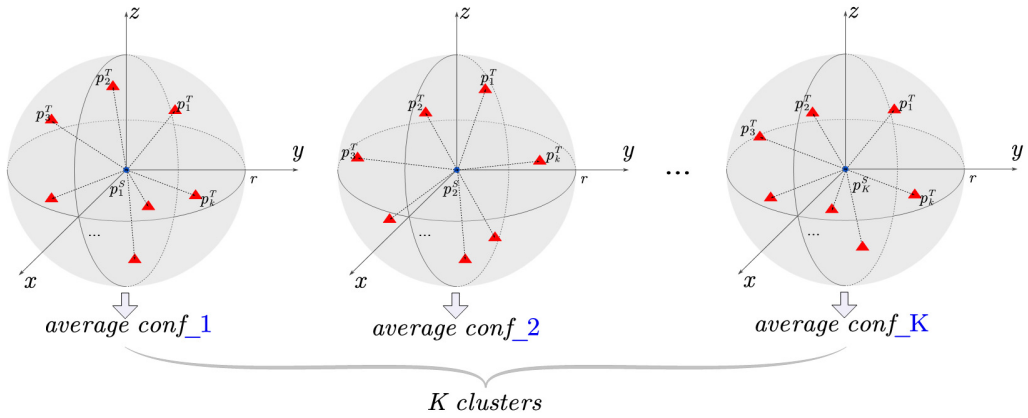


Figure 4. Illustration of positive and hard negative sample. For the K points in source point cloud domain, we find k neighborhood points in the target point cloud domain and calculate the average confidence of the cross-domain cluster through the confidence of the correspondence of each keypoint. Then the cluster with highest average confidence is regarded as a positive example, and the second to fifth highest random is regarded as a hard negative example. Among them, the blue circle is the source point cloud and the red triangle is the target point cloud.

3.4. Registration Block

Through the processing of the above block, some accurate inliers are obtained, which are the inputs of the registration block. For each accurate inlier $p_i^S, i = 1, 2, ..., M$ in source point cloud domain, we calculate the cosine similarity between its features f_i^S and all interior point features $f_1^T, f_2^T, ..., f_{M-1}^T$ in the target point cloud domain, where M is the number of inliers. Consequently, the $M \times M$ similarity matrix S can be constructed by the similarity scores s_{ij} between all interior point features:

$$s_{ij} = \frac{< f_i^S, f_j^T >}{||f_i^S||_2 ||f_j^T||_2}, \tag{4}$$

where $< \cdot >$ and $|| \cdot ||_2$ represent inner product and L_2 norm, respectively.

The registration block is shown in Figure 5. When the corresponding similarity matrix is established, we use a strategy based ICP algorithm to calculate the best rotation and translation parameter, as follows:

1. The m pairs $\psi_1, \psi_2, \psi_3, ..., \psi_m$ with high corresponding similarity scores, called "control pairs", are selected from the set of matched pairs, where ψ_z consists of three point pairs because three pairs can determine a transformation matrix, defined as $\psi_z = (p_i^S, p_i^T), (p_j^S, p_j^T), (p_k^S, p_k^T), z = 1, 2, ..., m$.
2. The transformations $(R, t)_1, (R, t)_2, ..., (R, t)_m$ corresponding to m pairs are calculated by singular value decomposition [36,37] (SVD). First, we define $\overline{P^S}$ and $\overline{P^T}$ as centroids of point cloud P^S and P^T. The covariance matrix Cov is computed by:

$$Cov = (P^T - \overline{P^T})(P^S - \overline{P^S})^T. \tag{5}$$

Then, we conduct the SVD to decompose Cov to U, V:

$$USV^T = SVD(Cov). \tag{6}$$

Subsequently, the rotation matrix R and translation vector t can be given by the following formula:

$$R = VU^T, \tag{7}$$

$$t = -R \cdot \overline{P^S} + \overline{P^T}. \tag{8}$$

Finally, in order to distinguish which transformation is the optimal initial transformation, we align the source point cloud with target point cloud by the given R, t and calculate the Euclidean distance between the aligned source point cloud and the target point cloud. The smaller the distance, the better the initial transformation.

3. After obtaining the optimal initial transformation matrix, the ICP algorithm is used to continuously optimize the initial transformation and get the best transformation $\widehat{R}, \widehat{t}$.

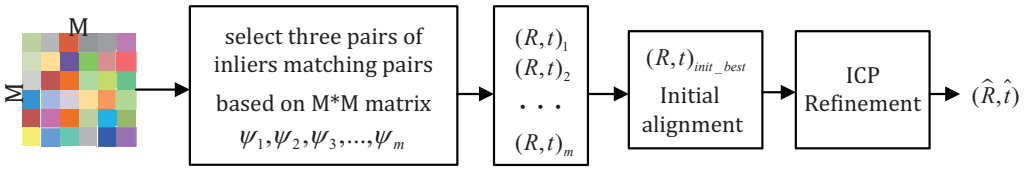


Figure 5. The process of registration block, which selects the control pairs based on similarity and obtains the optimal initial transformation for subsequent iterative optimization.

3.5. Loss Function

The loss function $\mathcal{L} = \alpha\mathcal{L}_{trans} + \beta\mathcal{L}_{rot} + \gamma\mathcal{L}_{cl}$, where $\mathcal{L}_{trans}$ is translation loss, $\mathcal{L}_{rot}$ is rotation loss and $\mathcal{L}_{cl}$ is contrastive loss. When the model predicts the best transformation $\widehat{R}, \widehat{t}$, the truth transformation R, t is known, $\mathcal{L}_{trans}$ and $\mathcal{L}_{rot}$ are defined by Equations (9) and (10), respectively, and $\mathcal{L}_{cl}$ is defined by Equation (3).

$$\mathcal{L}_{trans} = ||t - \widehat{t}||_2, \tag{9}$$

$$\mathcal{L}_{rot} = ||\widehat{R}^T R - I||_2, \tag{10}$$

where I denote identity matrix.

4. Experiments

4.1. Experiment Setting

Datasets: to prove the advanced performance of the model, we perform extensive experiments on the two large-scale automatic driving LiDAR point cloud benchmark datasets. One is KITTI odometry dataset [38], which is captured with a Velodyne HDL64 LiDAR in Karlsruhe, Germany, and ground truth provided by a GNSS/INS integrated navigation system. It is composed of 11 sequences with real ground posture, we regard two point clouds with an interval of 10 frames as a registration pair, and take 60% frames of each sequence as the training data, the 20% as the validation data and the 20% as the testing data. The other is the nuScenes dataset [39], which is captured with a full sensor suite, such as 1 × LiDAR, 5 × RADAR, 6 × camera, IMU, and GPS. It consists of 1000 scenes, among them, the first 600 scenes are used as training data, and the remaining 400 scenes are used as validation data and testing data. During the training process, we implement enhancement steps based on random transformation and noise, more specially, we create the pairs to be registered by duplicating the target point cloud and applying random transformation and noise to the source point cloud.

Implementation details: during data preprocessing, we use voxel mesh filter and FPS algorithm to sample 32,768 points on the KITTI odometry dataset and 16,384 points on the nuScenes dataset. The network is implemented based on the PyTorch [40] framework, it uses Adam [41] as the gradient optimizer, and the learning rate is set to 0.0001 and the exponential decayed factor is set to 0.99. The optimal value of the hyperparameter α, β, γ is set by a repetitive experiment. The training of the whole model is divided into two steps. The first step is to pre-train the feature extraction block, and the second step is to freeze the feature extraction block and train the whole model on the pre-training features. The whole model is trained and tested on GPU NVIDIA TITAN RTX 24 G workstation.

Baselines: we compare the proposed SCRnet against many strong baselines from traditional methods to recent learning-based approaches.

Traditional methods: to verify the performance of the proposed SCRnet, we compare with the following three typical traditional methods. (1) The ICP algorithm aligns the corresponding point set through the initial alignment, then calculates the new corresponding point set according to the nearest point and continuous iteration. (2) Fast global registration (FGR) [42] achieves a reliable result through a well initialized local optimization algorithm, and the computational cost is more than one order of magnitude lower than the previous rough global alignment algorithms. (3) RANSAC is used to estimate the transformation matrix without iteration, which learns an independent feature extraction network by a separated training process.

Learning-based methods: for the learning-based methods, we select three networks for comparison. (1) Deep closet point (DCP) [43]: It proves the role of local and global features in the point cloud registration, and considers the relationship between two point clouds by employing the embedding generated by the attention module. Moreover, a differentiable SVD decomposition layer is proposed to solve rigid transformation. (2) Feature-metric registration (FMR) [44]: it believes that the features extracted from point clouds with different poses are different, so it takes the differences of the point cloud features with different poses as the objective functions for iterative solutions. (3) Deep global registration (DGR) [13]: it constructs a likelihood prediction network for the interior point correspondence based on the designed six-dimensional convolution, and estimates the transformation by the weighted procrustes algorithm.

4.2. Evaluation

Evaluation metrics: we use three classical evaluation metrics in the point cloud registration. (1) Registration recall (RR), it is the percentage of successful matching under a certain rotation error and translation error threshold. (2) Rotation error (RE) is calculated as the degrees by the inverse cosine function. (3) Translation error (TE) represents the distance between the predicted translation vector and the ground truth translation vector. RE and TE are defined by Equations (11) and (12):

$$RE(\widehat{R}) = \arccos \frac{Tr(\widehat{R}^T R) - 1}{2}, \tag{11}$$

$$TE(\widehat{t}) = \left\| \widehat{t} - t \right\|_2, \tag{12}$$

where R is the ground truth of rotation matrix, t is the ground truth of translation vector, $\widehat{R}$ and $\widehat{t}$ denote the estimated optimal transformation. When the RE is less than 5 degrees and TE is less than $2m$, the registration is commonly considered successful. The RE and TE are only computed on successfully aligned pairs.

Evaluation results: the quantitative comparisons with some advanced methods are summarized in Table 1, and some successful registration cases are shown in Figures 6 and 7. According to the results, ICP algorithm and FGR fail to achieve a relatively reasonable transformation due to many outliers in most cases. RANSAC estimates the optimal transformation by iteration in the dataset containing outliers, which achieves the best performance among the traditional methods due to its ability to filter out outliers with ratio or reciprocity tests. For the learning-based methods, it can be seen that the registration recall of DCP is both higher than that of ICP and FGR on KITTI and nuScenes dataset, but far lower than that of RANSAC, and the TE or RE is also higher than the three traditional methods. The registration recall of FMR is more than 90%, but it is much lower than ours, and its TE and RE are still higher than RANSAC, which means that it can recall many accurate interior point, but the registration accuracy is not high enough. Among all baseline methods, DGR achieves the best registration recall rate, but the voxelization of point cloud loses much information of the original point cloud and hinders the registration accuracy. As shown in rows 4 and 7 of the Table 1, the TE of DGR on KITTI dataset is almost three times that of RANSAC. By contrast, our proposed SCRnet has achieved the best registration recall

due to strict outlier rejection, and the precision also has reached or even exceeded the best results of other methods.

Table 1. Registration performance comparision on KITTI and nuScenes dataset.

Methods	KITTI Dataset			nuScenes Dataset		
	TE (m)	RE (deg)	RR	TE (m)	RE (deg)	RR
ICP [6]	0.04 ± 0.05	0.11 ± 0.09	14.3%	0.25 ± 0.51	0.25 ± 0.50	18.0%
FGR [42]	0.93 ± 0.59	0.96 ± 0.81	39.4%	0.71 ± 0.62	1.01 ± 0.92	32.2%
RANSAC [29]	0.13 ± 0.07	0.54 ± 0.40	91.9%	0.21 ± 0.19	0.74 ± 0.70	60.9%
DCP [43]	1.03 ± 0.51	2.07 ± 1.19	47.3%	1.09 ± 0.49	2.07 ± 1.14	58.6%
FMR [44]	0.66 ± 0.42	1.49 ± 0.85	90.6%	0.60 ± 0.39	1.61 ± 0.97	92.1%
DGR [13]	0.32 ± 0.32	0.37 ± 0.30	98.7%	0.21 ± 0.18	0.48 ± 0.43	98.4%
SCRnet	**0.13** ± **0.13**	**0.30** ± **0.25**	**99.4**%	**0.20** ± **0.18**	**0.46** ± **0.30**	**99.5**%

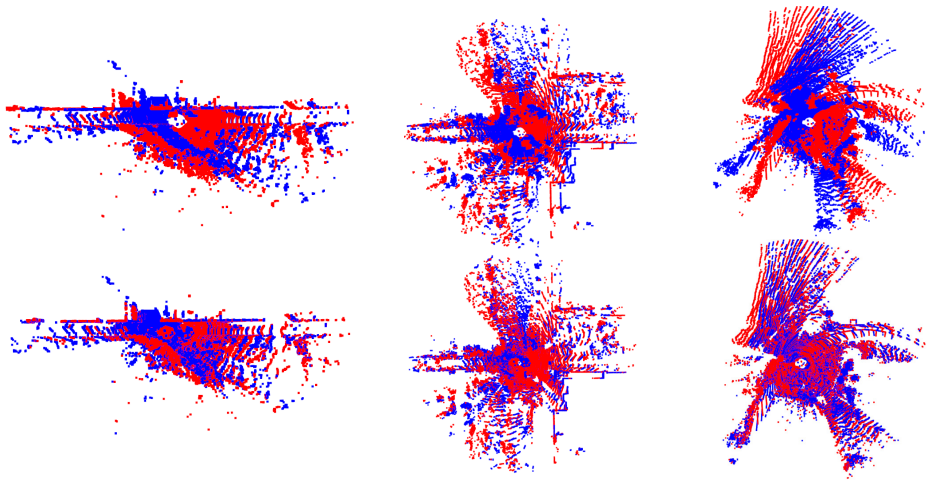


Figure 6. Successful registration examples from the KITTI dataset. The first row: blue represents the source point cloud, red represents the target point cloud. The second row: blue represents the aligned source point cloud using SCRnet, red represents the target point cloud.

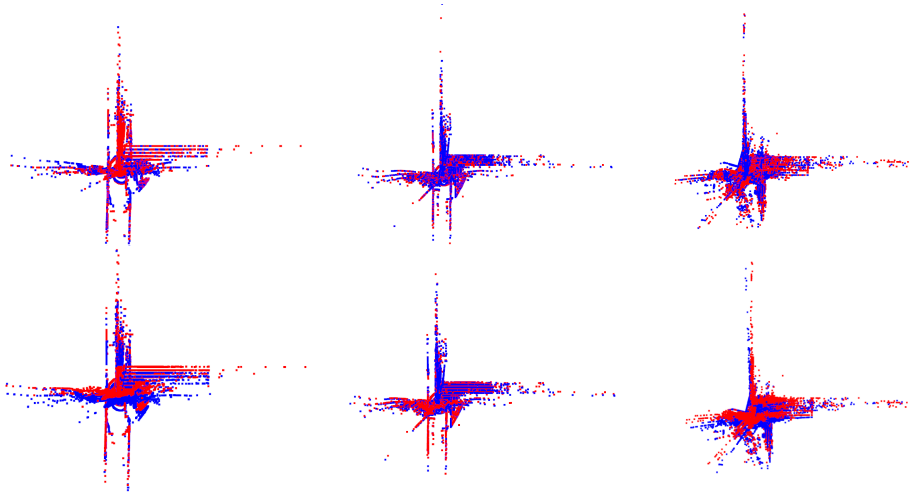


Figure 7. Successful registration examples from the nuScenes dataset. The first row: blue represents the source point cloud, red represents the target point cloud. The second row: blue represents the aligned source point cloud using SCRnet, red represents the target point cloud.

In the registration block, the optimal initial alignment can be used as the initial value of the ICP algorithm. That is, the proposed SCRnet first obtains the optimal initial transformation, and then iteratively refines the transformation through the ICP algorithm, which

ensures the learned transformation is strong and robust. To demonstrate this standpoint, we continuously change the maximum rotation angle to plot the curves of the translation error (TE) and rotation error (RE) in Figure 8. As we can see, the TE and RE of SCRnet are very stable with the increases of the maximum rotation angle, reflecting the reliable and robust registration power of SCRnet.

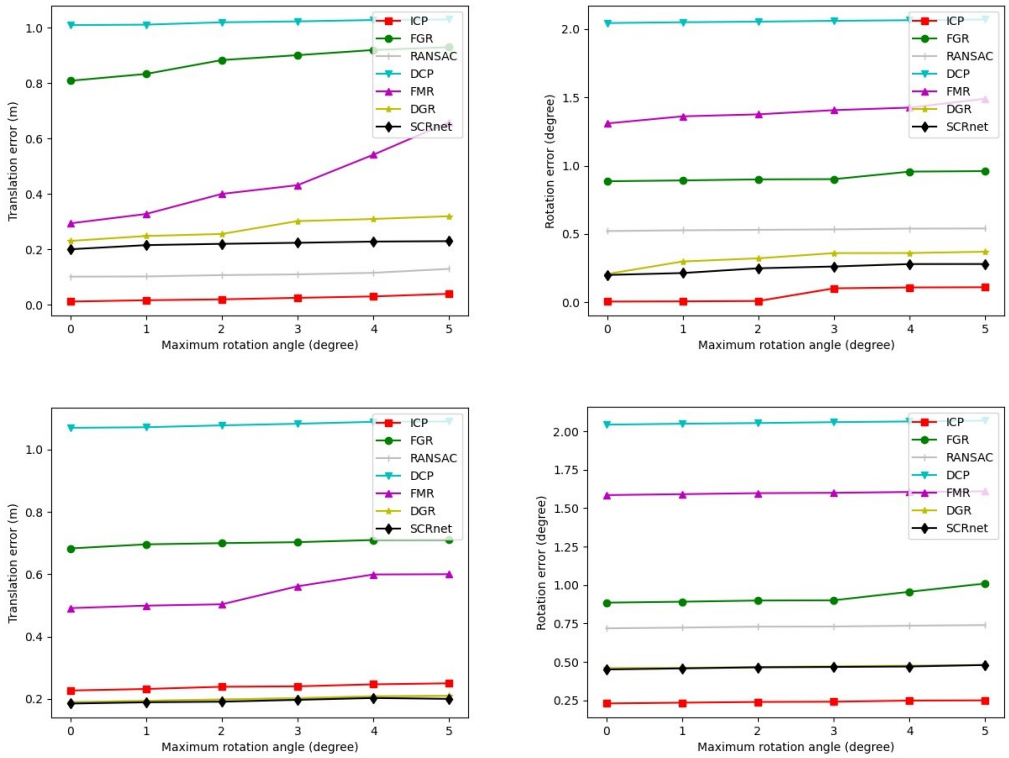


Figure 8. The plots of the maximum rotation angle against the translation error and rotation error, where the first row was measured on the KITTI dataset, the second row was measured on nuScenes dataset.

4.3. Ablation Study

We perform a detailed ablation experiment on the KITTI dataset to verify the role of MLPs, local spatial coding, contrastive learning, and the ICP algorithm. The results are summarized in Table 2. The network only with a shared MLP provides a benchmark, which is 79.6%. Using local spatial to coding geometric features for confidence estimation of the corresponding keypoints, the registration recall is improved to 85.5%. Then, a great improvement of 5.3% is because of contrastive learning. Moreover, the ICP algorithm for fine-tuning transformation leads to a good gain of 3.2%. However, with 2 MLPs and 3 MLPs, the results are improved to 93.2% and 93.6%, respectively. Finally, the best result is 99.4% by using the ICP algorithm, and the best registration recall can be achieved when using 2 MLPs, which saves model parameters compared with using 3 MLPs.

ICP algorithm is very significant for obtaining more reliable and solid alignment, which helps our network more stable and tighter. Importantly, a good initial alignment is often the key to the success of ICP algorithm. In this experiment, we add Gaussian noise with a mean of zero and a standard deviation of 0.01 to the point cloud to generate a noisy point cloud, then we align the noise source point cloud with the noise free target point cloud. The registration results are shown in Figure 9, it shows three registration methods: SCRnet without using the ICP algorithm, SCRnet using the ICP algorithm with the random initial value, and SCRnet using the ICP algorithm with an optimal initial value. We can see that the SCRnet using the ICP algorithm is robust to Gaussian noise and obtains a slightly better result compared with not using the ICP algorithm. Moreover, we can find that the SCRnet using the ICP algorithm with an optimal initial value achieves the

best performance because the ICP algorithm may fail to align well without the excellent initialization. Obviously, a fine registration block-based ICP algorithm can reduce the error, which further proves that good initial alignment is important and the strict outlier rejection is the core of registration task.

Table 2. Ablation study on KITTI dataset.

Point	MLP	Local Spatial Coding	Contrastive Learning	ICP Algorithm	RR (%)
32,768	1				79.6
32,768	1	✓			85.5
32,768	1	✓	✓		90.8
32,768	2	✓	✓		93.2
32,768	2	✓	✓	✓	99.4
32,768	3	✓	✓		93.6
32,768	3	✓	✓	✓	99.4

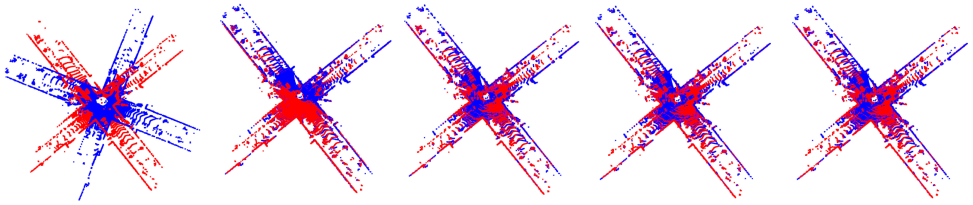


Figure 9. Ablation study about the importance of initial values in ICP algorithm.

5. Discussion

5.1. Role of Contrastive Learning

Outlier rejection is a very important step in the point cloud registration. It is very clever to treat outer point removal as an inner and outer point classification problem in 3DRegNet. Our method fully absorbs this idea, carries out the aggregation analysis of local features through the confidence estimation block, and estimates the confidence of the correspondence of keypoints. However, as shown in Figure 10, owing to that the scene point cloud has a highly repetitive structure and homogeneous architectural layouts, there may be more than one high or even equal confidence in corresponding point pairs, so it is necessary to remove highly similar non corresponding point pairs. The contrastive learning in the proposed SCRnet is designed for this problem.

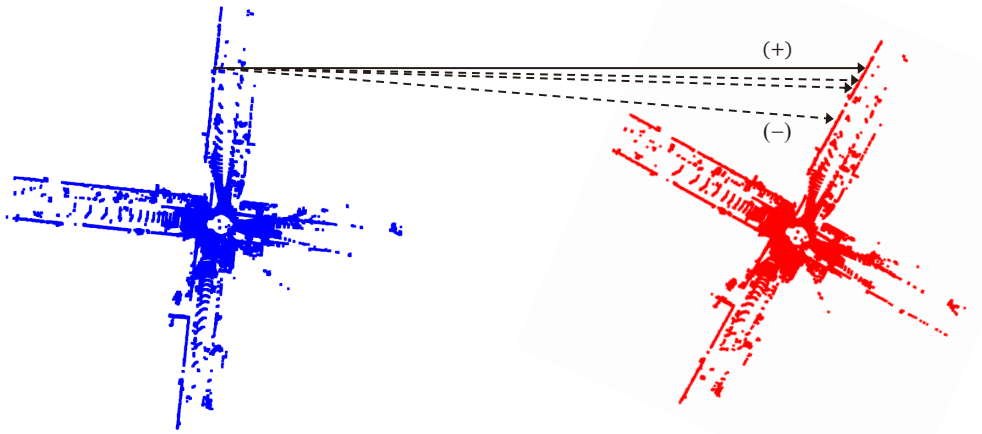


Figure 10. Illustration of positive and hard negative corresponding point pairs.

5.2. Limitations of SCRnet

In general, we see that SCRnet works extremely well for the registration. However, there exist some recent exemplary networks that have a better effect. From the perspective of feature extraction block and confidence estimation block, we can find that SCRnet uses the distance and angle of the neighborhood to encode the local spatial information. However, there are many existing superior geometric feature descriptors, which may be of great benefit to the performance improvement of the network. In addition, it is observed that we use an empirical value, as the size of the cluster and the k value of nearest neighbor when constructing local spatial features of neighborhood and applying k-NN. Therefore, specific empirical values need to be set for point clouds in different scenes due to the variability of the scene.

6. Conclusions

In this paper, a spatial consistency guided network using contrastive learning for point cloud registration, called SCRnet, was proposed. We found that the contrastive learning block is of great benefit to remove hard outliers, and thanks to the design of the confidence estimation block, we can easily construct positive point pairs and hard point pairs through the confidence between keypoint correspondences. Moreover, it is noteworthy that the best transformation can be identified by the optimal initial alignment and ICP algorithm in registration block, which drives the model to output tighter alignment. The extensive experiments on the KITTI and nuScenes dataset show that our method achieves significant improvement compared with state-of-the-art methods. Nevertheless, there are still some problems. The ICP algorithm is used to refine the transformation in a registration block, which limits the inference speed of the model. In the future, we will attempt to use a faster ICP algorithm to accelerate. Another possibility for future work is to improve the scalability of SCRnet to deal with large-scale real LiDAR point clouds. Furthermore, we hope our proposed method can be incorporated into larger pipelines and find more applications beyond the cases shown in this paper.

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

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: All datasets used to support the findings of this study were supplied by the publicly available databases.

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

References

1. Geiger, A.; Ziegler, J.; Stiller, C. Stereoscan: Dense 3d reconstruction in real-time. In Proceedings of the 2011 IEEE Intelligent Vehicles Symposium (IV), Baden-Baden, Germany, 5–9 June 2011; pp. 963–968.
2. Besl, P.J.; McKay, N.D. Method for registration of 3-d shapes. In *Sensor Fusion IV: Control Paradigms and Data Structures*; International Society for Optics and Photonics: Bellingham, DC, USA, 1992; Volume 1611, pp. 586–606.
3. Chen, S.; Liu, B.; Feng, C.; Vallespi-Gonzalez, C.; Wellington, C. 3D Point Cloud Processing and Learning for Autonomous Driving: Impacting Map Creation, Localization, and Perception. *IEEE Signal. Proc. Mag.* **2020**, *38*, 68–86. [CrossRef]
4. Deschaud, J.E. Imls-slam: Scan-to-model matching based on 3d data. In Proceedings of the IEEE International Conference on Robotics and Automation (ICRA), Brisbane, Australia, 21–25 May 2018; pp. 2480–2485.
5. Han, L.; Xu, L.; Bobkov, D.; Steinbach, E.; Fang, L. Real-time global registration for globally consistent rgb-d slam. *IEEE Trans. Robot.* **2019**, *35*, 498–508. [CrossRef]

6. Segal, A.; Haehnel, D.; Thrun, S. Generalized-icp. In *Robotics: Science and Systems*; The MIT Press: Cambridge, MA, USA, 2009; Volume 2, p. 435.
7. Yang, J.; Li, H.; Campbell, D.; Jia, Y. Go-icp: A globally optimal solution to 3d icp point-set registration. *IEEE Trans. Pattern Anal. Mach. Intell.* **2015**, *38*, 2241–2254. [CrossRef] [PubMed]
8. Li, H.; Hartley, R. The 3D-3D registration problem revisited. In Proceedings of the International Conference on Computer Vision (ICCV), Rio de Janeiro, Brazil, 14–20 October 2007; pp. 1–8.
9. Yang, J.; Cao, Z.; Zhang, Q. A fast and robust local descriptor for 3D point cloud registration. *Inf. Sci.* **2016**, *346*, 163–179. [CrossRef]
10. Rusu, R.B.; Blodow, N.; Beetz, M. Fast point feature histograms (FPFH) for 3D registration. In Proceedings of the 2009 IEEE International Conference on Robotics and Automation, Kobe, Japan, 12–17 May 2009; pp. 3212–3217.
11. Salti, S.; Tombari, F.; Stefano, L.D. Shot: Unique signatures of histograms for surface and texture description. *Comput. Vis. Image Underst.* **2014**, *125*, 251–264. [CrossRef]
12. Ao, S.; Hu, Q.; Yang, B. SpinNet: Learning a General Surface Descriptor for 3D Point Cloud Registration. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 19–25 June 2021; pp. 11753–11762.
13. Choy, C.; Dong, W.; Koltun, V. Deep global registration. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 2514–2523.
14. Zhang, Z.; Dai, Y.; Sun, J. Deep learning based point cloud registration: An overview. *Virtual Real. Intell. Hardw.* **2020**, *2*, 222–246. [CrossRef]
15. Pais, G.D.; Ramalingam, S.; Govindu, V.M.; Nascimento, J.C.; Chellappa, R.; Miraldo, P. 3dregnet: A deep neural network for 3d point registration. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 7193–7203.
16. Lu, F.; Chen, G.; Liu, Y. HRegNet: A Hierarchical Network for Large-scale Outdoor LiDAR Point Cloud Registration. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Nashville, TN, USA, 19–25 June 2021; pp. 16014–16023.
17. Efraim, A.; Francos, J.M. Dual Transformation and Manifold Distances Voting for Outlier Rejection in Point Cloud Registration. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 4204–4212.
18. Li, J. A Practical O (N2) Outlier Removal Method for Point Cloud Registration. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**. [CrossRef]
19. Li, Y.; Bu, R.; Sun, M.; Wu, W.; Di, X.; Chen, B. PointCNN: Convolution on X-transformed points. *Adv. Neural Inf. Process. Syst.* **2018**, *31*, 828–838.
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, Honolulu, HI, USA, 21–26 July 2017; pp. 652–660.
21. Qi, C.R.; Yi, L.; Su, H.; Guibas, L.J. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *arXiv* **2017**, arXiv:1706.02413.
22. Quan, S.; Yang, J. Compatibility-guided sampling consensus for 3-d point cloud registration. *IEEE Trans. Geosci. Remote Sens.* **2020**, *58*, 7380–7392. [CrossRef]
23. Ge, X.; Hu, H. Object-based incremental registration of terrestrial point clouds in an urban environment. *ISPRS J. Photogramm. Remote Sens.* **2020**, *161*, 218–232. [CrossRef]
24. Zhou, Y.; Wan, G.; Hou, S.; Yu, L.; Wang, G.; Rui, X.; Song, S. Da4ad: End-to-end deep attention-based visual localization for autonomous driving. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020; pp. 271–289.
25. Chen, T.; Kornblith, S.; Norouzi, M.; Hinton, G. A simple framework for contrastive learning of visual representations. *arXiv* **2020**, arXiv:2002.05709.
26. van den Oord, A.; Li, Y.; Vinyals, O. Representation learning with contrastive predictive coding. *arXiv* **2018**, arXiv:1807.03748.
27. Zeng, A.; Song, S.; Nießner, M.; Fisher, M.; Xiao, J.; Funkhouser, T. 3DMatch: Learning local geometric descriptors from RGB-D reconstructions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017.
28. Liu, W.; Wang, Z.; Liu, X. A survey of deep neural network architectures and their applications. *Neurocomputing* **2017**, *234*, 11–26. [CrossRef]
29. 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. [CrossRef]
30. Bustos, A.P.; Chin, T. Guaranteed outlier removal for point cloud registration with correspondences. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *40*, 2868–2882. [CrossRef] [PubMed]
31. Yang, H.; Carlone, L. A polynomial-time solution for robust registration with extreme outlier rates. *arXiv* **2019**, arXiv:1903.08588.
32. Gower, J.C. Generalized procrustes analysis. *Psychometrika* **1975**, *2*, 4. [CrossRef]
33. Bai, X.; Luo, Z.; Zhou, L. PointDSC: Robust Point Cloud Registration using Deep Spatial Consistency. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 19–25 June 2021; pp. 15859–15869.
34. Hjelm, R.D.; Fedorov, A.; Lavoie-Marchildon, S.; Grewal, K.; Trischler, A.; Bengio, Y. Learning deep representations by mutual information estimation and maximization. *arXiv* **2019**, arXiv:1808.06670.

35. Xie, S.; Gu, J.; Guo, D.; Qi, C.R.; Guibas, L.; Litany, O. Pointcontrast: Unsupervised pre-training for 3d point cloud understanding. In Proceedings of the European Conference on Computer Vision, Munich, Germany, 8–14 September 2020; pp. 574-591.
36. Bes, P.J.; McKay, N.D. A method for registration of 3-d shapes. *IEEE Trans. Pattern Anal. Mach. Intell.* **1992**, *14*, 239–256.
37. Billings, S.D.; Boctor, E.M.; Taylor, R.H. Iterative most-likely point registration (imlp): A robust algorithm for computing optimal shape alignment. *PLoS ONE* **2015**, *10*, e0117688.
38. Geiger, A.; Lenz, P.; Urtasun, R. Are we ready for autonomous driving? The kitti vision benchmark suite. In Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, 16–21 June 2012; pp. 3354–3361.
39. Liong, V.E.; Xu, Q.; Krishnan, A.; Pan, Y.; Baldan, G.; Beijbom, O. nuscenes: A multimodal dataset for autonomous driving. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 14–19 June 2020; pp. 11621–11631.
40. DValsesia, I.; Fracastoro, G.; Paszke, E.M.L.l.A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; et al. Pytorch: An imperative style, high-performance deep learning library. *Adv. Neural Inf. Process. Syst.* **2019**, *32*, 8024–8035.
41. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. In Proceedings of the 3rd International Conference on Learning Representations, San Diego, CA, USA, 7–9 May 2015.
42. Zhou, Q.; Park, J.; Koltun, V. Fast global registration. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 11–14 October 2016; pp. 766–782.
43. 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, Seoul, Korea, 27–28 October 2019; pp. 3523–3532.
44. Huang, X.; Mei, G.; Zhang, J. Featuremetric registration: A fast semi-supervised approach for robust point cloud registration without correspondences. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 14–19 June 2020; pp. 11366–11374.