Pattern Recognition Letters 151 (2021) 120–126

Contents lists available at ScienceDirect

Pattern Recognition Letters

journal homepage: www.elsevier.com/locate/patrec

SCANet: A Spatial and Channel Attention based Network for Partial-to-Partial Point Cloud Registration

Ruqin Zhou[a], Xixing Li[b], Wanshou Jiang[a,*]

[a] State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan 430072, China
[b] China National Digital Switching System Engineering and Technological Research Center, Zhengzhou 450000, China

ARTICLE INFO

Article history:
Received 18 March 2021
Revised 7 July 2021
Accepted 7 August 2021
Available online 14 August 2021

Edited by Jiwen LU, Ph.D.

ABSTRACT

Point cloud registration plays an essential role in many areas, such as computer vision and robotics. However, traditional feature-based registration requires handcrafted descriptors for various scenarios, which is of low efficiency and flexibility; ICP and its locally optimal variants are sensitive to initialization, while globally optimal methods are of high computational time to overcome noise, outliers, and partial overlap. Learning-based registration can automatically and flexibly learn shape representation for different objects, but existing methods are of either low efficiency or low precision, and poorly perform in partial-to-partial point cloud registration. Thus, we present a simple spatial and channel attention based network, named SCANet, for partial-to-partial point cloud registration. A spatial self-attention aggregation (SSA) module is applied in a feature extraction sub-network to efficiently make use of the inter and global information of each point cloud in different levels, while a channel cross-attention regression (CCR) module is adopted in a pose estimation sub-network for information interaction between two input global feature vectors, enhancing relevant information and suppressing redundant information. Experimental results show that our SCANet achieves state-of-the-art performances in both accuracy and efficiency compared to existing non-deep learning and learning-based methods on partial visibility with Gaussian noise. Our source code is available at the project website https://github.com/zhouruqin/SCANet.

1. Introduction

With a rapid development of LiDAR, stereo cameras and structured light sensors, 3D point clouds are ubiquitous with a wide range of applications [1], for example, autonomous driving, robotics, heritage modeling, augmented reality, environment survey, and urban planning. Accordingly, there has been a growing interest in improving performances of classification, segmentation, detection, reconstruction, and tracking etc. However, it is difficult for traditional algorithms to meet the requirements of practical application with enormous point clouds in various scenarios. Recently, great breakthrough has been made by deep learning in both efficiency and accuracy of many fields, especially Natural Language Processing (NLP) and image processing [2], which brings both opportunities and challenges to the data processing of point clouds.

Point cloud registration is a task to find a rigid transformation between two given partially overlapped point clouds [3]. It is indispensable yet challenging in many computer vision and robotics areas, e.g. robot and object pose estimation, point cloud-based odometry and mapping, LiDAR SLAM, 3D reconstruction [4]. Up to now, the most commonly used methods for point cloud registration are non-deep learning based methods. One is handcrafted feature-based method, finding correspondences through well-designed descriptors (such as SI [5], 3DSC [6], FPFH [7], SHOT [8], RoPS [9], LoVS [10], 3DHoPD [11]), and then eliminating mismatches by a random sample consensus (RANSAC) [12] or least square (LS) algorithm. The other is global optimal methods (such as the Go-ICP [13], the convex relaxation [14], and the mixed-integer programming [15]), attempt to find a good optima with the ICP [16]. However, these feature-based methods require handcrafted descriptor, which is of low efficiency and flexibility [17], while most global optimal methods are time consuming, rendering them unsuitable for real time applications [18]. Researches on learning-based registration started relatively later than other tasks (such as classification and segmentation), and there has been a growing interest in recent three years. A set of networks were proposed, such as DCP [19], PRNet [20], RPMNet [21], PointNetLK [18], PCRNet [22], which could automatically learn shape representations. However, most of them are of either low efficiency or low precision, and poorly perform in partial-to-partial registration. A comprehensive literature will be given in Section.2.

* Corresponding author
E-mail address: jws@whu.edu.cn (W. Jiang).

https://doi.org/10.1016/j.patrec.2021.08.002
0167-8655/

Recently, various attention mechanisms, especially self-attention and cross-attention, have sprung up in NLP and image processing, and show better performances than traditional neural networks. For example, Vaswani et al. [23] proposed a solely self-attention based network to compute representations by relating different positions of a single sequence, which could be trained significantly faster than recurrent or convolutional architectures for translation tasks; Wang et al. [24] proposed a non-local network with the self-attention mechanism to efficiently and dynamically focus on cores of the image through the attention weight, which showed high accuracy in object detection, segmentation and pose estimation; Hou et al. [25] introduced a cross-attention module to highlight the target object regions for few-shot classification by generating cross-attention maps between class features and query sample features, which is effective and computationally efficient; Chen et al. [26] proposed a dual-branch transformer for image classification, where a cross-attention module is used in fusion to exchange information with the other branch by linear-time generation of the cross-attention maps. Thus, inspired by those, we introduce spatial self-attention and channel cross-attention mechanisms into partial-to-partial point cloud registration, named SCANet. It is worth noting that our network is fully differentiable and directly processes point clouds without iteration, handcrafted features, voxelization, region querying and correspondences, resulting in computational efficiency and robustness to noise and occlusion. To sum up, our key contributions are as follows:

- A spatial self-attention aggregation (SSA) module is applied in the feature extraction sub-network to efficiently make use of the inter and global information of point clouds in different levels;
  A channel cross-attention regression (CCR) module is innovatively adopted in the pose estimation sub-network for information interaction between two input features, enhancing relevant information and suppressing redundant information, which can greatly improve the registration accuracy and reduce a large number of weights;
  The method achieves state-of-the-art performances by a thorough experimental validation, compared against traditional methods and latest learning-based methods;
  We release our code to facilitate reproducibility and future research.

2. Related works

As mentioned above, learning-based methods for point cloud registration have sprung up mainly in the past three years, for example, DCP [19], PRNet [20], RPMNet [21], PointNetLK [18], PCRNet [22]. According to whether correspondences are estimated or not, existing learning-based methods can be divided into two categories: (1) correspondence-based methods, usually composed of keypoint detection, feature extraction, correspondence matching and registration; and (2) non-correspondence-based methods, adopting a global optimal procedure without expensive correspondence computation.

Correspondence-based: DeepVCP [27] was the first end-to-end deep neural network for point cloud registration based on learned matching probabilities among a group of candidates with a learning-based keypoint detector. Deep Closest Point (DCP) [19] used a transformer network to incorporate global and inter point cloud information, and an attention-based module with a pointer generation layer to predict correspondences between two point clouds. However, it was hard to handle partial-to-partial point cloud registration [20]. PRNet [20] was well-designed framework for partial point cloud registration, where a Gumbel-Softmax with a straight-through gradient estimator were used to sample keypoint correspondences, and distant point clouds were coarsely matched by a diffuse (fuzzy) matching. RPMNet [21] was a less sensitive to initialization and more robust deep learning-based approach, where a differentiable Sinkhorn layer and annealing were applied to get soft assignments. This method handled missing correspondences and point clouds with partial visibility, but it used an iterative inference pipeline to achieve high precision, and required additional normal information. 3DRegNet [28] was a deep neural network to classify point correspondences into inliers/outliers, and regress the motion parameters with a Procrustes approach. IDAM [29] presented an iterative distance-aware similarity matrix convolution layer to find correspondences based on the entire geometric features and Euclidean offset, which could improve computational efficiency and reduce false positive correspondences. Besides, it could be easily integrated with both traditional (e.g. FPFH [7]) and learning-based features. CorsNet [30] fed global features from PointNet [31] to per-point local features to make effective use of point cloud information, and then the correspondences were assigned by fully connected layers, finally a rigid transform was estimated by SVD (Singular Value Decomposition).

Non-correspondence-based: PointNetLK [18] opened up new paths for learning based point cloud registration, utilizing PointNet [31] to compute a global representation and then optimizing the transforms by a modified Lucas Kanade (LK) [32] form in iteration. However, it heavily relied upon the estimation of a gradient through finite differentiation, which is inherently ill-conditioned and highly sensitive to the step-size choice [33]. Inspired by [18], PCRNet [22] presented a fully differentiable framework that used the PointNet [31] representation to align point clouds with fully connected (FC) layers. Experiments showed that it is robust to noise and initial misalignment in data. Further, a deterministic derivation of PointNetLK [33] was advocated, allowing for the derivation of an analytical Jacobian matrix that can be decomposed into “feature” and “warp” components. This approach well solved the inherent memory and efficiency issues taken by PointNetLK [18]. Deep-3Daligner [34] introduced a new Spatial Correlation Representation (SCR) feature optimizer with a transformation decoder network, which were jointly updated towards the minimization of an unsupervised alignment loss. This method was of high accuracy but low efficiency.

In short, the majority of reported learning-based researches were focused on correspondence-based registration, which could achieve high accuracy but low efficiency. Non-correspondence based methods with few studies reported gives rise to substantial advantages in robustness and efficiency, however, most of them used a iteration process to pursuit higher accuracy.

3. Method

3.1. Architecture

A diagram of SCANet is shown in Fig.1. The SCANet is composed of a feature extraction sub-network and a pose estimation sub-network. The point clouds obtained from a sensor are referred to as the source, while the point clouds of the known model of the object are regarded as the target.

In the feature extraction sub-network, different from an original PointNet [31] adopted in the PointNetLK [18] and the PCRNet [22], a spatial self-attention aggregation (SSA) module is adopted to simultaneously make use of the inter and global information of each point cloud. Two input point clouds are firstly upsampled into 64 channels, and then fed into three spatial self-attention blocks with a size of (64, 64, 128). Outputs of three blocks are aggregated to represent information in different levels. Finally, a k max-pooling function is respectively used to select k (k=4) points with the most

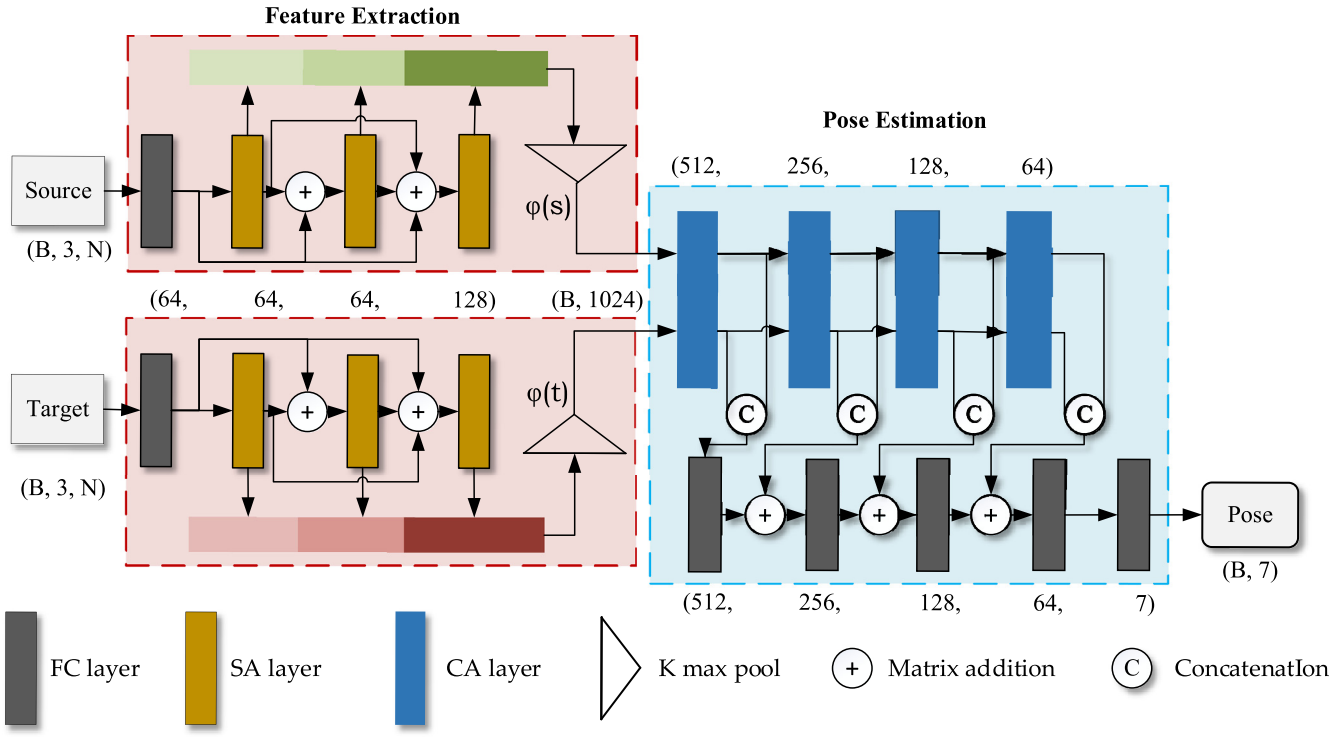

Fig. 1. SCANet architecture. The source and target point clouds are firstly fed into a spatial self-attention aggregation (SSA) module with a size of (64, 64, 64, 128) to extract pointwise features with a size (*B*, 256, *N*), which is arranged in a Siamese structure with shared weights; following, a *k* max-pooling function (*k*=4) is used to find the global feature vectors φ(s) and φ(t) with a size (B, 1024). Secondly, two global features are given into a channel cross-attention regression (CCR) module with a size of (512, 256, 128, 64, 7) to predict the pose with a size 7, where the first four values of *T* represent the rotation quaternion $q \in R^4$, $q^T q=1$ and last three represents the translation vector $t \in R^3$.

distinguish features of two point clouds, forming a global feature vectors φ(s) and φ(t) with a size (1024, 1).

In the pose estimation sub-network, different from a modified LK algorithm [32] in the PointNetLK [18] and all fully connected (FC) layers in the PCRNet [22], a channel cross-attention regression (CCR) module is proposed for information exchange between two features during regression, enhancing relevant information and suppressing redundant information. Given two feature vectors of the source and target point clouds, four channel cross-attention blocks with a size (512, 256, 128, 64, 32) and five FC layers with a size (512, 256, 128, 64, 32, 7) are applied to estimate the transformation T, where the first four values of T represent the rotation quaternion $q \in R^4$, $q^T q=1$ and last three represents the translation vector $t \in R^3$.

3.2. Feature Extraction

Many learning-based registration adopted a PointNet [31] to extract high dimensional information of each point, however, it lacks geometric information. Although PointNet++ [35] could well solve above problems, but it involves a lot of irregular accesses, resulting in low computational efficiency. As the self-attention can efficiently capture global information [24], therefore, we adopt a spatial self-attention aggregation (SSA) module to simultaneously and efficiently utilize the inter and global information of each point cloud in different levels.

As shown in above Fig.1, the SSA module mainly includes four components: (1) a basic convolution to upsample the channel from 3 to 64; (2) three self-attention blocks to extract the inter and global information of each point clouds; (3) a feature aggregation to obtain features in different levels. (4) a *k* max pooling to select *k* points with the most distinguish features, forming a global feature vectors φ. Among them, the core is the self-attention mechanism.

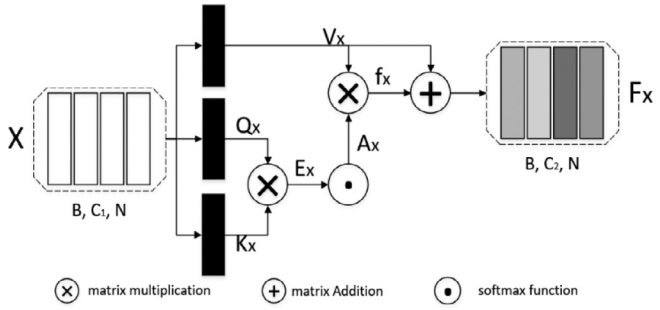

Fig. 2. The self-attention mechanism: The shape of input features *X* is (B, C_1, N), where *B* is the batch size, C_1 is the channels and *N* is the point number.

As shown in Fig.2, the operations of a self-attention mechanism [24] on feature map are mainly divided into three categories: *query* (*Q*), *key* (*K*), and *value* (*V*). Specifically, given a source feature map *X*, by interactively multiplying the *query* (*Q*) in row *i* and the *key* (*K*) in column *j*, a self-attention map (A_x) can be obtained after a softmax function Eq. 1-(2). Secondly, by respectively multiplying the *value* (*V*) and the attention map, the attention-based feature maps (f_x) are obtained as Eq. 3. It is note that, for simplification, the operations of *query, key* share the weights.

$$Q_x = (W_a X)^T, K_x = Q_x^T, \ V_x = W_c X \tag{1}$$

$$E_x = Q_x K_x^T, A_x = softmax(E_x) \tag{2}$$

$$f_x = V_x A_x, F_x = V_x + \alpha f_x \tag{3}$$

Where W_a, W_c denote the weights, α is a learnable weight.

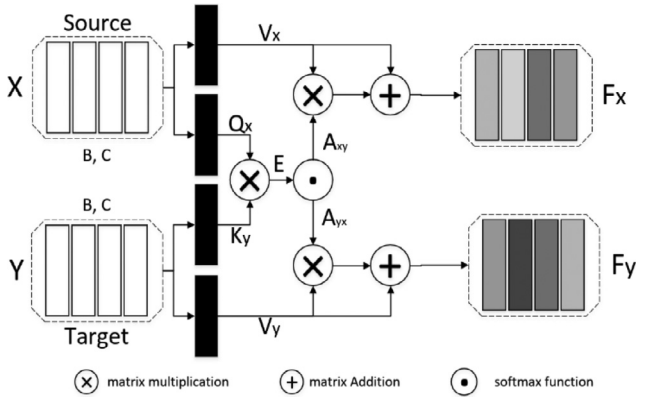

Fig. 3. The cross-attention mechanism: The shape of input features *X, Y* is (*B, C*), where *B* is the batch size, *C* is the channels.

Through such a SSA structure, the point clouds can be directly extracted the inter and global information of each point cloud in different levels. It does not involve any query of the neighborhood of the point, not generating an irregular access.

3.3. Pose Estimation

In the pose estimation, PointNetLK [18] adopted a modified Lucas Kanade (LK) algorithm [32] to iteratively predict transformations, but it heavily relied upon the estimation of a gradient through finite differentiation; PCRNet [22] adopted fully connected layers, however, there are a large number of weights. For registration, the information of overlapping regions is significant, while the information of other regions will even interfere with the final result. Considering that cross-attention can highlight more relevant information by generating cross attention maps for two input features [25], hence, we innovatively adopt a channel cross-attention regression (CCR) module in the pose estimation sub-network for information interaction between two feature vectors.

As shown in Fig.1, the CCR module is composed of two branches: (1) the upper line is four channel cross-attention blocks for information interaction between two features during regression, enhancing relevant information and suppressing redundant information; (2) the bottom line is five FC layers, concatenating both source and target representations for pose regression. Among them, the key is the channel cross-attention mechanism.

$$Q_x = W_a X,\ K_y = W_a Y \tag{4}$$

$$V_x = W_b X,\ V_y = W_b Y \tag{5}$$

$$E = (W_a X)^T (W_a Y) \tag{6}$$

$$A_{xy} = softmax(E),\ A_{yx} = softmax\left(E^T\right) \tag{7}$$

$$F_x = V_x + aA_{xy}V_x, F_y = V_y + aA_{yx}V_y \tag{8}$$

Where W_a denotes the weights, α is a learnable weight.

The channel cross-attention mechanism achieves simultaneously information interaction between the source and target representations by calculating the cross-attention map at all channels. As shown in Fig.3, we divide the input source feature X into *query* and *value* of X, and the input target feature Y into *key* and *value* of Y. Then, the energy matrix between X and Y is calculated by Eq.6. The cross-attention matrix A_{xy} transforms the source attention space to the target attention space (vice versa for A_{yx}) by Eq.7.

Based on the above attention weights, the attention-based feature vectors of the source and target are calculated as Eq.8. It is noted that, for simplification, the operations of *query, key* and *value* share the weights.

Thus, through the channel cross-attention mechanism, attention will be gradually focused on much more relevant information, while redundant features will be gradually suppressed, greatly improving the registration accuracy and saving a large number of network parameters.

3.4. Loss Function

A transformation matrix error between the predicted transformation and the truth transformation is considered in the definition of loss to train the network, including a rotation error *err(R)* and a translation error *err(t)*. The rotation and translation errors are defined as Eq.9, where q_1 and q_2 are quaternions of the predicted and the truth rotation matrices R_1 and R_2, respectively, and t_1 and t_2 are the predicted and the truth translation, respectively. By minimizing the loss, the pose parameters are directly predicted through the network.

$$err(R) = ||q_1 - q_2||_2^2, err(t) = ||t_1 - t_2||_2^2 \tag{9}$$

$$Loss = err(R) + err(t) \tag{10}$$

4. Experiments

4.1. Experimental Setup

Following experiments are carried on the ModelNet40 dataset [36], including 9843 training shapes and 2468 testing shapes from 40 object categories. Train data are repeated five times for data augmentation with a rotation perturbation. For a given shape, we randomly sample 1024 points to form a point cloud with randomly generate rotations within [0°, 45°] and translation in [−0.5, 0.5]. To generate partially overlapped point clouds, similar to [21], we fix a random point far away from the point clouds, and preserve 717 points (approximately 70%) closest to the far point for each point cloud. To generate noise point clouds, we randomly jitter the points in both point clouds by noises sampled from $N(0, 0.01)$ and clipped to [-0.05, 0.05] on each axis.

The networks are trained with a Adam optimizer for 250 epochs, using a cosine annealing schedule with the original learning rate 0.001 and the minimum learning rate 0.000001. The network parameters are updated on a single NVIDIA GeForce GTX 1080 Ti GPU and an Intel(R) Xeon(R) CPU E5-2630 v4 at 2.20GHz.

For quantitatively evaluation, similar to [20], we measure mean squared error (MSE), root mean squared error (RMSE), mean absolute error (MAE), and coefficient of determination (R^2) in both rotation and translation. Errors of rotation are in units of degrees. For comparison, a set of experiments are carried out with traditional methods (ICP [16], Go-ICP [13], FPFH [7]+RANSAC [12], FGR [37] DCP [19]) and latest learning based networks (PRNet [20], RPMNet [21], PointNetLK [18], PCRNet [22], IDAM [29]). It is worth noting that, in following tables, the data in blue are quoted from [20], while the data in green are quoted from [29]. The results of PCRNet [22] and RPMNet [21] are obtained from their provided source code or trained weights.

4.2. Partial Visibility with Gaussian

Following experiment tests the ability of our SCANet on partial visibility with Gaussian noise. As listed in Tab.1, it is obvious that, with occlusion and noise interfered, our SCANet greatly exceeds not only traditional methods (FPFH [7]+RANSAC [12], ICP

Table 1
Test on partially visible point clouds with Gaussian noise.

Model	MSE(R)↓	RMSE(R)↓	MAE(R)↓	R2(R)↑	MSE(t) ↓	RMSE(t)↓	MAE(t)↓	R2(t)↑	Time*(s)
ICP [16]	1229.670	35.067	25.564	-6.252	0.0860	0.294	0.250	-0.045	0.095
Go-ICP [13]	150.320	12.261	2.845	0.112	0.0008	0.028	0.029	0.991	/
FGR [37]	764.671	27.635	13.794	-3.491	0.0048	0.070	0.039	0.941	0.123
PointNetLK [21]	397.575	19.939	9.076	-1.343	0.0032	0.057	0.032	0.960	0.082
DCP-v2 [19]	47.378	6.883	4.534	0.718	0.0008	0.028	0.021	0.991	0.015
PRNet [20]	18.691	4.323	2.051	0.889	0.0003	0.017	0.012	0.995	0.022
FPFH [7]+RANSAC [12]	25.604	5.06	4.19	/	0.0004	0.021	0.018	/	0.159
FPFH [7]+IDAM [29]	201.924	14.21	7.52	/	0.0045	0.067	0.042	/	0.050
GNN [38]+IDAM [29]	13.838	3.72	1.85	/	0.0005	0.023	0.011	/	0.026
RPMNet [21]	10.778	3.283	1.625	0.940	0.0009	0.0304	0.0167	0.989	0.063
PCRNet [22]	27.0809	5.2039	3.5269	0.8485	0.0032	0.0567	0.0372	0.9616	0.017
SCANet	15.9613	3.9952	2.4485	0.9111	0.0013	0.0363	0.0236	0.9843	0.023

Time* is the speed on point clouds with 1024, and it is measured in seconds-per-frame. The data in orange are quoted from [28].

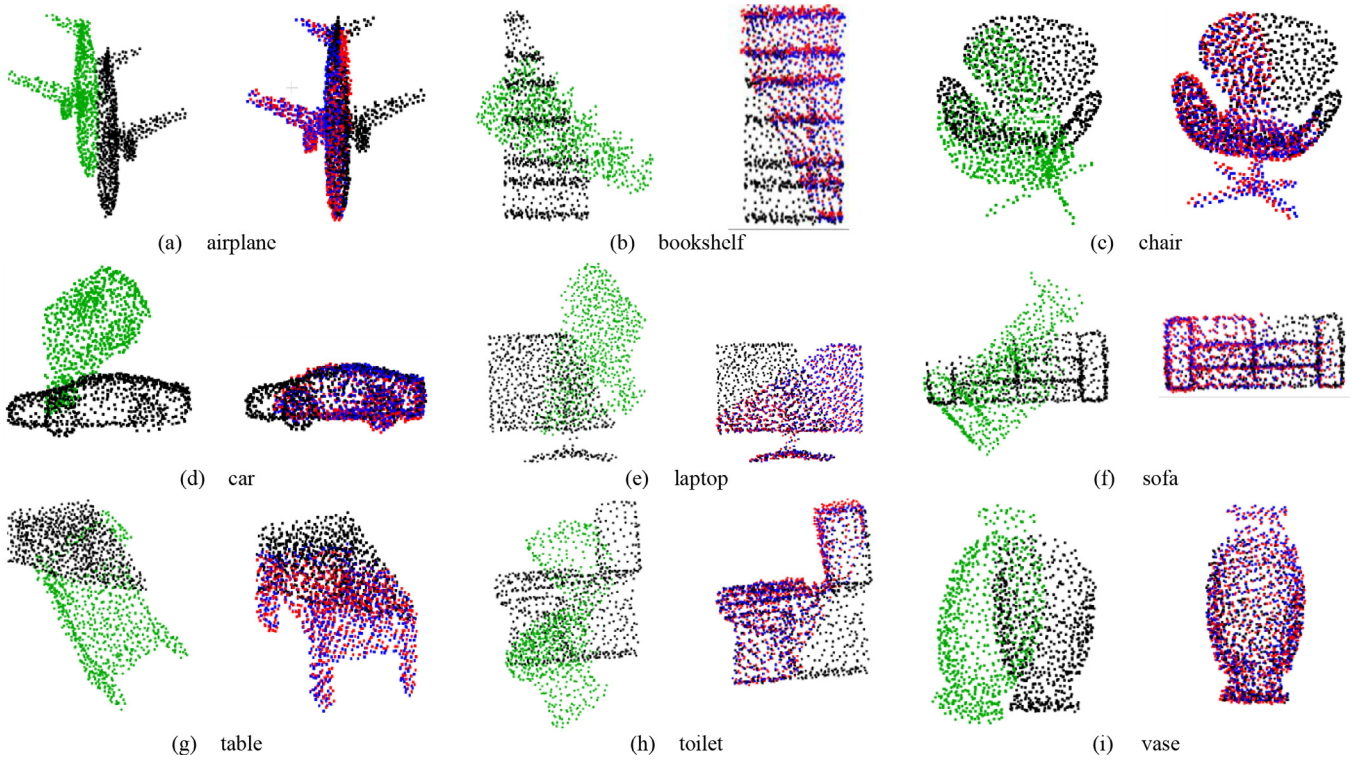

Fig. 4. Example results of our method on partially visible data with Gaussian noise. The original source points in green, target points are in black, ground truth points are in blue, and predicted points are in red.

[16], Go-ICP [13]), but also latest learning-based networks (PRNet [20], PointNetLK [18], PCRNet [22]), especially in rotation. However, there is still a little gap between our method and GNN [38]+ IDAM [29] and RPMNet [21]. It is worth noting that GNN [38]+ IDAM [29] and RPMNet [21] are both correspondence-based methods, which mean a long processing flow. And both two methods require several iterations to achieve such good performances (three iterations for GNN [38]+ IDAM [29], and five iterations RPMNet [21]). Furthermore, RPMNet [21] used additional handcrafted features (XYZ, dxyz and PPF) as inputs, while the input of our SCANet is only the coordinate information of point clouds, not containing any other information and involving no iteration. Example results of our method on partially visible data with Gaussian noise are shown in Fig.4.

4.3. Noise

Following experiment tests the ability of our SCANet on partial visibility (completeness = 70%) with varying Gaussian noise. It is worth noting that we use the trained model in Section.B (noise = 0.01, completeness = 70%) to predict. We randomly jitter the predicted points in both point clouds by noises sampled from $N(0, 0.01)$, $N(0, 0.02)$, $N(0, 0.03)$, $N(0, 0.04)$, $N(0, 0.05)$ and respectively clipped to [-0.05, 0.05], [-0.1, 0.1], [-0.15, 0.15], [-0.2, 0.2], [-0.25, 0.25] on each axis.

The registration accuracy is listed in Tab.2 and example results are shown in Fig.5. It is obvious that, with the increase of noise, the shapes of point clouds are deformed, and the differences between the source and target point clouds become larger, resulting in both rotation and translation accuracy of our SCANet slightly decreasing. However, it is still in good performances. It shows that our SCANet is robust to noise to some extent, owing to the aggregation of inter and global information by the SSA module.

4.4. Overlap

Following experiment tests the ability of our SCANet on noisy (noise = 0.01) point clouds with varying overlaps. It is worth not-

Table 2
Test on partially visible point clouds with varying Gaussian noise.

noise	MSE(R)↓	RMSE(R)↓	MAE(R)↓	R2(R)↑	MSE(t) ↓	RMSE(t)↓	MAE(t)↓	R2(t)↑
0.01	15.9613	3.9952	2.4485	0.9111	0.0013	0.0363	0.0236	0.9843
0.02	16.5131	4.0636	2.5424	0.9080	0.0014	0.0369	0.0243	0.9837
0.03	17.3978	4.1711	2.6825	0.9031	0.0015	0.0384	0.0256	0.9824
0.04	18.7338	4.3283	2.8655	0.8955	0.0016	0.0406	0.0276	0.9804
0.05	20.5612	4.5344	3.0867	0.8852	0.0019	0.0436	0.0301	0.9774

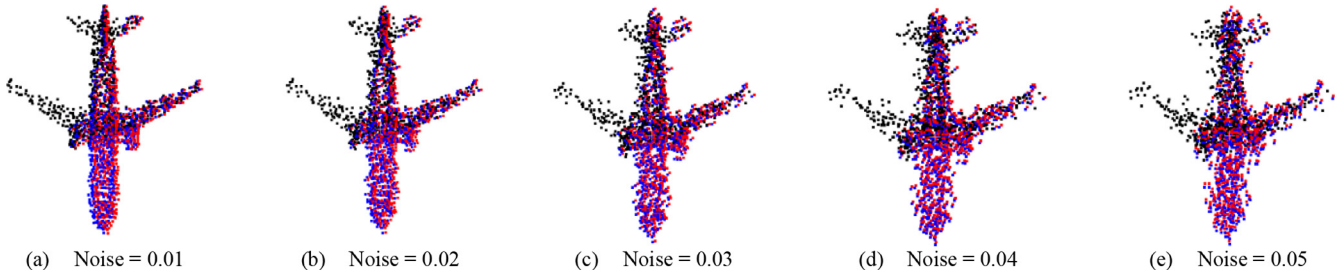

Fig. 5. Example results of our method on partially visible data with varying Gaussian noise: The target points are in black, ground truth points are in blue, and predicted points are in red.

Table 3
Test on noisy point clouds with varying overlaps.

completeness	MSE(R)↓	RMSE(R)↓	MAE(R)↓	R2(R)↑	MSE(t) ↓	RMSE(t)↓	MAE(t)↓	R2(t)↑
75%	14.6768	3.8310	2.3127	0.9183	0.0012	0.0340	0.0223	0.9862
70%	15.9613	3.9952	2.4485	0.9111	0.0013	0.0363	0.0236	0.9843
65%	18.1313	4.2581	2.6141	0.8990	0.0017	0.0414	0.0274	0.9796
60%	20.2838	4.5038	2.8310	0.8868	0.0025	0.0499	0.0335	0.9703
55%	22.4535	4.7385	3.0719	0.8744	0.0040	0.0635	0.0426	0.9519

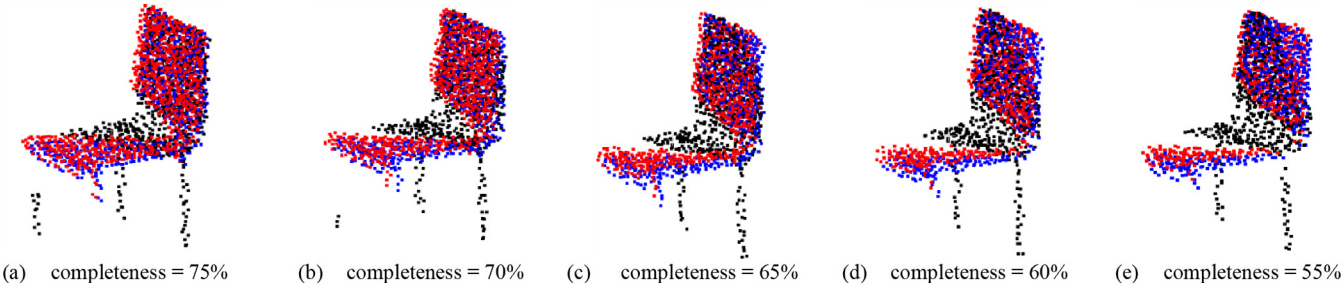

Fig. 6. Example results of our method on noisy data with varying overlaps: The target points are in black, ground truth points are in blue, and predicted points are in red.

ing that we use the trained model in Section.B (noise = 0.01, completeness = 70%) to predict. We randomly sample planes that cut away 25%, 30%, 35%, 40%, 45% of the predicted points to obtain point clouds with 75%, 70%, 65%, 60%, 55% completeness.

The registration accuracy is listed in Tab.3 and example results are shown in Fig.6. It shows that, with the completeness decreasing, the overlap regions become smaller and the information for registration is less. As a consequence, when the completeness reduces from 75% to 60%, the registration accuracy of our SCANet slightly decrease, but from 60% to 55%, the precision drops sharply, especially in translation. It shows that our SCANet can well handle small overlaps, thanks to the information interaction between two representations by the CCR module. However, it poorly performs when the overlap is too small.

4.5. Ablation Studies

To better understand how various choices affect the performance of the network, we compare our SSA module with the PointNet [31] in the feature extraction sub-network, and the proposed CCR with the FC layers in the pose estimation sub-network. It worth noting that the PointNet [31] with FC is the PCRNet [22]. All studies in this section are evaluated on the partial visibility (completeness = 70%) with Gaussian noise (noise = 0.01).

The results are listed in Tab.4. By comparing Model A&B, it is found that the SSA can improve the registration accuracy in both rotation and translation, which is mainly attributed to the fact that the proposed SSA module makes full use of the inter and global information of each point cloud. Through the comparison between Model A&C, it shows that the CCR can increase the regression accuracy in rotation, meanwhile, it can save a large number of network parameters. This is benefited from that the CCR allows information interaction between two input representations, which can enhance relevant information and suppress redundant information. It is obvious that Model D combines the superiority of both SSA and CCR modules, making great improvements in both rotation and translation and saving a large number of weights.

5. Conclusion

This work presents a simple, novel and efficient network, named SCANet, based on spatial and channel attention mechanisms for partial point cloud registration. We creatively propose two new modules: (1) a spatial self-attention aggregation (SSA)

Table 4
Test on unseen point clouds with Gaussian noise of different combinations.

Mode	Feature Extraction PointNet	Feature Extraction SSA	Pose estimation FC	Pose estimation CCR	Weights↓	Time(s)↓	MSE(R)↓	RMSE(R)↓	MAE(R)↓	R2(R)↑	MSE(t)↓	RMSE(t)↓	MAE(t)↓	R2(t)↑
A	√		√		4216653	34	27.0809	5.2039	3.5269	0.8485	0.0032	0.0567	0.0372	0.9616
B		√	√		4138186	43	22.5121	4.7447	3.1351	0.8744	0.0021	0.0461	0.0305	0.9747
C	√			√	1542353	39	17.2144	4.1490	2.5774	0.9040	0.0018	0.0422	0.0266	0.9787
D		√		√	1463886	48	15.9613	3.9952	2.4485	0.9111	0.0013	0.0363	0.0236	0.9843

module to efficiently extract the inter and global information of each point cloud; (2) a channel cross-attention regression (CCR) module in the pose estimation for information interaction and redundant information suppressing between two input features. Our SCANet is fully differentiable and directly process point clouds without handcrafted features, voxelization, region querying and correspondences, resulting in computational and storage efficiency and robustness to noise and occlusion.

However, at present, we still deal with point clouds with small number in the simple scene. How to extend our method to point clouds with large scale and complex scenarios, especially the vehicle point cloud with large and inhomogeneous density, is a core of our follow-up research.

Declaration of Competing Interest

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

Acknowledgments

This work is supported by the National Natural Science Foundation of China (Grant Nos. 91638203, 91738302).

References

- [1] Z. Liang, Y. Guo, Y. Feng, W. Chen, L. Qiao, L. Zhou, J. Zhang, and H. Liu, “Stereo matching using multi-level cost volume and multi-scale feature constancy,” IEEE TPAMI, 2019.
- [2] Y. Guo, H. Wang, Q. Hu, H. Liu, L. Liu and M. Bennamoun, " Deep Learning for 3D Point Clouds: A Survey," in IEEE Transactions on Pattern Analysis and Machine Intelligence, doi: 10.1109/TPAMI.2020.3005434.
- [3] B. Eckart, K. Kim, J. Kautz, HGMR: Hierarchical Gaussian Mixtures for Adaptive 3D Registration, in: Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 705–721.
- [4] W. Lu, G. Wan, Y. Zhou, X. Fu, P. Yuan, S. Song, arXiv preprint, 2019.
- [5] A.E. Johnson, M. Hebert, Using spin images for efficient object recognition in cluttered 3D scenes, IEEE Trans. Pattern Anal. Mach. Intell. 21 (5) (1999) 433–449.
- [6] A. Frome, D. Huber, R. Kolluri, et al., Recognizing objects in range data using regional point descriptors, in: European Conference on Computer Vision, Springer, Berlin Heidelberg, 2004, pp. 224–237.
- [7] R.B. Rusu, N. Blodow, M. Beetz, Fast point feature histograms (FPFH) for 3d registration, in: Robotics and Automation, 2009. ICRA'09. IEEE International Conference on. IEEE, 2009, pp. 3212–3217.
- [8] F. Tombari, S. Salti, L. Distefano, Unique signatures of histograms for local surface description, in: European Conference on Computer Vision, Springer, Berlin Heidelberg, 2010, pp. 356–369.
- [9] Y. Guo, F. Sohel, M. Bennamoun, et al., Rotational projection statistics for 3D local surface description and object recognition, Int. J. Comput. Vision 105 (1) (2013) 63–86.
- [10] S. Quan, J. Ma, F. Hu, B. Fang, T. Ma, Local voxelized structure for 3D binary feature representation and robust registration of point clouds from low-cost sensors, Inf. Sci. (Ny). 444 (2018) 153–171, doi:10.1016/j.ins.2018.02.070.
- [11] S.M. Prakhya, J. Lin, V. Chandrasekhar, W. Lin, B. Liu, ‘3DHoPD: A fast low-dimensional 3-D descriptor, IEEE Robot. Autom. Lett. 2 (3) (Jul. 2017) 1472–1479.
- [12] M.A. Fischler, R.C. Bolles, Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography, Communications of The ACM (1981).
- [13] J. Yang, H. Li, Y. Jia, Go-ICP: Solving 3d registration efficiently and globally optimally, in: 2013 IEEE International Conference on Computer Vision (ICCV), Dec 2013, pp. 1457–1464. pages.
- [14] Haggai Maron, Nadav Dym, Itay Kezurer, Shahar Kovalsky, Yaron Lipman, Point registration via efficient convex relaxation, ACM Transactions on Graphics (2016).
- [15] Tom Guerout, Yacine Gaoua, Christian Artigues, Georges Da Costa, Pierre Lopez, Thierry Monteil, Mixed integer linear programming for quality of service optimization in clouds, Future Generation Computer Systems (2017).
- [16] P.J. Besl, N.D. McKay, A method for registration of 3d shapes, IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) 14 (2) (1992) 239–256, doi:10.1109/34.121791.
- [17] H. Wang, L. Wang, Beyond Joints: Learning Representations from Primitive Geometries for Skeleton-Based Action Recognition and Detection, IEEE Trans. Image Process. 27 (2018) 4382–4394, doi:10.1109/TIP.2018.2837386.
- [18] Y. Aoki, H. Goforth, R.A. Srivatsan, S. Lucey, PointNetLK: robust & efficient point cloud registration using PointNet, in: Proceedings of the IEEE Conference on Com- puter Vision and Pattern Recognition (CVPR), 2019, pp. 7163–7172. pages.
- [19] Y. Wang, J.M. Solomon, Deep closest point: Learning representations for point cloud registration, in: Proceedings of the IEEE International Conference on Com- puter Vision, 2019, pp. 3523–3532.
- [20] Y. Wang, J.M. Solomon, Prnet: Self-supervised learning for partial-to-partial registration, in: Advances in Neural Information Processing Systems, 2019, pp. 8812–8824.
- [21] Zi Jian Yew, Gim Hee Lee, Rpm-net: Robust point matching using learned features, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 11824–11833. pages.
- [22] V. Sarode, X. Li, H. Goforth, Y. Aoki, R.A. Srivatsan, S. Lucey, H. Choset, arXiv preprint, 2019.
- [23] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszko- reit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, Illia Polosukhin, Attention is all you need, in: Advances in Neural Information Processing Systems, 2017, pp. 5998–6008. pages.
- [24] Xiaolong Wang, Ross Girshick, Abhinav Gupta, Kaiming He; Non-local Neural Networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 7794–7803.
- [25] R Hou, H Chang, B Ma, et al., Cross Attention Network for Few-shot Classification, NeurIPS (2019).
- [26] C F Chen, Q Fan, R. Panda, arXiv preprint, 2021.
- [27] Weixin Lu, Guowei Wan, Yao Zhou, Xiangyu Fu, Pengfei Yuan, Shiyu Song, Deepvcp: An end-to-end deep neural network for point cloud registration, in: IEEE Int'l Conf. Computer Vision (ICCV), 2019, pp. 3523–3532. pages.
- [28] G.D. Pais, S. Ramalingam, V.M. Govindu, J.C. Nascimento, R. Chellappa, P. Miraldo, 3DRegNet: a deep neural network for 3D point registration, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 7193–7203. pages.
- [29] J. Li, C. Zhang, Z. Xu, H. Zhou, C. Zhang, Iterative Distance-Aware Similarity Matrix Convolution with Mutual-Supervised Point Elimination for Efficient Point Cloud Registration, European Conference on Computer Vision (ECCV), 2020.
- [30] Akiyoshi Kurobe, Yusuke Sekikawa, Kohta Ishikawa, Hideo Saito, Corsnet: 3d point cloud registration by deep neural network, IEEE Robotics and Automation Letters 5 (3) (2020) 3960–3966.
- [31] C.R. Qi, H. Su, K. Mo, L.J. Guibas, PointNet: Deep learning on point sets for 3D classification and segmentation, in: CVPR, 2017, pp. 652–660.
- [32] D.D. Lucas, An iterrative image registration technique with an application to stereo vision, in: Proc. of Imaging Understanding Workshop, 1981.
- [33] Xueqian Li, Jhony Kaesemodel Pontes, Simon Lucey. Deterministic PointNetLK for Generalized Registration. 2020.
- [34] Wang, L.; Li, X.; Fang, Y. Deep-3DAligner : Unsupervised 3D Point Set Registration Network With Optimizable Latent Vector. 2020, 1–7.
- [35] C.R. Qi, L. Yi, H. Su, L.J. Guibas, PointNet++: Deep hierarchical feature learning on point sets in a metric space, in: NeurIPS, 2017, pp. 5099–5108.
- [36] Z. Wu, S. Song, A. Khosla, F. Yu, L. Zhang, X. Tang, J. Xiao, 3d shapenets: A deep representation for volumetric shapes, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 1912–1920.
- [37] Q.-Y. Zhou, J. Park, V. Koltun, Fast global registration, in: European Conference on Computer Vision, Springer, 2016, pp. 766–782. pages.
- [38] Z. Wu, S. Pan, F. Chen, G. Long, C. Zhang, P.S. Yu, arXiv preprint, 2019.