GPNF:A Point Cloud Registration Framework Using Sharp Global Linear Attention Prior and Neighborhood Filtering Strategy

Congyang Zhu, Mengxiao Yin✉, Junjie Liao, Zhijie Liang, and Kan Chang

Guangxi University, Nanning, China
zcy1234562022120163.com, ymx@gxu.edu.cn

Abstract. Robust point features are essential when registering point cloud scenes with numerous instances. To enhance the point features, we propose KPConvFormer module. It leverages the advantages of attention mechanisms to focus on important features, considers the feature and position differences among points in point convolution simultaneously, and pre-weights the neighborhood points in the convolution region. The pre-weight process filters out irrelevant points from other instances near the boundaries and noisy points within the convolution region, correcting the point convolution factors in each neighborhood to help aggregate more accurate point features. Addressing the incorrect registration caused by the similar structure of point clouds, we designed a Shareped-Linear-Self-Attention module. It learns a sharp global prior, efficiently capturing fine-grained global structural information. This module distinguishes similarity structures in the point clouds to be registered from a larger receptive field, providing a global prior for subsequent convolution operations. Compared to existing state-of-the-art methods, our approach achieves superior performance on most registration metrics across the 3DMatch, 3DLoMatch, and KITTI datasets.

Keywords: Point Cloud Registration · 3D Representation Learning · Attention Mechanism.

1 Introduction

Point cloud registration is a fundamental task in graphics, vision, and robotics. Given two partially overlapping 3D point clouds, point cloud registration estimates a rigid transformation to align two input point clouds. With the rapid development of 3D point representation learning, this task has regained significant attention.

Point cloud registration methods based on correspondence [4,11,13,16,19,55] train neural networks to extract point correspondences between two input point clouds and then use a robust estimator (e.g., RANSAC) to compute the alignment transformation. Following this scheme, Qin et al.[37] downsampled the input point clouds into superpoints and then match them through testing whether their local neighborhood overlaps. These matched superpoints are propagated

to individual points, generating dense point correspondences for estimating the transformation.

The accuracy of the above registration method [37] mainly depends on the results of superpoint matching. We design a well-generalizing point convolution method called KPConvFormer to obtain robust superpoints and achieve accurate matching. When learning complex point cloud scene representations, irrelevant points from other instances or noise may be introduced within a convolution neighborhood. The generalization theory of CNNs indicates that points with stronger feature correlations should be included in the same neighborhood [26]. Inspired by this theory, we utilize the idea of attention in KPConv[41] to perceive the positional and feature differences of convolutional points, achieving high-dimensional modeling among points. The computed difference values are used for calculating weighting scores to pre-weight the neighborhood points so that the subsequent KPConv[41] can focus on more critical neighborhood points.

Due to the locality of convolution operations, most convolution-based point cloud representation learning methods have limited capability in capturing long-range dependencies, whereas dependencies over a more extensive scope are crucial for learning global structural information, especially for real-world captured data [43]. When dealing with point cloud data containing numerous similar structures, capturing long-range dependencies and learning broader global contextual information can help distinguish these structures and achieve accurate registration results.

Some vision tasks try to use linear-complexity attention mechanisms to capture long-range dependencies in a lightweight way. However, work [18] points out some existing linear attention approaches may introduce additional computational overhead from complex mapping functions, resulting in degraded model performance. For point cloud registration, some methods attempt to separate point convolution and attention [37,54], first using multiple layers of point convolution to obtain sparse high-dimensional features and then applying attention to these features. These approaches can reduce computational costs, but obtaining sparse point features typically requires a multi-layer convolutional network, and the completeness of the information may not be comparable to the original dense point cloud information.

Inspired by previous work [18], we design an efficient global prior learning structure (Sharp-Linear-Self-Attention) before the KPConvFormer to model the complete global dependencies. This structure captures the global information at a fine-grained level for the dense point cloud, ensuring that each point participating in the subsequent point convolution has clear global prior knowledge, thereby enhancing the performance of the superpoints and dense points used for point cloud registration.

Prior work [18] has indicated that traditional linear attention mechanisms often result in feature distributions that are overly smooth and homogeneous. In our Sharp-Linear-Self-Attention module, we introduce a simple mapping function to adjust the feature directions of queries and keys, leading to a sharper distribution of attention weights. This allows for better aggregation of impor-

tant information. To further enhance the ability of linear attention to capture diverse features, we leverage the spatial locality of independence convolution to adjust the outputs at different positions within the original attention matrix, alleviating the limitations imposed by the original linear attention.

The contributions of our research can be summarized as follows:

- We propose a KPConvFormer module. This module introduces an attention-based neighborhood filtering strategy within KPConv [41], weakening the influence of irrelevant points and noise points in the convolution region. It inherently integrates attention and point convolution operations.
- We propose a Sharp-Linear-Self-Attention module for point clouds. This module captures fine-grained global information, providing prior knowledge for the subsequent point convolution.
- Based on the KPConvFormer module and the Sharp-Linear-Self-Attention module, we design a superpoints and dense points feature extraction network called KPFLAN and integrated it into our built point cloud registration framework, GPNF. Through extensive experiments on the 3DMatch, 3DLoMatch, and KITTI datasets, our framework achieves state-of-the-art results compared to existing methods across most evaluation metrics.

2 Related Work

Correspondence-Based Registration. The correspondence-based methods [12,13,14,16] first obtain the correspondence points between the two point clouds and then use a robust pose estimator (e.g., RANSAC) to recover the transformation. Owing to the robust estimator, they achieve state-of-the-art performance in indoor and outdoor scene registration tasks. Based on the ways of extracting correspondences, these methods can be further categorized into two types. The first type detects more repeatable keypoints [4,19] and learns more powerful descriptors for the keypoints [1,12,44]. The second type [55] retrieves correspondences without keypoint detection by considering all possible matches. Our work follows the detection-free approach and utilizes global priors and neighborhood filtering strategies to improve the accuracy of the correspondences.

Direct Registration Methods. Direct registration methods use a neural network to estimate the transformation in an end-to-end manner. These methods can be categorized into two classes based on how the alignment transformation is estimated. The first class [14,25,45,46,53,57,59] utilizes the idea of ICP [6], iteratively establishing soft correspondences and employing differentiable weighted SVD to compute the transformation. The second class [2,20,49,50] extracts global feature vectors for each point cloud and then regresses the transformation using these global feature vectors. Although promising results have been achieved on single synthetic shapes, direct registration methods can still fail in large-scale scenes, as discussed in [19].

Deep Robust Estimators. Traditional robust estimators, such as RANSAC, exhibit slow convergence and instability when dealing with high outlier ratios. Deep robust estimators [3,10,32] are proposed as alternatives. Deep robust estimators typically consist of a classification network that rejects outliers and an estimation network that computes the transformation. Compared to traditional estimators, they achieve improvements in accuracy and speed but require training specific networks.

Point cloud understanding and attention. Based on the different point cloud modeling methods, the methods of learning point cloud representations can be divided into three categories: projection-based, voxel-based, and point-based methods. Projection-based methods project 3D points onto image planes and use 2D CNNs for feature extraction [9,22,24,40]. Voxel-based approaches convert point clouds into voxel grids for 3D convolutions [31,39], but they often struggle with large kernel sizes. In contrast, point-based methods directly process point clouds [30,33,34,41,60] and have recently shifted towards Transformer-based architectures [17,38,47,48,51,61]. The KPConvFormer module for optimizing point cloud representations utilizes the idea of attention to achieve pre-filtering of the convolution neighborhood, improving the generalization of point cloud representations. At the same time, to efficiently capture long-range dependencies in point cloud scenes, inspired by the use of linear attention in the image domain to solve high-complexity problems [18,21], we design a Sharp-Linear-Self-Attention module at the front-end of the backbone network for obtaining superpoints, efficiently learning complete global priors.

3 Method

Given two point clouds $\boldsymbol{P} = \{p_1, p_2, \ldots, p_n\} \subseteq \mathbb{R}^3$ and $\boldsymbol{Q} = \{q_1, q_2, \ldots, q_m\} \subseteq \mathbb{R}^3$, respectively. Our point cloud registration framework GPNF recovers a transformation consisting of a rotation matrix $\mathbf{R} \in SO(3)$ and a translation vector $\mathbf{t} \in \mathbb{R}^3$ that aligns $\boldsymbol{P}$ and $\boldsymbol{Q}$. The rigid transformation $\mathbf{T} = \{\mathbf{R}, \mathbf{t}\}$ is solved by the following formula:

$$\min_{\mathbf{R},\mathbf{t}} \sum_{(p_i^*, q_j^*) \in C^*} \left\| \mathbf{R} \cdot p_i^* + \mathbf{t} - q_j^* \right\|_2^2, \tag{1}$$

where C^* is the set of ground-truth correspondences between $\boldsymbol{P}$ and $\boldsymbol{Q}$. Since C^* is unknown in reality, we need to first extract the point-by-point features used for matching to obtain point correspondences in $\boldsymbol{P}$ and $\boldsymbol{Q}$ and then estimate the alignment transformation.

We apply the backbone KPFLAN on the left side of Fig. 1 to learn robust point features and build a coarse-to-fine point cloud registration framework GPNF based on the excellent work of Geometric [37]. The pipeline of GPNF is shown in Fig. 1. KPFConvBlock and KPFResBlock transform point clouds $\boldsymbol{P}$ and $\boldsymbol{Q}$ into higher-dimensional feature spaces. The SLN-Block learns and

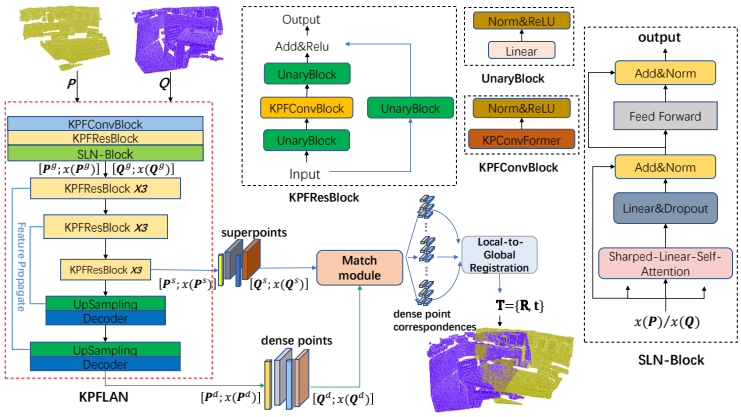

Fig. 1. The pipeline of GPNF. The multi-level Backbone KPFLAN in the left red dashed box learns robust superpoints and dense points features. The KPConvFormer used in the KPFConvBlock and the KPFResBlock is a point convolution module with neighborhood filtering. The Sharp-Linear-Self-Attention used in the SLN-Block can capture dense global priors. $[\boldsymbol{P}^g; x(\boldsymbol{P}^g)]$ and $[\boldsymbol{Q}^g; x(\boldsymbol{Q}^g)]$ have complete global prior knowledge. The Match module extracts dense correspondences between $\boldsymbol{P}$ and $\boldsymbol{Q}$. Local-to-Global Registration solves the transformation $\mathbf{T}$ relay on the dense point correspondences.

backpropagates the complete global associations in the feature space, obtaining point cloud features $x(\boldsymbol{P}^g)$ and $x(\boldsymbol{Q}^g)$ with global priors. We use a sequence of three KPFResBlocks to downsample and aggregate superpoints. Two levels of UpSampling and Decoder are used for recovering dense points. The coarse-grained superpoints and fine-grained dense points are used for obtaining dense point correspondences through the Match module [37]. Subsequently, Local-to-Global Registration(LGR) [37] solves for $\mathbf{R}$ and $\mathbf{t}$ on the dense corresponding points.

As shown in Fig. 1, KPFResBlock contains KPFConvBlock and UnaryBlock. UnaryBlock is composed of a Linear layer and Norm&ReLU block. KPFConvBlock contains KPConvFormer and Norm&ReLU block. KPConvFormer(Fig. 2) combines KPConv[41] and attention, using a neighborhood filtering strategy to obtain a clean convolution neighborhood (Sec. 3.1). Based on the original Transformer architecture, we build a SLN-Block using Sharp-Linear-Self-Attention(Fig. 3), which achieves efficient information exchange across the global range, learning dense priors (Sec. 3.2).

3.1 Neighborhood Filtering Strategy(KPConvFormer)

To explore the generalizability of point convolution, we have noticed the following constraints proved in [5]:

$$\hat{G}_N(F) \leq C \max_{p' \in \mathcal{N}(p)} \sqrt{\mathbb{E}_{x,p}[(x(p) - x(p'))^2]}, \tag{2}$$

where $\hat{G}_N(F)$ is the empirical Gaussian complexity about the function class F: a single-layer CNN followed by a fully connected layer, and C is a constant. The smaller the Gaussian complexity, the better the generalization [5]. To minimize the constraint in Eq. 2, points with higher feature correlations should be selected within the same neighborhood. For images, neighboring pixels are usually considered to have higher color correlations [27]. Therefore conventional CNNs achieve better generalization by choosing smaller local neighborhoods (e.g., 3x3).

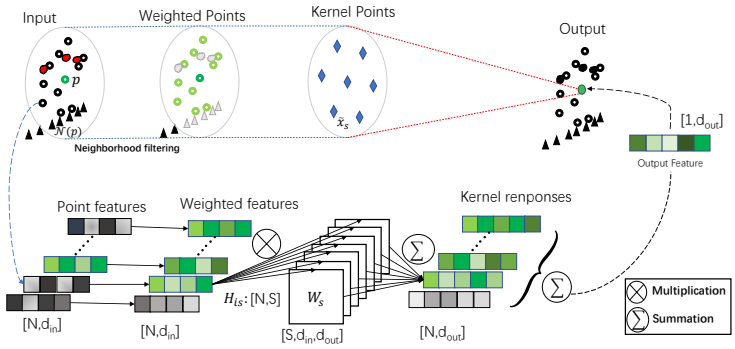

Fig. 2. The point convolution process of the KPConvFormer. The input spherical convolution region $\mathcal{N}(p)$ is the neighborhood of the dark green circular points, the hollow black circular points are the relevant points, the red points are noise points, and the triangles represent irrelevant points from other instances near the boundary. H_{is} is the association score between the neighborhood points and kernel points defined in KPConv[41], determined by their position difference.

For 3D point cloud scenes, the same convolution neighborhood may contain irrelevant points from other instances and noisy points. We expect to use the attention mechanism to explicitly examine the magnitude of $x(p)-x(p')$, minimize $\max_{p'\in\mathcal{N}(p)}\sqrt{\mathbb{E}_{x,p}\left[(x(p)-x(p'))^2\right]}$, and filter out unnecessary points within the $\mathcal{N}(p)$ neighborhood.

Inspired by the above generalization theory, we design a point convolution operator called KPConvFormer to learn robust superpoints and apply it to point cloud registration tasks. The KPConvFormer operation for a point p with a spherical convolution neighborhood $\mathcal{N}(p)$ is expressed as Eq. 3:

$$x'_p = \sum_{p_i\in\mathcal{N}(p)} g(p_i-p)\psi((x(p_i)-x(p),p_i-p))x(p_i), \tag{3}$$

$$g(y_i) = \sum_{s<S} h(y_i,\tilde{x}_s)W_s, \tag{4}$$

$$h(y_i,\tilde{x}_s) = \max\left(0,1-\frac{\|y_i-\tilde{x}_s\|}{\sigma}\right), \tag{5}$$

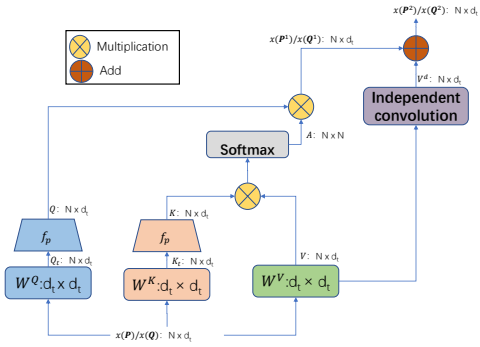

Fig. 3. The computation graph of Sharp-Linear-Self-Attention. W^Q, W^K, and W^V represent the projection matrices for the query Q, key K, and value V, respectively.f_p represents the focusing function.

where the scalar function $\psi(\cdot)$ is a learner based on the feature difference $x(p_i) - x(p)$ and the position difference $p_i - p$. $\psi(\cdot)$ is simulated by a multi-layer perceptron and an activation layer, learning how to select useful points in the neighborhood, which works similarly to the multi-head attention in Transformer. The kernel function $g(p_i - p)$ in KPConv [41] learns the convolution weights of the distance of neighborhood points relative to the kernel points within $\mathcal{N}(p)$, with $\{\tilde{x}_s | s < S\}$ set as the kernel points, $\{W_s | s < S\}$ as the weight matrix related to the kernel points. $|\tilde{x}_s| = s < |\mathcal{N}(p)| = S$. The influence distance of the kernel points σ is determined by the input density.

As shown in Fig. 2, the noisy and irrelevant points in the input region are weakened (shown in light gray) after neighborhood filtering, while the relevant points are assigned higher weights (shown in light green). The weighted convolution neighborhood is then convolved with multiple kernel points(defined in KPConv[41]) based on the distance difference, learns different regional features of the point cloud and obtains the dark green neighborhood output. This KPConvFormer considers both the position and feature differences of the neighborhood points, pre-evaluates the degree of attention to different points within the neighborhood, and avoids the impact of indistinguishable irrelevant points, noise points, and relevant points that are spatially close.

3.2 Sharp Global Linear Attention Prior(Sharp-Linear-Self-Attention)

Inspired by linear attention in the image domain [18], we utilize a simple and effective focusing function and independence convolution in 3D space to enhance the ability of linear attention to perceive critical features in point clouds, obtaining a sharp attention distribution and diversified features. The computation graph of Sharp-Linear-Self-Attention is shown in Fig. 3. The input point cloud features $x(\boldsymbol{P})/x(\boldsymbol{Q})$ are mapped into high-dimensional Q_t, K_t, and V after the

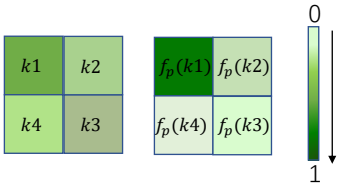

Fig. 4. Calculate the attention scores of query q and keys $k1$,$k2$,$k3$, and $k4$ using vanilla linear attention: [0.29, 0.17, 0.23, 0.12] (*left*). Calculate the linear attention scores of query q and keys $k1$,$k2$,$k3$, and $k4$ after using f_p: [0.67, 0.13, 0.10, 0.04] (*right*).

corresponding projection matrices W^Q, W^K, and W^V. The focusing function f_p performs further mapping on Q_t and K_t to produce Q and K. To achieve linear computational complexity, the attention scores A computed from K and V are multiplied with Q to obtain the temporary features $x(\boldsymbol{P}^1)/x(\boldsymbol{Q}^1)$. After applying independence convolution on V, the final outputs $x(\boldsymbol{P}^2)/x(\boldsymbol{Q}^2)$ are obtained by summing $x(\boldsymbol{P}^1)/x(\boldsymbol{Q}^1)$ and the diversified V^d. Each point in the point cloud learns effective global dependencies through this processing.

A simple and effective focusing function. Softmax attention provides a nonlinear re-weighting mechanism that makes it easier to focus on important features [7,35]. To obtain a sharp distribution similar to softmax attention in linear attention, inspired by [18], we use focusing function f_p to re-map the query Q and key K in the attention and then calculate the similarity score $\mathrm{Sim}(Q_i, K_j)$:

$$\mathrm{Sim}(Q_i, K_j) = \phi_p(Q_i)\phi_p(K_j)^T, \tag{6}$$

$$\phi_p(y) = f_p(\mathrm{ReLU}(y)), \quad f_p(y) = \frac{\|y\|}{\|y^p\|} y^p, \tag{7}$$

where y^p represents element-wise power p of y. The ReLU function is used for ensuring the non-negativity of input. From Eq. 7, we observed the norm of the feature is preserved after the mapping, i.e.,$\|y\| = \|f_p(y)\|$, indicating that only the feature direction is adjusted.

As shown in Fig. 4, f_p achieves a more obvious distinction between similar and dissimilar Query-Key pairs, simulating the sharp attention distribution of the softmax attention, which helps the network focus on important features of the point cloud.

Independence convolution. Han et al.[18,56] pointed out that the low rankness of the linear attention matrix limits its ability to capture diverse features, and the rank of this matrix is constrained by the number of tokens N and the channel dimension d of each attention head:

$$\mathrm{rank}(\phi(Q)\phi(K)^T) \leq \min\{\mathrm{rank}(\phi(Q)), \mathrm{rank}(\phi(K))\} \leq \min\{N, d\}, \tag{8}$$

where d is commonly smaller than N in vision Transformer, e.g., d=64, N=196 in DeiT[42] and d=32, N=49 in Swin Transformer[28].Therefore, the upper bound of the attention matrix rank is restricted at a lower ratio, indicating that many attention map rows are homogenized, which leads to the resemblance among the output features. We additionally use 3D depth-wise convolution (3d-DWC) to increase the feature diversity of the linear attention. The complete formulation of Sharp-Linear-Self-Attention is as follows:

$$O = \text{Sim}(Q, K)V = \phi_p(Q)\phi_p(K)^T V + \text{3d-DWC}(V). \tag{9}$$

From the perspective of the attention mechanism, 3d-DWC means that each query only focuses on the neighboring features in the point cloud space, and the non-adjacent features maintain independence during the convolution operation, avoiding the ambiguity caused by attending to all features V. The locality of 3d-DWC ensures that even if the linear attention values corresponding to two queries are the same, different outputs can still be obtained from different local features, thereby maintaining feature diversity.

4 Loss Function

Following the Geometric [37], we use the Overlap-Aware Circle loss $\mathcal{L}_{oc}$ and the point matching loss $\mathcal{L}_p$ for the loss function of point cloud registration:

$$\mathcal{L} = \mathcal{L}_{oc} + \mathcal{L}_p. \tag{10}$$

$\mathcal{L}_{oc}$ supervises the quality of matching between superpoints. It ensures that the superpoint correspondences learned by the model during training are not only consistent in feature space but also geometrically matched. $\mathcal{L}_p$ supervises the accuracy of dense point-level matching. It can optimize the model's matching prediction for each point, ensuring that the local details between point clouds can be accurately aligned.

5 Experiments

We evaluate the GPNF framework on the indoor 3DMatch and the 3DLoMatch benchmarks[19,58] (Sec. 5.1), as well as the outdoor KITTI odometry benchmark[15] (Sec. 5.2). We also compare our method with the recent Deep Robust Estimators (Sec. 5.3). The ablation studies on the framework details are designed(Sec. 5.4). All three datasets are added with noise to verify the framework's robustness.

5.1 Indoor Benchmarks: 3DMatch & 3DLoMatch

Dataset. 3DMatch contains 62 indoor scenes, with 46 for training, 8 for validation, and 8 for testing. We use the preprocessed training data from [19] for our experiments. The point cloud pairs in 3DMatch have an overlap rate above 30%, while the point cloud pairs in 3DLoMatch have an overlap rate between 10% and 30%.

Table 1. Correspondence results on 3DMatch and 3DLoMatch. Boldfaced numbers highlight the best and the second best are underlined

	3DMatch			3DLoMatch			3DMatch			3DLoMatch		
Samples	2500	1000	500	2500	1000	500	2500	1000	500	2500	1000	500
	Fetaure Matching Recall(%)↑						Inlier Ratio(%)↑					
PerfectMatch[16]	94.3	92.9	90.1	61.7	53.6	45.2	32.5	26.4	21.5	10.1	8.0	6.4
FCGF[11]	97.3	97.0	96.7	75.4	74.2	71.7	54.2	48.7	42.5	20.0	17.2	14.8
D3Feat[4]	95.4	94.5	94.1	66.7	67.0	66.7	38.8	40.4	41.5	13.1	14.0	14.6
SpinNet[1]	97.2	96.8	95.5	74.9	72.5	70.0	44.7	39.4	33.9	19.0	16.3	13.8
Predator[19]	96.6	96.5	96.3	77.4	76.3	75.7	58.4	57.1	54.1	28.1	28.3	27.5
YOHO[44]	97.6	97.5	97.7	78.1	76.3	73.8	60.7	55.7	46.4	23.3	22.6	18.2
CoFiNet[55]	**98.3**	98.1	**98.2**	83.5	83.3	83.1	51.2	51.9	52.2	25.9	26.7	26.8
Geo[36]	98.1	98.1	**98.2**	**87.7**	87.8	88.0	75.9	76.8	82.8	45.8	46.7	53.3
GPNF(ours)	97.9	**98.2**	**98.2**	**87.7**	**88.5**	**88.4**	**78.6**	**83.1**	**85.0**	**48.5**	**54.6**	**57.2**

Metrics. Following [4,19], we evaluate the performance with three metrics: (1) Inlier Ratio (IR), the fraction of putative correspondences whose residuals are below a certain threshold (i.e., 0.1m) under the ground-truth transformation, (2) Feature Matching Recall (FMR), the fraction of point cloud pairs whose inlier ratio is above a certain threshold (i.e., 5%), and (3) Registration Recall (RR), the fraction of point cloud pairs whose transformation error is smaller than a certain threshold (i.e., RMSE $<$ 0.2m).

Correspondence Results. Tab. 1 shows the correspondence results of our method with the recent state of the arts: PerfectMatch[16], D3Feat[4], FCGF[11], CoFiNet[55], Predator[19], SpinNet[1], YOHO[44], and Geotransformer[36]. Following [4,19], we report the results with different numbers of correspondences. For FMR, our method achieves a relatively significant improvement in the low-overlap 3DLoMatch dataset. IR consistently achieves more significant improvements on both 3DMatch and 3DLoMatch, surpassing the baselines by 0.6% to 6.3% on 3DMatch and improving by 1.1% to 7.9% on 3DLoMatch. It indicates that our method extracts more reliable correspondences.

Registration Results. Tab. 2 shows the registration results without RANSAC. For the weighted SVD estimator [6], we use the top 250 correspondences with the highest confidence to estimate the transformation, as required by Geotransformer [36]. Most of the methods struggled to obtain reasonable results, while our method achieved the best performance with an 88.0% registration recall rate on 3DMatch and the second-best result on 3DLoMatch. It's worth noting that we did not use RANSAC here, and the high registration recall rate is primarily attributed to a higher inlier ratio. However, a high inlier ratio does not necessarily lead to a high registration recall rate, as the correspondences may cluster together, as pointed out in [19]. Nevertheless, our method performs well in extracting reliable and well-distributed point correspondences without using

Table 2. Registration results w/o RANSAC on 3DMatch and 3DLoMatch.Boldfaced numbers highlight the best and the second best are underlined

Model	Estimator	Samples	Registration Recall(%)↑	
			3DMatch	3DLoMatch
FCGF[11]	Weighted SVD	250	42.1	3.9
D3Feat[4]	Weighted SVD	250	37.4	2.8
SpinNet[1]	Weighted SVD	250	34.0	2.5
Predator[19]	Weighted SVD	250	50.0	6.4
CoFiNet[55]	Weighted SVD	250	64.6	21.6
Geo[36]	Weighted SVD	250	<u>86.7</u>	**60.5**
GPNF(ours)	Weighted SVD	250	**88.0**	<u>60.1</u>
CoFiNet[55]	LGR	all	87.6	64.8
Geo[36]	LGR	all	<u>91.8</u>	**74.5**
GPNF(ours)	LGR	all	**91.9**	<u>73.4</u>

RANSAC. For the local-to-global estimator(LGR), our method improves the registration recall rate on 3DMatch to 91.9%, far exceeding all the methods using the weighted SVD estimator, and achieves the second-best result on 3DLoMatch.

Qualitative Results. Fig. 5 shows some qualitative results of our method on 3DLoMatch. Our method can achieve accurate registration results even on the low-overlap 3DLoMatch dataset. Relying on the learning of global priors and the neighborhood filtering strategy, our method can distinguish similar structures at different locations (such as the backrest and seat cushion of a sofa, similar walls and floors), achieving registration closer to the ground-truth. Additional qualitative results can be found in the supplementary materials.

5.2 Outdoor Benchmark: KITTI Odometry

Dataset. KITTI odometry [15] consists of 11 sequences of outdoor driving scenes scanned by LiDAR. We follow works[4,11,19] and use sequences 0-5 for training, 6-7 for validation and 8-10 for testing. The ground-truth poses are refined with ICP, and we only use point cloud pairs that are at least 10m away for evaluation.

Metrics. We follow Geotransformer[36] to evaluate our GPNF with three metrics: (1) Relative Rotation Error (RRE), the geodesic distance between estimated and ground-truth rotation matrices, (2) Relative Translation Error (RTE), the Euclidean distance between estimated and ground-truth translation vectors, and (3) Registration Recall (RR), the fraction of point cloud pairs whose RRE and RTE are both below certain thresholds (i.e., RRE$<5^\circ$ and RTE$<$2m).

Registration Results. As shown in Tab. 3, the top of the table lists the state-of-the-art RANSAC-based methods, including D3Feat [4], FCGF [11], CoFiNet

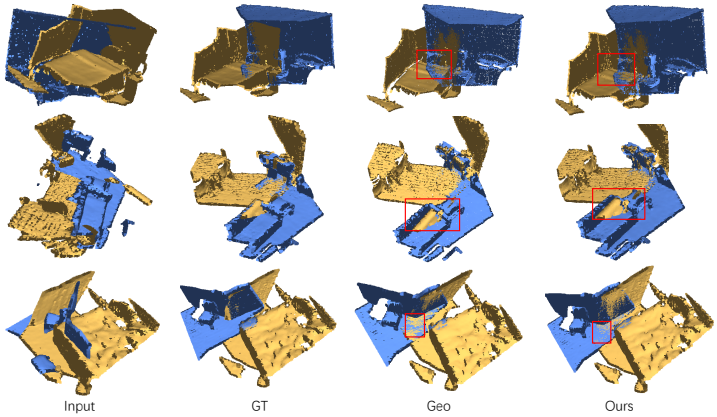

Fig. 5. Comparison of the registration results on 3DLoMatch. As seen from the red boxes in the figures, our method can achieve registration results closer to the ground truth.

[55], Predator [19], SpinNet [1], 3DFeat-Net [52] and Geotransformer [36]. The results show that our method achieves the best results on all three metrics. The bottom of Tab. 3 lists some RANSAC-free methods, such as FMR [20], DGR [10], HRegNet [29] and Geotransformer [36]. The results show that our method also achieves the best registration metrics. It should be noted that the improvement of our method on the RTE metric, which reflects the registration accuracy, is more significant.

5.3 Comparison with Deep Robust Estimators

We compare the performance of GPNF with the recent deep robust estimators on the 3DMatch, 3DLoMatch, and KITTI odometry benchmarks, including 3DRegNet [32], DGR [10], PointDSC [3], DHVR [23], and PCAM [8]. To make a fair comparison, we follow the common practice and compare the three metrics of RTE (Relative Translation Error), RRE (Relative Rotation Error), and RR (Registration Recall) on the above datasets. Here RR is defined as in Section 5.2 but with different thresholds. The RTE threshold is 30 cm on 3DMatch and 60 cm on KITTI, while the RRE threshold is 15° on 3DMatch and 5° on KITTI.

The results are shown in Tab. 4 and Tab. 5. Our method significantly outperforms all the baselines on the three benchmarks, even though the other methods use other types of correspondence extractors.

5.4 Ablation Study

We conduct ablation studies on the 3DLoMatch benchmark. To evaluate superpoint (patch) matching, we introduce another metric Patch Inlier Ratio (PIR) which is the fraction of patch matches with actual overlap. The FMR is reported with all dense point correspondences, with LGR being used for registration.

Table 3. Registration results on KITTI odometry. Boldfaced numbers highlight the best, and the second best are underlined.

Model	RTE(cm)↓	RRE(°)↓	RR(%)↑
3DFeat-Net[52]	25.9	0.25	96.0
FCGF[11]	9.5	0.30	96.6
D3Feat[4]	7.2	0.30	**99.8**
SpinNet[1]	9.9	0.47	99.1
Predator[19]	6.8	0.27	**99.8**
CoFiNet[55]	8.2	0.41	**99.8**
Geo(RANSAC)[36]	7.4	0.27	**99.8**
GPNF(ours)	**6.0**	**0.23**	**99.8**
FMR[20]	∼66	1.49	90.6
DGR[10]	∼32	0.37	98.7
HRegNet[29]	∼12	0.29	99.7
Geo(LGR)[36]	6.8	0.24	**99.8**
GPNF(ours)	**6.0**	**0.23**	**99.8**

Table 4. Comparison with deep robust estimators on 3DMatch and KITTI. Boldfaced numbers highlight the best, and the second best are underlined.

Model	RTE(cm)↓	RRE(°)↓	RR(%)↑	RTE(cm)↓	RRE(°) ↓	RR(%)↑
	3DMatch			KITTI		
FCGF+3DRegNet[32]	8.13	2.74	77.8		\	
FCGF+DGR[10]	7.36	2.33	86.5	21.7	0.34	96.9
FCGF+PointDSC[3]	6.55	2.06	93.3	20.9	0.33	98.2
FCGF+DHVR[23]	6.61	2.08	91.4	19.8	0.29	99.1
PCAM[8]	∼7	2.16	92.4	∼8	0.33	97.2
GPNF(Ours)	**5.60**	**1.91**	**95.7**	**5.8**	**0.22**	**99.5**

Neighborhood filtering strategy and global prior. We separately validated the effectiveness of KPConvFormer and SLN-Block. As shown in Tab. 6, both KPConvFormer and SLN-Block improve the precision of matching and registration to different degrees. Between them, the performance gain of SLN-Block is the most prominent, indicating the importance of global prior information in the point cloud registration task. The neighborhood filtering effect of KPConvFormer is mainly concentrated around a relatively small number of instance boundaries and noisy points in the point cloud scene, and its own improvement effect is not as significant as SLN-Block. However, it is still necessary for complex point cloud registration tasks with high-precision requirements.

The number of attention heads. We test the impact of the number of attention heads in KPConvFormer on the experimental results. As shown in Tab. 7, 8-heads attention significantly improved the accuracy of matching and registration compared to 4-heads attention. 8-heads attention can learn richer spatial representations from more fine-grained perspectives compared to 4-heads attention, and the model's flexibility in integrating these different perspectives is

Table 5. Comparison with deep robust estimators on 3DLoMatch.Boldfaced numbers highlight the best, and the second best are underlined.

Model	RTE(cm)↓	RRE(°)↓	RR(%)↑
3DLoMatch			
FCGF+PointDSC[3]	<u>10.50</u>	<u>3.82</u>	<u>56.2</u>
FCGF+DHVR[23]	11.76	3.88	55.6
GPNF(Ours)	**8.50**	**2.97**	**77.2**

Table 6. Ablation studies on A: KPConvFormer and B: SLN-Block.(Double the original noise level)

A	B	PIR(%)↑	FMR(%)↑	RR(%)↑
×	×	51.7	83.5	69.6
✓	×	53.2	85.9	71.5
×	✓	52.8	86.6	71.7
✓	✓	**54.6**	**88.2**	**72.8**

Table 7. Ablation study on the number of attention heads in the neighborhood filtering strategy.

Head numbers	PIR(%)↑	FMR(%)↑	RR(%)↑
4	54.4	87.0	72.2
8	**55.7**	**88.9**	73.4
16	55.6	87.5	**73.6**

also higher. However, the metrics slightly declined when the number of attention heads increased to 16. We ultimately chose to use an 8-heads attention for neighborhood filtering.

6 Conclusion

To improve the generalization of point features for point cloud registration, the KPConvFormer introduces the idea of attention mechanism in point convolution to filter out noisy and irrelevant points in the convolution neighborhood. The SLN-Block learns complete global dependencies in fine-grained point clouds, providing prior knowledge for subsequent point convolution, which can distinguish similar structures in the target point cloud from a broader context. Therefore, GPNF can also be seen as an initial attempt to integrate point convolution and attention in point cloud registration tasks. We hope that this idea of integrating point convolution and attention can inspire and provide insights for other types of downstream point cloud tasks.

Acknowledgments. This work was supported by the National Natural Science Foundation of China (NSFC) under Grant 62171145, and also by Guangxi Key R&D Program under Grant AB23075106.

References

1. 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. pp. 11753–11762 (2021)
2. Aoki, Y., Goforth, H., Srivatsan, R.A., Lucey, S.: Pointnetlk: Robust & efficient point cloud registration using pointnet. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 7163–7172 (2019)
3. 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. pp. 15859–15869 (2021)
4. Bai, X., Luo, Z., Zhou, L., Fu, H., Quan, L., Tai, C.L.: D3feat: Joint learning of dense detection and description of 3d local features. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 6359–6367 (2020)
5. Bartlett, P.L., Mendelson, S.: Rademacher and gaussian complexities: Risk bounds and structural results. Journal of Machine Learning Research **3**(Nov), 463–482 (2002)
6. Besl, P.J., McKay, N.D.: Method for registration of 3-d shapes. In: Sensor fusion IV: control paradigms and data structures. vol. 1611, pp. 586–606. Spie (1992)
7. Cai, H., Gan, C., Han, S.: Efficientvit: Enhanced linear attention for high-resolution low-computation visual recognition. arXiv preprint arXiv:2205.14756 **3** (2022)
8. 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. pp. 13229–13238 (2021)
9. Chen, X., Ma, H., Wan, J., Li, B., Xia, T.: Multi-view 3d object detection network for autonomous driving. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. pp. 1907–1915 (2017)
10. Choy, C., Dong, W., Koltun, V.: Deep global registration. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 2514–2523 (2020)
11. Choy, C., Park, J., Koltun, V.: Fully convolutional geometric features. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 8958–8966 (2019)
12. Deng, H., Birdal, T., Ilic, S.: Ppf-foldnet: Unsupervised learning of rotation invariant 3d local descriptors. In: Proceedings of the European conference on computer vision (ECCV). pp. 602–618 (2018)
13. Deng, H., Birdal, T., Ilic, S.: Ppfnet: Global context aware local features for robust 3d point matching. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 195–205 (2018)
14. Fu, K., Liu, S., Luo, X., Wang, M.: Robust point cloud registration framework based on deep graph matching. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8893–8902 (2021)
15. Geiger, A., Lenz, P., Urtasun, R.: Are we ready for autonomous driving? the kitti vision benchmark suite. In: 2012 IEEE conference on computer vision and pattern recognition. pp. 3354–3361. IEEE (2012)
16. Gojcic, Z., Zhou, C., Wegner, J.D., Wieser, A.: The perfect match: 3d point cloud matching with smoothed densities. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5545–5554 (2019)

17. Guo, M.H., Cai, J.X., Liu, Z.N., Mu, T.J., Martin, R.R., Hu, S.M.: Pct: Point cloud transformer. Computational Visual Media **7**, 187–199 (2021)
18. Han, D., Pan, X., Han, Y., Song, S., Huang, G.: Flatten transformer: Vision transformer using focused linear attention. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 5961–5971 (2023)
19. Huang, S., Gojcic, Z., Usvyatsov, M., Wieser, A., Schindler, K.: Predator: Registration of 3d point clouds with low overlap. In: Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition. pp. 4267–4276 (2021)
20. Huang, X., Mei, G., Zhang, J.: Feature-metric 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. pp. 11366–11374 (2020)
21. Katharopoulos, A., Vyas, A., Pappas, N., Fleuret, F.: Transformers are rnns: Fast autoregressive transformers with linear attention. In: International conference on machine learning. pp. 5156–5165. PMLR (2020)
22. Lang, A.H., Vora, S., Caesar, H., Zhou, L., Yang, J., Beijbom, O.: Pointpillars: Fast encoders for object detection from point clouds. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 12697–12705 (2019)
23. Lee, J., Kim, S., Cho, M., Park, J.: Deep hough voting for robust global registration. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 15994–16003 (2021)
24. Li, B., Zhang, T., Xia, T.: Vehicle detection from 3d lidar using fully convolutional network. arXiv preprint arXiv:1608.07916 (2016)
25. Li, J., Zhang, C., Xu, Z., Zhou, H., Zhang, C.: Iterative distance-aware similarity matrix convolution with mutual-supervised point elimination for efficient point cloud registration. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIV 16. pp. 378–394. Springer (2020)
26. Li, X., Li, F., Fern, X., Raich, R.: Filter shaping for convolutional neural networks. In: International Conference on Learning Representations (2022)
27. Li, X., Li, F., Fern, X., Raich, R.: Filter shaping for convolutional neural networks. In: International Conference on Learning Representations (2022)
28. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 10012–10022 (2021)
29. Lu, F., Chen, G., Liu, Y., Zhang, L., Qu, S., Liu, S., Gu, R.: Hregnet: A hierarchical network for large-scale outdoor lidar point cloud registration. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 16014–16023 (2021)
30. Ma, X., Qin, C., You, H., Ran, H., Fu, Y.: Rethinking network design and local geometry in point cloud: A simple residual mlp framework. arXiv preprint arXiv:2202.07123 (2022)
31. Maturana, D., Scherer, S.: Voxnet: A 3d convolutional neural network for real-time object recognition. In: 2015 IEEE/RSJ international conference on intelligent robots and systems (IROS). pp. 922–928. IEEE (2015)
32. 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. pp. 7193–7203 (2020)

33. 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. pp. 652–660 (2017)
34. Qi, C.R., Yi, L., Su, H., Guibas, L.J.: Pointnet++: Deep hierarchical feature learning on point sets in a metric space. Advances in neural information processing systems **30** (2017)
35. Qin, Z., Sun, W., Deng, H., Li, D., Wei, Y., Lv, B., Yan, J., Kong, L., Zhong, Y.: cosformer: Rethinking softmax in attention. arXiv preprint arXiv:2202.08791 (2022)
36. Qin, Z., Yu, H., Wang, C., Guo, Y., Peng, Y., Ilic, S., Hu, D., Xu, K.: Geotransformer: Fast and robust point cloud registration with geometric transformer. IEEE Transactions on Pattern Analysis and Machine Intelligence **45**(8), 9806–9821 (2023)
37. Qin, Z., Yu, H., Wang, C., Guo, Y., Peng, Y., Xu, K.: Geometric transformer for fast and robust point cloud registration. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11143–11152 (2022)
38. Robert, D., Raguet, H., Landrieu, L.: Efficient 3d semantic segmentation with superpoint transformer. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 17195–17204 (2023)
39. Song, S., Yu, F., Zeng, A., Chang, A.X., Savva, M., Funkhouser, T.: Semantic scene completion from a single depth image. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1746–1754 (2017)
40. Su, H., Maji, S., Kalogerakis, E., Learned-Miller, E.: Multi-view convolutional neural networks for 3d shape recognition. In: Proceedings of the IEEE international conference on computer vision. pp. 945–953 (2015)
41. Thomas, H., Qi, C.R., Deschaud, J.E., Marcotegui, B., Goulette, F., Guibas, L.J.: Kpconv: Flexible and deformable convolution for point clouds. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 6411–6420 (2019)
42. Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., Jégou, H.: Training data-efficient image transformers & distillation through attention. In: International conference on machine learning. pp. 10347–10357. PMLR (2021)
43. Uy, M.A., Pham, Q.H., Hua, B.S., Nguyen, T., Yeung, S.K.: Revisiting point cloud classification: A new benchmark dataset and classification model on real-world data. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 1588–1597 (2019)
44. Wang, H., Liu, Y., Dong, Z., Wang, W.: You only hypothesize once: Point cloud registration with rotation-equivariant descriptors. In: Proceedings of the 30th ACM International Conference on Multimedia. pp. 1630–1641 (2022)
45. 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. pp. 3523–3532 (2019)
46. Wang, Y., Solomon, J.M.: Prnet: Self-supervised learning for partial-to-partial registration. Advances in neural information processing systems **32** (2019)
47. Wu, X., Jiang, L., Wang, P.S., Liu, Z., Liu, X., Qiao, Y., Ouyang, W., He, T., Zhao, H.: Point transformer v3: Simpler faster stronger. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4840–4851 (2024)
48. Wu, X., Lao, Y., Jiang, L., Liu, X., Zhao, H.: Point transformer v2: Grouped vector attention and partition-based pooling. Advances in Neural Information Processing Systems **35**, 33330–33342 (2022)

49. Xu, H., Liu, S., Wang, G., Liu, G., Zeng, B.: Omnet: Learning overlapping mask for partial-to-partial point cloud registration. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 3132–3141 (2021)
50. Xu, H., Ye, N., Liu, G., Zeng, B., Liu, S.: Finet: Dual branches feature interaction for partial-to-partial point cloud registration. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 36, pp. 2848–2856 (2022)
51. Yang, Y.Q., Guo, Y.X., Xiong, J.Y., Liu, Y., Pan, H., Wang, P.S., Tong, X., Guo, B.: Swin3d: A pretrained transformer backbone for 3d indoor scene understanding. arXiv preprint arXiv:2304.06906 (2023)
52. Yew, Z.J., Lee, G.H.: 3dfeat-net: Weakly supervised local 3d features for point cloud registration. In: Proceedings of the European conference on computer vision (ECCV). pp. 607–623 (2018)
53. Yew, Z.J., Lee, G.H.: Rpm-net: Robust point matching using learned features. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11824–11833 (2020)
54. Yew, Z.J., Lee, G.H.: Regtr: End-to-end point cloud correspondences with transformers. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 6677–6686 (2022)
55. Yu, H., Li, F., Saleh, M., Busam, B., Ilic, S.: Cofinet: Reliable coarse-to-fine correspondences for robust point cloud registration. Advances in Neural Information Processing Systems **34**, 23872–23884 (2021)
56. Yu, T., Khalitov, R., Cheng, L., Yang, Z.: Paramixer: Parameterizing mixing links in sparse factors works better than dot-product self-attention. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 691–700 (2022)
57. Yuan, W., Eckart, B., Kim, K., Jampani, V., Fox, D., Kautz, J.: Deepgmr: Learning latent gaussian mixture models for registration. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part V 16. pp. 733–750. Springer (2020)
58. 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. pp. 1802–1811 (2017)
59. Zhang, Z., Sun, J., Dai, Y., Zhou, D., Song, X., He, M.: End-to-end learning the partial permutation matrix for robust 3d point cloud registration. In: Proceedings of the AAAI conference on Artificial Intelligence. vol. 36, pp. 3399–3407 (2022)
60. Zhao, H., Jiang, L., Fu, C.W., Jia, J.: Pointweb: Enhancing local neighborhood features for point cloud processing. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5565–5573 (2019)
61. Zhao, H., Jiang, L., Jia, J., Torr, P.H., Koltun, V.: Point transformer. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 16259–16268 (2021)