A Coarse-to-Fine Indoor Point Cloud Registration Method Guided by Prior Correspondences

Mengchong Sun[1], Juntao Yang[1*], Yutao Zhang[1], Sa Li,[1] Dandan Liu[1], Xue Zhang[2]

[1]College of Geodesy and Geomatics, Shandong University of Science and Technology, Qingdao 266590, China - smc20010713@163.com, jtyang@sdust.edu.cn, zhangyutao060116@outlook.com, 744853961@qq.com, 1870739192@qq.com
[2]College of Computer Science and Engineering, Shandong University of Science and Technology, Qingdao 266590, China - xuezhang@sdust.edu.cn

Keywords: Indoor Scenes, Point Cloud Registration, Coarse-to-Fine, Multiple Geometric Information, Sparse Mixture-of-Experts Model.

Abstract:

Superpoint matching is a critical step in coarse-to-fine point cloud registration, and its performance directly affects the accuracy of subsequent point matching and pose estimation. However, most existing methods establish correspondences mainly relying on feature similarity, without explicit modeling of spatial structure, which easily leads to unstable matching in complex scenarios such as noise, occlusion, and low overlap. To address these issues, this paper proposes a coarse-to-fine point cloud registration method guided by prior correspondences. First, prior superpoint correspondences are constructed using rigid transformations estimated by existing SOTA methods, and are serially encoded via a prior encoding module to provide explicit constraints for feature learning. Furthermore, multiple geometric information including pairwise distances, angles, and normals is introduced and uniformly encoded to enhance spatial structure representation. On this basis, a prior-guided sparse mixture-of-experts attention mechanism is designed to differentially model features in overlapping and non-overlapping regions, thereby improving feature discriminability and structural consistency. Using the learned features, the model gradually establishes correspondences through superpoint matching and point matching, and estimates the final rigid transformation with RANSAC. Experiments on the 3DMatch dataset show that when sampling 1000 point correspondences, the proposed method achieves an inlier ratio of 80.7% and a registration recall of 92.9%, which are 5.5% and 1.1% higher than the baseline method respectively, verifying the effectiveness of the proposed method in terms of accuracy and robustness.

1. Introduction

With the continuous development of LiDAR technology, the acquisition of high-precision point cloud data has become increasingly efficient, and point clouds have gradually become a key data form in 3D applications, making point cloud data processing crucial. As an important part of point cloud processing, point cloud registration aims to compute a transformation matrix to align local point clouds from different coordinate systems into a unified coordinate system, thereby reconstructing a complete 3D scene (B. Yang et al., 2024). Depending on the application environment and scene characteristics, point cloud registration can generally be divided into two typical scenarios: indoor and outdoor. Compared with outdoor environments, indoor scenes often feature more complex spatial structures and functional layouts, severe local occlusions, and a large number of geometrically smooth and repetitive structures. These factors significantly increase the difficulty of point cloud feature extraction and matching. In addition, typical indoor applications, such as indoor 3D reconstruction (T. Yang et al., 2022) and virtual/augmented reality (Nguyen et al., 2024), usually impose higher requirements on the accuracy and stability of registration results, further elevating the challenges of the point cloud registration task.

In recent years, leveraging the remarkable advantages of deep learning in high-dimensional feature representation and nonlinear modeling, researchers have proposed numerous learning-based point cloud registration methods, which have significantly advanced the development of this field. In general, existing approaches can be broadly classified into two categories: one is correspondence-based registration methods centered on deep feature extraction and matching (Bai et al., 2020; Choy et al., 2019; Yu et al., 2024), and the other is pose regression-based methods that directly regress rigid transformation parameters through end-to-end networks(Wang and Solomon, 2019; W. Yuan et al., 2020; Y. Yuan et al., 2023).Among them, pose regression-based methods typically rely on global features for holistic representation and can achieve satisfactory performance in scenarios with relatively complete structures and sufficient overlap. However, in complex scenes with low overlap, severe occlusion, or large-scale variations, such methods often struggle to extract stable and discriminative global features, leading to inaccurate or even failed transformation estimation. In contrast, correspondence-based registration methods explicitly construct matching relationships between points or superpoints, transforming the registration problem into a process of local correspondence screening and optimization. They exhibit stronger robustness and generalization ability in complex environments, and have thus gradually become the mainstream direction of current research.

In such methods, how to improve feature representation ability and the reliability of correspondences has become crucial. To address these issues, SACF-Net (Wu et al., 2023) introduces a skip-attention-based matching filtering strategy that selectively fuses multi-level features in a multi-resolution encoder‑decoder architecture. This allows the network to better focus on overlapping regions during decoding, effectively filtering high-quality correspondences and improving performance in low-overlap scenarios. GeoTransformer (Qin et al., 2023) further leverages

* Corresponding author: jtyang@sdust.edu.cn

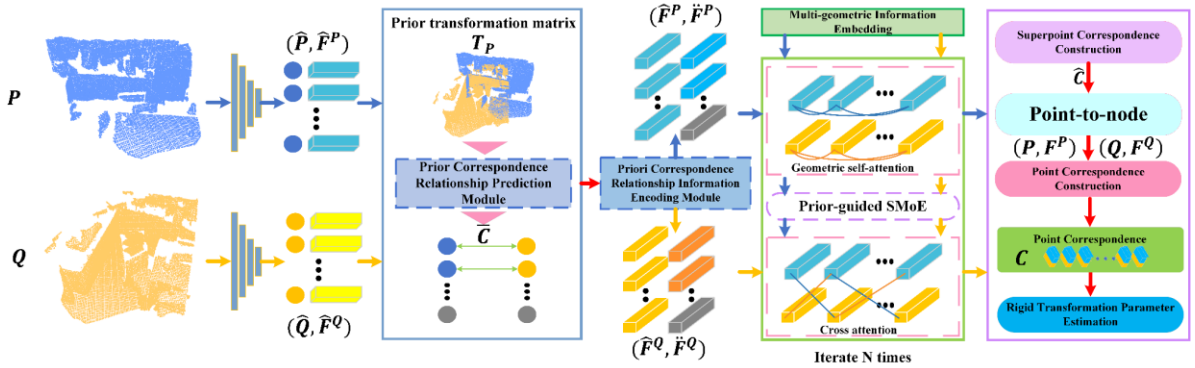


Figure 1. Overall workflow of the proposed method

geometric information by designing a geometric self-attention mechanism that incorporates relative positional and angular cues into attention computation, enhancing structural awareness. It also introduces an overlap-aware coarse matching loss to improve the confidence and stability of correct superpoint matches. On this basis, RoITr (Yu et al., 2023) extends this idea by integrating geometric priors into a Transformer framework. It employs rotation-invariant Point Pair Features (PPF) to model both local and global structural relationships, enabling robust feature interaction under large rotations, low overlap, and noise, and thus producing more stable and discriminative matching results.

Although existing methods have achieved significant progress in feature representation and matching robustness, their ability to explicitly model and distinguish overlapping regions remains limited. To address this issue, we introduce a mechanism that enables region-aware feature modeling. The Sparse Mixture-of-Experts (SMoE) (Fedus et al., 2022) provides a promising solution: if the routing network can leverage both feature distribution and spatial relationships to distinguish tokens in overlapping and non-overlapping regions, different experts can be assigned to process different feature subsets. This allows specialized feature extraction for different regions, enhancing discriminative representations in overlapping areas while reducing interference from non-overlapping regions, thereby improving correspondence reliability and registration accuracy. However, without prior guidance, the routing network struggles to accurately identify overlapping regions. Therefore, we incorporate prior correspondences into the attention and SMoE framework, and propose a coarse-to-fine point cloud registration method for indoor scenes. The main contributions are summarized as follows:

(1) Propose a prior correspondence encoding module. We construct prior superpoint correspondences based on rigid transformations estimated by existing methods, and embed them into raw features through sequential encoding, enabling the network to explicitly exploit potential matching information during feature learning and provide a stable initialization for coarse matching.
(2) Introduce a multi-geometric information embedding mechanism. We uniformly encode geometric attributes such as distances, angles, and normal directions between point clouds to construct learnable geometric structure representations, which are integrated into feature embedding to provide explicit spatial structure priors for subsequent feature learning.
(3) Design a prior-guided sparse mixture-of-experts attention mechanism. Guided by prior correspondences, we build a multi-expert feature learning framework that performs specialized modeling on different feature subspaces and adaptively fuses expert outputs, achieving differential processing of features in overlapping and non-overlapping regions.

2. Methodology

Given a pair of point clouds $P = \{p_i \in \mathbb{R}^3 | i = 1, \dots, N\}$ and $Q = \{q_i \in \mathbb{R}^3 | j = 1, \dots, M\}$, the objective is to estimate a rigid transformation $T = \{R, t\}$, where $R \in \mathcal{SO}(3)$ denotes the rotation and $t \in \mathbb{R}^3$ denotes the translation, to align them. To estimate T, it is necessary to establish point correspondences C between P and Q. In our method, the correspondences are progressively constructed through a coarse-to-fine strategy. At the superpoint matching stage, prior superpoint correspondences are first obtained using existing methods, and a prior encoding module is employed to encode them into sequences, providing prior information for subsequent feature learning. Furthermore, multiple geometric cues, including inter-point distances, angles, and normal directions, are incorporated and uniformly encoded to form a learnable geometric structural representation. Based on this, we design a prior-guided sparse mixture-of-experts attention mechanism, where geometric information is integrated into the attention computation. Under the guidance of prior correspondences, different experts are encouraged to learn discriminative features for overlapping and non-overlapping regions, respectively. Finally, the point correspondences are established through both superpoint-level and point-level matching modules, and the transformation matrix is estimated using the RANSAC algorithm. The overall pipeline of the proposed method is illustrated in Figure 1.

2.1 Prior Correspondence Encoding

We employ the encoding layers of KPConv to perform feature extraction and downsampling on the input raw point cloud data, which reduces computational complexity while preserving key information of the point clouds. Through this process, we obtain superpoints $\hat{P}$ and $\hat{Q}$, as well as the corresponding superpoint features $\hat{F}^P \in \mathbb{R}^{|\hat{P}| \times b}$ and $\hat{F}^Q \in \mathbb{R}^{|\hat{Q}| \times b}$.

(1) Prior Correspondence Prediction:
Superpoint correspondences provide not only overlap cues between two point clouds but also explicit matching relationships between source and target superpoints. These cues can serve as valuable prior information to guide the routing network in selecting appropriate experts, thereby improving feature discrimination. However, reliable correspondence estimation depends on strong feature representations, while learning discriminative features also requires accurate correspondences. This forms a mutually dependent relationship between feature learning and correspondence estimation. To address this issue, we adopt the method of Sun et al. (2025) to estimate a prior rigid transformation $T_p = \{R_p, t_p\}$. With the prior transformation, the source point cloud is transformed and aligned with the target point cloud,

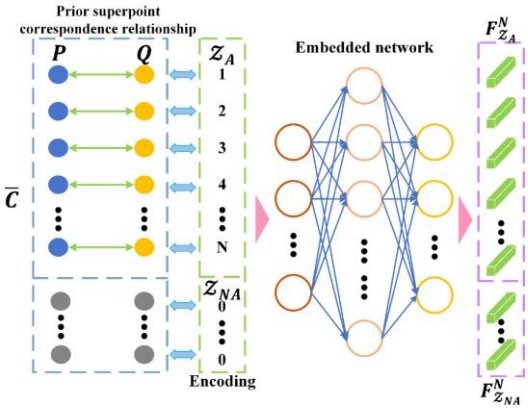


Figure 2. Prioritization Correspondence Coding Process

enabling the estimation of an overlap probability matrix $\hat{O} \in \mathbb{R}^{|\hat{P}|\times|\hat{Q}|}$ between the transformed superpoint sets $\hat{P}$ and $\hat{Q}$, where each element represents the overlap probability between a pair of superpoints. Subsequently, superpoint pairs with overlap probabilities greater than a predefined threshold τ_o are selected from $\hat{O}$ as prior superpoint correspondences $\bar{C}$, and their corresponding overlap probabilities are denoted as $\bar{O}$.

(2) Encoding of Prior Correspondences:
To encode prior correspondence information in a learnable form, we design a prior correspondence encoding module to explicitly represent predicted superpoint matches (as illustrated in Figure 2). First, we construct an ordered index set $Z_A = \{1,2,\cdots, c'_p\}$to assign unique identifiers to each predicted correspondence pair $\bar{C}$, where $c'_p$is the number of predicted matches.To handle unmatched superpoints, we introduce a special index $Z_{NA} = 0$, which is assigned to superpoints in both $\hat{P}$and $\hat{Q}$that do not participate in any correspondence. Based on this scheme, all superpoints are encoded into a unified integer index representation, providing a consistent way to describe both matched and unmatched cases:

$$\mathcal{N} = \{i \in \mathbb{N} | i = 0,1,\cdots, c'_p\} \tag{1}$$

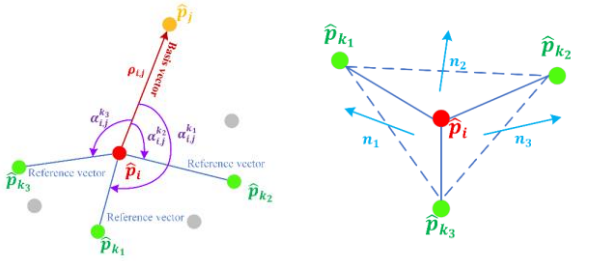
To map the above discrete indices into a continuous feature space, we adopt a sinusoidal positional encoding function to generate their embedding representations. This approach preserves the ordinal relationships of the indices while transforming discrete identifiers into continuous vector representations with structured properties:

$$\begin{cases} \Gamma^N_{i,2k} = \sin\left(\dfrac{i}{10000^{2k/d_t}}\right) \\ \Gamma^N_{i,2k+1} = \cos\left(\dfrac{i}{10000^{2k/d_t}}\right) \end{cases} \tag{2}$$

where d denotes the embedding dimension. Subsequently, an embedding layer is constructed using a multi-layer perceptron (MLP) to compute the prior embeddings:

$$F^N = MLP(\Gamma^N), \quad F^N \in \mathbb{R}^{(c'_p+1)\times d_t} \tag{3}$$

After obtaining the prior embeddings, we utilize the prior superpoint correspondences $\bar{C}$to select the source superpoints involved in matching from the superpoint set $\hat{P}$, forming an anchor set $A \subset \hat{P}$. In real-world scenarios, a source superpoint in Amay correspond to multiple target superpoints; that is, under the constraint of the threshold τ_o, multiple candidate matches may exist, forming a candidate correspondence set A'. In this case, a single source superpoint is associated with multiple prior embedding representations, which may introduce redundancy and uncertainty. To ensure the uniqueness and stability of the prior embeddings, we perform a weighted fusion of multiple embeddings based on their overlap probabilities. Specifically, for each superpoint in A, all its matching entries are retrieved from



(a) (b)

Figure 3. Multiple geometric information. (a) Is distance and angle information, (b) Is normal vector information

the prior correspondence set $\bar{C}$, and their indices are denoted as X_i. Finally, the prior embedding representation of all superpoints in the source point cloud $\hat{P}$, denoted as $\ddot{F}^P \in \mathbb{R}^{|\hat{P}|\times d}$, can be uniformly formulated as:

$$\ddot{F}^P_i = \begin{cases} softmax\left(\bar{O}(X_i)\right) \cdot F^N_{Z_A}(X_i), & |X_i| > 1 \\ F^N_{Z_A}(X_i), & |X_i| = 1 \\ F^N_{Z_{NA}}, & otherwise \end{cases} \tag{4}$$

where $F^N_{Z_A}$and $F^N_{Z_{NA}}$denote the prior embedding features corresponding to $Z_A$and Z_{NA}, respectively. For the superpoint features of the target point cloud $\ddot{F}^Q$, we apply the same encoding procedure to obtain the corresponding prior embeddings. Through this unified encoding strategy, a consistent prior representation is established between the source and target point clouds, thereby providing structured prior information to support subsequent feature modeling and routing network decisions.

2.2 Multi-Geometric Information Encoding

Superpoint matching is a key step in coarse-to-fine point cloud registration, but it is inherently sparse and uncertain due to downsampling. As a result, nearby superpoints may still differ significantly in geometry, making reliable correspondence difficult when relying only on spatial proximity or feature similarity, which is also sensitive to noise and occlusion. To address this, we introduce a multi-level geometric encoding strategy that captures structural cues such as distances, angles, and normal directions between superpoints (as illustrated in Figure 3). These geometric relationships are encoded and integrated into feature learning to enhance the reliability of matching.

Since these geometric relationships are determined by local spatial structures, they are highly consistent within overlapping regions and can provide reliable constraints for feature learning. For two superpoints $\hat{p}_i, \hat{p}_j \in \hat{P}$, we construct their geometric representation using three components: distance, angle, and normal embeddings. Specifically, the Euclidean distance is first computed and encoded as a distance embedding. The angle embedding captures their relative directional relationships within the local neighborhood. In addition, normal information is used to model orientation consistency and is encoded as a normal-based embedding. By combining these components, the spatial relationship between superpoints is represented from multiple geometric perspectives, leading to a more robust and discriminative feature representation. The embedding process can be formulated as follows:

(1) Distance Embedding.
For superpoints $\hat{p}_i$and $\hat{p}_j$, their Euclidean distance is defined as $\rho_{(i,j)} = \| \hat{p}_i - \hat{p}_j \|_2$. The distance embedding $\rho^D_{(i,j)} \in \mathbb{R}^{d_t}$ is then obtained using the same sinusoidal encoding function as described in Section 2.1:

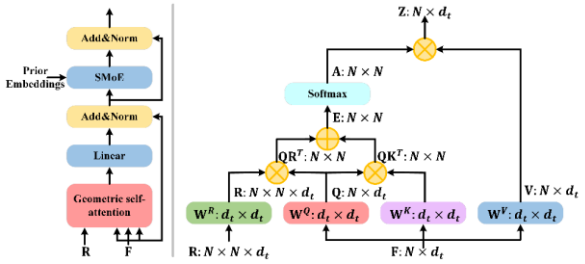


Figure 4. Combining the geometric self-attention mechanism of SMoE

$$\begin{cases} r^D_{i,j,2k} = \sin\left(\dfrac{\rho_{i,j}/\sigma_d}{10000^{2k/d_t}}\right) \\ r^D_{i,j,2k+1} = \cos\left(\dfrac{\rho_{i,j}/\sigma_d}{10000^{2k/d_t}}\right) \end{cases} \tag{5}$$

where σ_d is a scaling parameter used to control the sensitivity of the distance embedding to variations in distance, thereby affecting the scale of the encoded representation.

(2) Angle Embedding.

For angle embedding, we first select the k nearest neighbors K_i of $\hat{p}_i$ within its local neighborhood. For any point $\hat{p}_x \in K_i$, we take $\hat{p}_x$ as the reference point and construct direction vectors pointing to $\hat{p}_i$ and $\hat{p}_j$, respectively. The angle between these two direction vectors is then computed as $\alpha^x_{(i,j)}$, which characterizes the geometric relationship between local structure and global configuration. Subsequently, the angle value $\alpha^x_{(i,j)}$ is mapped into a high-dimensional space using a sinusoidal positional encoding scheme, yielding the corresponding angle embedding representation:

$$\begin{cases} r^A_{i,j,2l} = \sin\left(\dfrac{\alpha^x_{i,j}/\sigma_\alpha}{10000^{2l/d_t}}\right) \\ r^A_{i,j,2l+1} = \cos\left(\dfrac{\alpha^x_{i,j}/\sigma_\alpha}{10000^{2l/d_t}}\right) \end{cases} \tag{6}$$

Similarly, σ_α is a scaling parameter used to control the sensitivity of the angle embedding to variations in the angle, thereby regulating the scale of the encoded representation.

(3) Normal Embedding.

To fully exploit local geometric information in point clouds, we construct normal features based on neighborhood points and further encode them into embeddings. Specifically, for each point $\hat{p}_i$, we first select its K nearest neighbors, denoted as N_i. For any two points $\hat{p}_x, \hat{p}_y \in N_i$, we take $\hat{p}_i$ as the reference point and construct vectors $\mathbf{v}_x = \hat{p}_x - \hat{p}_i$ and $\mathbf{v}_y = \hat{p}_y - \hat{p}_i$. The local normal vector is then computed via the cross product:

$$\mathbf{n}_{xy} = \frac{\mathbf{v}_x \times \mathbf{v}_y}{\| \mathbf{v}_x \times \mathbf{v}_y \|} \tag{7}$$

To address the ambiguity of normal directions, we introduce a global reference direction to ensure consistency. Using the point cloud origin as a reference, we compute a unit vector toward it and flip normals whose dot product indicates opposite orientation. The consistent normals are then concatenated within each neighborhood to form an initial geometric descriptor. Finally, an MLP is applied to map this descriptor into a higher-dimensional feature space for learning:

$$\boldsymbol{f}^n_i = \mathrm{MLP}(n_i) \tag{8}$$

where the MLP consists of two fully connected layers with a ReLU activation function in between, enabling nonlinear mapping from geometric descriptors to a semantic feature space. For a point pair $(\hat{p}_{i'}\hat{p}_j)$, their corresponding normal embeddings are concatenated to form a pairwise geometric representation:

$$\boldsymbol{r}^N_{i,j} = \mathrm{Concat}([\boldsymbol{f}^n_i, \boldsymbol{f}^n_j]) \tag{9}$$

Finally, the distance embedding, angle embedding, and normal embedding are fused and aggregated to obtain a unified multi-level geometric embedding representation:

$$\boldsymbol{r_{i,j}} = r^D_{i,j} \cdot W^D + \max\{r^A_{i,j} \cdot W^A\} + \boldsymbol{r}^N_{i,j} \cdot W^N \tag{10}$$

where W^D, W^A, and $W^N \in \mathbb{R}^{d_t \times d_t}$ are projection matrices corresponding to the three types of embedding information, respectively.

2.3 Prior-guided Sparse Mixture-of-Experts Attention Mechanism

In point cloud registration, attention mechanisms capture long-range dependencies through global feature interaction. However, standard feed-forward networks use a uniform transformation, which cannot adapt well to the diverse regions and structures in point clouds.In practice, point clouds contain overlapping and non-overlapping regions with different geometric patterns and feature distributions. A single transformation strategy can lead to feature homogenization, reducing the discriminative ability of important regions and harming matching accuracy.

SMoE introduces a multi-expert architecture with routing, enabling dynamic token assignment and conditional feature transformation. It improves model expressiveness without significantly increasing computational cost. Replacing the Transformer feed-forward network with SMoE creates adaptive computation paths for different tokens, leading to more discriminative features. However, in point cloud registration, standard SMoE is limited because routing relies only on features and ignores geometric structure and overlap relationships, which may cause mismatched expert assignment. To address this, we propose a Prior-guided SMoE Attention Mechanism (PSMEAM), which incorporates prior correspondences into both routing and feature interaction (as illustrated in Figure 4). This helps the model focus on potential correspondences, improves representation in overlapping regions, and suppresses noise from non-overlapping areas, leading to more reliable matching.

We further apply geometric self-attention to the multi-geometric encoded superpoint features to capture global dependencies. By introducing geometric constraints into the attention process, the model jointly models feature similarity and spatial structure. Taking source superpoints as an example, the input feature $X \in \mathbb{R}^{|\hat{P}| \times d_t}$ is transformed into the output $Z \in \mathbb{R}^{|\hat{P}| \times d_t}$ through weighted aggregation of projected features.

$$z_i = \sum_j^{|\hat{P}|} a_{i,j}(x_j W^V) \tag{11}$$

where $W^V \in \mathbb{R}^{d_t \times d_t}$ is the projection matrix for values. The attention weight $a_{i,j}$ is obtained by applying a softmax operation over the attention score $e_{i,j}$. The computation of the attention score $e_{i,j}$ is formulated as follows:

$$e_{i,j} = \frac{(x_i W^Q)(x_j W^K + r_{i,j} W^R)^T}{\sqrt{d_t}} \tag{12}$$

where $W_Q, W_K, W_R \in \mathbb{R}^{d_t \times d_t}$ are the projection matrices for queries, keys, and geometric embeddings, respectively. $r_{i,j} \in \mathbb{R}^{d_t}$ denotes the geometric structural embedding between superpoints i and j.

After the geometric self-attention layer enhances feature interactions, we further introduce a Sparse Mixture-of-Experts (SMoE) module to adaptively refine feature representations. The encoded prior correspondence information is incorporated as guidance in the expert routing process, enabling differentiated learning

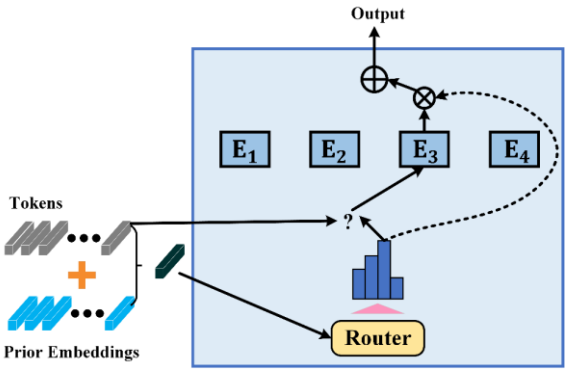


Figure 5. Prior Correspondence Relationship Information Guided SMoE

between overlapping and non-overlapping regions. Specifically, conventional Feed-Forward Networks (FFNs) in Transformers apply a uniform transformation to all tokens, which limits their ability to adapt to the heterogeneous nature of point cloud regions, such as overlapping versus non-overlapping areas. To address this limitation, we replace the FFN with an SMoE module, allowing each superpoint feature to be dynamically routed to different expert networks for conditional processing.

Given the superpoint features $Z \in \mathbb{R}^{|\hat{P}| \times d_t}$ produced by geometric self-attention and the corresponding prior-encoded features $\ddot{F}^P \in \mathbb{R}^{|\hat{P}| \times d_t}$, we fuse them as inputs to the SMoE routing network to guide expert selection (as illustrated in Figure 4). Specifically, for the i-th superpoint, the fused representation is formulated as $z_i + \ddot{f}_i$, where z_i denotes the self-attention output feature and $\ddot{f}_i$ represents the corresponding prior encoding. Based on this fused representation, routing scores over experts are computed, and a top-k strategy is applied to select the most relevant k experts for computation. In this work, we set $k = 2$, and the routing process can be formulated as follows:

$$g_i = top - k(softmax((W^g(z_i + \ddot{f}_i)))) \quad (13)$$

where $W^g \in \mathbb{R}^{d_t \times E}$ denotes the learnable parameters of the routing network, E is the number of experts, and $g_i \in \mathbb{R}^E$ represents the sparse gating vector for the i-th superpoint. Finally, the output feature of the i-th superpoint is obtained as a weighted combination of the selected experts, formulated as:

$$\hat{z}_i = \sum_{e \in \mathcal{G}_i} g_i^{(e)} \cdot \mathcal{E}_e(z_i) \quad (14)$$

where G_i denotes the set of experts selected by the top-k routing strategy, $g_i^{(e)}$ is the routing weight assigned to the e-th expert, and $E_e(\cdot)$ represents the e-th expert network. Through this mechanism, superpoint features in both overlapping and non-overlapping regions can be adaptively assigned to different experts according to their prior information and intrinsic features, thereby enhancing the diversity and discriminative power of the learned representations.

2.4 Loss function

To improve training stability and performance, we use a joint loss including three terms: overlap-aware loss L_{oc}, point matching loss L_f, and load balance loss L_b. Among them, L_f supervises point-wise correspondence accuracy, L_{oc} enhances overlap discrimination for better superpoint matching, and L_b balances expert utilization in SMoE. Overall, this multi-task objective improves both matching accuracy and feature representation.

In this section, we focus on the overlap-aware loss and the load balance loss, while the point matching loss follows the formulation of Qin et al. (2023).

Method	3DMatch					3DLoMatch				
	5000	2500	1000	500	250	5000	2500	1000	500	250
Feature Matching Recall (%)										
3DSN	95.0	94.3	92.9	90.1	82.9	63.6	61.7	53.6	45.2	34.2
FCGF	97.4	97.3	97.0	96.7	96.6	76.6	75.4	74.2	71.7	67.3
D3Feat	95.6	95.4	94.5	94.1	93.1	67.3	66.7	67.0	66.7	66.5
SpinNet	97.4	97.0	96.4	96.7	94.8	75.5	75.1	74.2	69.0	62.7
Predator	96.6	96.6	96.5	96.3	96.5	78.6	77.4	76.3	75.7	75.3
CoFiNet	**98.1**	**98.3**	**98.1**	**98.2**	**98.3**	83.1	83.5	83.3	83.1	82.6
RIGA	97.9	97.8	97.7	97.7	97.6	85.1	85.0	85.1	84.3	85.1
GeoTrans	97.9	97.9	97.9	97.9	97.6	**88.3**	**88.6**	**88.8**	**88.6**	**88.3**
Ours	97.9	97.9	97.9	97.9	97.6	87.2	87.6	87.9	87.6	87.4
Inlier Ratio (%)										
3DSN	36.0	32.5	26.4	21.5	16.4	11.4	10.1	8.0	6.4	4.8
FCGF	56.8	54.1	48.7	42.5	34.1	21.4	20.0	17.2	14.8	11.6
D3Feat	39.0	38.8	40.4	41.5	41.8	13.2	13.1	14.0	14.6	15.0
SpinNet	48.5	46.2	40.8	35.1	29.0	25.7	23.7	20.6	18.2	13.1
Predator	58.0	58.4	57.1	54.1	49.3	26.7	28.1	28.3	27.5	25.8
CoFiNet	49.8	51.2	51.9	52.2	52.2	24.4	25.9	26.7	26.8	26.9
RIGA	68.4	69.7	70.6	70.9	71.0	32.1	33.4	34.3	34.5	34.6
GeoTrans	71.9	75.2	76.0	82.2	85.1	43.5	45.3	46.2	52.9	57.7
Ours	**74.1**	**80.7**	**85.4**	**87.2**	**88.5**	**47.6**	**53.3**	**59.3**	**61.8**	**63.6**
Registration Recall (%)										
3DSN	78.4	76.2	71.4	67.6	50.8	33.0	29.0	23.3	17.0	11.0
FCGF	85.1	84.7	83.3	81.6	71.4	40.1	41.7	38.2	35.4	26.8
D3Feat	81.6	84.5	83.4	82.4	77.9	37.2	42.7	46.9	43.8	39.1
SpinNet	88.8	88.0	84.5	79.0	69.2	58.2	56.7	49.8	41.0	26.7
Predator	89.0	89.9	90.6	88.5	86.6	59.8	61.2	62.4	60.8	58.1
CoFiNet	89.3	88.9	88.4	87.4	87.0	67.5	66.2	64.2	63.1	61.0
RIGA	89.3	88.4	89.1	89.0	87.7	65.1	64.7	64.5	64.1	61.8
GeoTrans	92.0	91.8	91.8	91.4	91.2	**75.0**	**74.8**	**74.2**	**74.1**	**73.5**
Ours	**92.5**	**92.9**	**92.4**	**92.2**	**91.7**	74.4	74.1	73.6	73.9	72.9

Table 1 Evaluation Results of 3DMatch and 3DLoMatch Datasets Under Different Sample Sizes

(1) Overlap-aware loss

Existing methods usually treat superpoint matching as a multi-label classification problem optimized with cross-entropy loss. However, in multi-positive cases, overlaps vary significantly, and cross-entropy may weaken high-confidence matches, reducing stability. To address this, we adopt the overlap-aware loss from GeoTransformer. We first build an anchor set $\hat{A}$ from $\hat{P}$, where each anchor has at least one positive match in $\hat{Q}$. A pair is positive if its overlap ratio is ≥10%, negative if it is 0, and others are ignored. For each anchor G_i^P, its positive and negative sets are ε_p^i and ε_n^i. Here, G_i^P is obtained via the point-to-node strategy(Sun et al., 2025). Based on this, the overlap-aware loss on $\hat{P}$ is defined as:

$$\mathcal{L}_{oc}^P = \frac{1}{|\mathcal{A}|} \sum_{G_i^P \in \mathcal{A}} \log \left[1 + \sum_{G_j^Q \in \varepsilon_p^i} e^{\lambda_i^j \beta_p^{i,j} (d_i^j - \Delta_p)} \cdot \sum_{G_k^Q \in \varepsilon_n^i} e^{\beta_n^{i,k} (\Delta_n - d_i^k)} \right] \quad (15)$$

where $d_i^j = \| \hat{h}_i^P - \hat{h}_j^Q \|_2$ denotes the Euclidean distance in the feature space, and $\lambda_i^j = (o_i^j)^{1/2}$, where o_i^j represents the overlap ratio between G_i^P and G_j^Q. The weights for positive and negative samples are computed individually as $\beta_p^{(i,j)} = \gamma(d_i^j - \Delta_p)$ and $\beta_n^{(i,k)} = \gamma(\Delta_n - d_i^k)$, respectively. The margin hyperparameters Δ_p and Δ_n are set to 0.1 and 1.4, respectively. The overlap-aware loss further reweights the loss terms over ε_p^i according to the overlap ratio, assigning larger weights to superpoint pairs with higher overlap. The loss on the target point cloud $\hat{Q}$, denoted as L_{oc}^Q, is computed in the same manner. The final overlap-aware loss is defined as: $\mathcal{L}_{oc} = (\mathcal{L}_{oc}^P + \mathcal{L}_{oc}^Q)/2$。

(2) Load balance loss

The basic routing mechanism often suffers from imbalanced

Method	Estimator	3DMatch			3DLoMatch		
		RR	RRE	RTE	RR	RRE	RTE
Predator	weighted SVD	50.0	3.89	0.122	6.4	>10	>1
CoFiNet	weighted SVD	64.6	2.92	0.089	21.6	6.34	0.154
GeoTrans	weighted SVD	86.5	2.13	0.070	59.9	3.88	0.105
Ours	weighted SVD	**91.9**	**1.82**	**0.062**	**73.6**	**2.88**	**0.088**
CoFiNet	LGR	87.6	2.23	0.071	64.8	3.49	0.124
GeoTrans	LGR	91.5	1.91	0.068	74.0	2.95	0.090
Ours	LGR	**92.7**	**1.73**	**0.061**	**75.0**	**2.66**	**0.084**

Table 2 The model's evaluation results on the 3DMatch and 3DLoMatch datasets without using the RANSAC algorithm

expert utilization. During training, the model tends to rely on a few experts, which receive more updates and become further reinforced, leading to routing collapse. This reduces feature diversity and degrades performance. To address this, we introduce a load balance loss to regularize expert usage. For each mini-batch, we compute two statistics: (i) the load f_i, representing the proportion of tokens assigned to expert i, and (ii) the importance P_i, representing the average normalized gating probability of expert i. These are formed into vectors of size E, and the load balance loss is defined using their dot product, scaled by the number of experts:

$$\mathcal{L}_b = E \cdot \sum_{i=1}^{E} f_i \cdot P_i \tag{16}$$

When both the load and importance of each expert approach a uniform distribution (i.e., $f_i \approx 1/E$ and $P_i \approx 1/E$), the loss approaches 1. In contrast, when only a few experts dominate the routing process, the resulting product increases, leading to a higher loss value. By minimizing this loss, the gating network is encouraged to distribute tokens more evenly across all experts, thereby preventing excessive concentration on a subset of experts. This loss is used as an auxiliary term and is added to the main task loss, with a weighting coefficient $\lambda = 0.01$during optimization to jointly guide gradient updates.

Finally, the overall loss function can be formulated as: $\mathcal{L} = \mathcal{L}_{oc} + \mathcal{L}_f + \lambda\mathcal{L}_b$.

3. Experiments

We evaluate the proposed method on the indoor datasets 3DMatch (with overlap > 30%) and 3DLoMatch (with overlap between 10% and 30%), and compare it with several state-of-the-art methods, including 3DSN (Gojcic et al., 2019), FCGF (Choy et al., 2019), D3Feat (Bai et al., 2020), SpinNet (Ao et al., 2021), Predator (Huang et al., 2021), CoFiNet (Yu et al., 2021), RIGA (Yu et al., 2024), and GeoTransformer (Qin et al., 2023) (denoted as GeoTrans in the table).

3.1 Evaluation on 3DMatch and 3DMLoMatch

Evaluation Metrics: Following prior works, we adopt Registration Recall (RR), Inlier Ratio (IR), and Feature Matching Recall (FMR) to evaluate the performance of our method. Specifically, (1) Registration Recall (RR) measures the success rate of point cloud registration, defined as the percentage of point cloud pairs whose estimated rigid transformation (via RANSAC) has an error below a certain threshold (e.g., RMSE < 0.2 m) compared to the ground truth. (2) Inlier Ratio (IR) evaluates the reliability of correspondences, defined as the proportion of matched point pairs whose geometric residuals are below a predefined threshold (e.g., $\tau_1 = 0.1$m) under the ground-truth transformation. (3) Feature Matching Recall (FMR) measures the quality of feature matching, defined as the percentage of point cloud pairs whose Inlier Ratio exceeds a given threshold (e.g., $\tau_2 = 5\%$).

Registration Results: To evaluate robustness under different correspondence sparsity levels, we follow Bai et al. (2020)and test our method with varying numbers of sampled correspondences (Table 1).As the number of correspondences decreases from 5000 to 250, our method consistently outperforms existing methods. It achieves higher IR improvements of 2.2%–9.4% on 3DMatch and 4.1%–13.1% on 3DLoMatch, and also shows consistently better FMR. For RR, our method surpasses most baselines on both datasets, with clear gains on 3DMatch and competitive results on 3DLoMatch.Overall, the results show that our method maintains strong stability and robust matching performance even under sparse correspondence settings.

To further validate the proposed method, we provide a visual comparison with CoFiNet, which also adopts a coarse-to-fine framework. As shown in Figure 6, the first two rows are from 3DMatch and the last two are from 3DLoMatch. In cases where CoFiNet fails (e.g., the first and fourth rows), our method still achieves correct alignment, showing better robustness. When both methods succeed (e.g., the second row), our method produces more precise alignment with more consistent errors. Overall, our approach is more stable in complex and low-overlap scenes, improving both registration accuracy and success rate.

3.2 Ablation Study

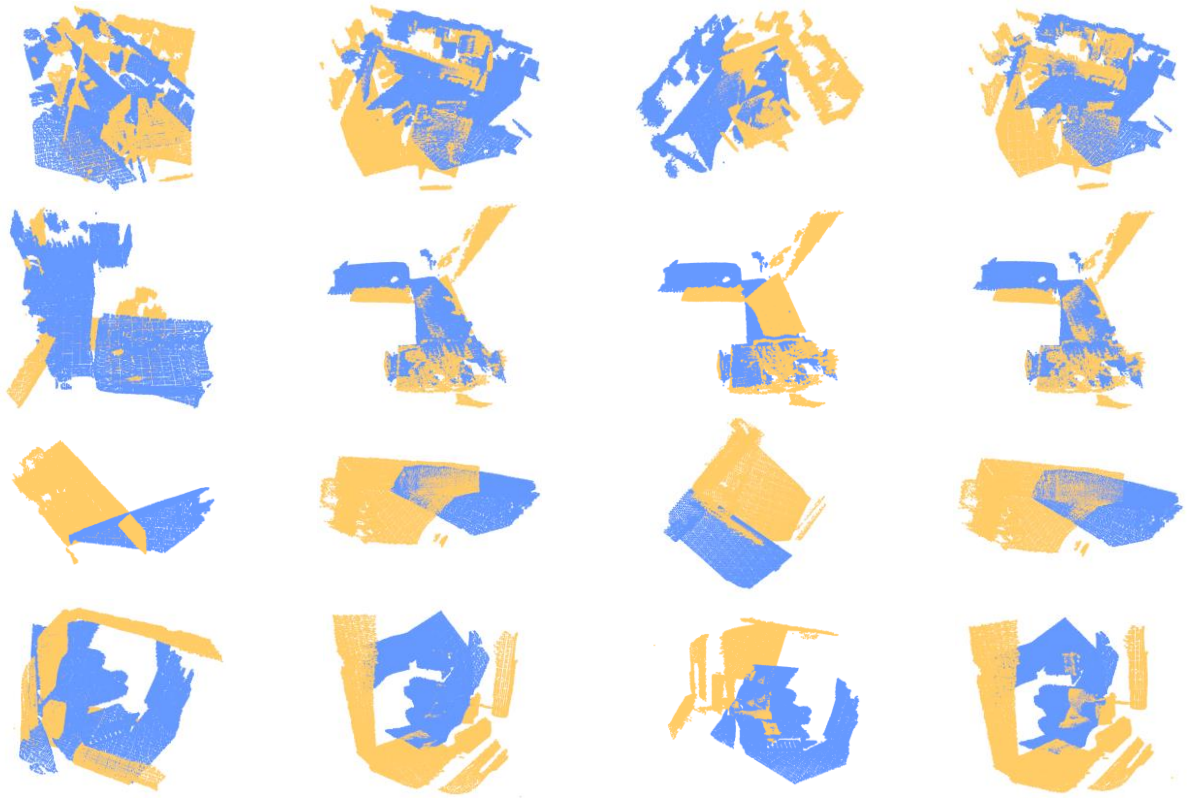
(1) Registration without RANSAC:
In this section, we evaluate registration performance without RANSAC (Table 2) using RR (%), RRE (°), and RTE (m). We first use weighted SVD for rigid transformation estimation. Under this setting, most methods degrade significantly. Predator suffers a large drop in RR and nearly fails in low-overlap scenes, while CoFiNet is also sensitive to outliers. GeoTrans remains relatively stable across both datasets.In contrast, our method maintains strong performance without RANSAC. On 3DMatch, RR exceeds 90% with low RRE and RTE. On 3DLoMatch, the improvement is more evident, with about 10% higher RR than GeoTrans. These results show that our superpoint correspondences have a higher inlier ratio and better spatial distribution, enabling robust registration even without explicit outlier filtering.

To further evaluate robustness, we replace weighted SVD with Local-to-Global Registration (LGR). Under this setting, all methods show performance improvements. CoFiNet and GeoTrans achieve notable gains in RR on both datasets, and the method from Chapter 3 also improves accordingly. Building upon this, our method still achieves the best performance: on 3DMatch, RR reaches 92.7%, with RRE and RTE reduced to 1.73° and 0.061 m, respectively; on 3DLoMatch, RR reaches 75.0%, with RRE and RTE of 2.66° and 0.084 m. Compared with GeoTrans, our method further improves both RR and error metrics, demonstrating that it can achieve more stable and accurate pose estimation when combined with a robust estimator.

(2) Ablation Study on Different Modules:

Model	3DMatch				3DLoMatch			
	PIR	FMR	IR	RR	PIR	FMR	IR	RR
vanilla self-attention	79.6	97.9	60.1	89.0	45.2	85.6	32.6	68.4
geometric self-attention	86.1	97.7	70.3	91.5	54.9	**88.1**	43.3	74.0
geometric self-attention w/ normal	87.0	**98.3**	71.2	91.7	55.3	87.7	44.1	72.0
PSMEAM	**90.6**	97.6	**72.7**	**92.7**	**64.4**	87.7	**47.4**	**75.0**

Table 3 Ablation studies of different modules



(a) Input (b) Ground Truth (c) CoFiNet (d) Ours

Figure 6. Visualization result graph (the first two rows show the registration results on 3DMatch, and the last two rows show the registration results on 3DLoMatch)

Ablation studies on 3DMatch and 3DLoMatch (Table 3) use RR, RRE, and RTE as metrics. Compared with vanilla self-attention, geometric self-attention significantly improves all results, with PIR and IR increasing by about 7%–10%, showing better structure modeling and correspondence quality. Adding normal information further improves performance, reaching 87.0% PIR, 71.2% IR, and 91.7% RR on 3DMatch, and 55.3% PIR and 44.1% IR on 3DLoMatch. Finally, introducing the SMoE module (PSMEAM) further boosts performance and robustness. On 3DMatch, metrics improve by about 2%–4%, while on 3DLoMatch the gains are more significant, with PIR increasing by nearly 10% and IR by about 3%. Although FMR slightly fluctuates, overall matching quality and stability are improved. These results confirm that SMoE is particularly effective in low-overlap scenarios.

4. Summary and outlook

Building on a coarse-to-fine point cloud registration framework, we propose an indoor registration method that integrates prior correspondence encoding, multi-level geometric information encoding, and a sparse mixture-of-experts (SMoE) attention mechanism to address the limitations of superpoint correspondence quality and feature representation. By introducing prior geometric constraints and structured feature modeling, together with an SMoE routing mechanism, our method improves superpoint matching and significantly enhances correspondence accuracy and registration robustness. Experiments on 3DMatch and 3DLoMatch demonstrate that our approach outperforms existing methods in registration accuracy, inlier ratio, and matching stability, with particularly notable gains under low-overlap conditions, confirming the effectiveness of each component. In future work, we will explore more efficient dynamic routing strategies to improve the efficiency and scalability of SMoE in large-scale point cloud scenarios, and investigate stronger cross-scale geometric modeling to further enhance robustness under extreme low-overlap and occlusion conditions.

Acknowledgements

This work is jointly supported by the National Natural Science Foundation of China (No. 42201486, 42371453, 62302278), the Natural Science Foundation of Shandong Province under Grant ZR2023QF014, the Qingdao Natural Science Foundation under Grant 23-2-1-112-zyyd-jch, the Higher Education Institutions Youth Innovation and Science & Technology Support Program of Shandong Province under Grant 2024KJH066, the Young Talent of Lifting engineering for Science and Technology in Shandong, China [SDAST2024QTA055].

Reference

Ao, S., Hu, Q., Yang, B., Markham, A., Guo, Y., 2021. 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).

Bai, X., Luo, Z., Zhou, L., Fu, H., Quan, L., Tai, C.-L., 2020. 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).

Choy, C., Park, J., Koltun, V., 2019. Fully convolutional geometric features. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 8958-8966).

Fedus, W., Zoph, B., Shazeer, N., 2022. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research, 23*(120), 1-39.

Gojcic, Z., Zhou, C., Wegner, J. D., Wieser, A., 2019. 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).

Huang, S., Gojcic, Z., Usvyatsov, M., Wieser, A., Schindler, K., 2021. 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).

Nguyen, A.-D., Choi, S., Kim, W., Kim, J., Oh, H., Kang, J., Lee, S., 2024. Single-Image 3-D Reconstruction: Rethinking Point Cloud Deformation. *IEEE Transactions on Neural Networks and Learning Systems, 35*(5), 6613-6627.

Qin, Z., Yu, H., Wang, C., Guo, Y., Peng, Y., Ilic, S., Hu, D., Xu, K., 2023. Geotransformer: Fast and robust point cloud registration with geometric transformer. *IEEE Transactions on pattern analysis and machine intelligence, 45*(8), 9806-9821.

Sun, M., Tan, J., Zhang, Y., Yang, J., Zhang, X., Liu, Y., Chen, J., 2025. Coarse-to-fine Point Cloud Registration Based on Superpoint Overlap Prediction. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences, 10*, 147-154.

Wang, Y., Solomon, J. M., 2019. Prnet: Self-supervised learning for partial-to-partial registration. *Advances in Neural Information Processing Systems, 32*.

Wu, Y., Hu, X., Zhang, Y., Gong, M., Ma, W., Miao, Q., 2023. SACF-Net: Skip-attention based correspondence filtering network for point cloud registration. *IEEE Transactions on Circuits and Systems for Video Technology, 33*(8), 3585-3595.

Yang, B., Dong, Z., Liang, F., Mi, X., 2024. *Ubiquitous Point Cloud: Theory, Model, and Applications*: CRC Press.

Yang, T., Ye, J., Zhou, S., Xu, A., Yin, J., 2022. 3D reconstruction method for tree seedlings based on point cloud self-registration. *Computers and Electronics in Agriculture, 200*, 107210.

Yu, H., Hou, J., Qin, Z., Saleh, M., Shugurov, I., Wang, K., Busam, B., Ilic, S., 2024. Riga: Rotation-invariant and globally-aware descriptors for point cloud registration. *IEEE Transactions on pattern analysis and machine intelligence, 46*(5), 3796-3812.

Yu, H., Li, F., Saleh, M., Busam, B., Ilic, S., 2021. Cofinet: Reliable coarse-to-fine correspondences for robust pointcloud registration. *Advances in Neural Information Processing Systems, 34*, 23872-23884.

Yu, H., Qin, Z., Hou, J., Saleh, M., Li, D., Busam, B., Ilic, S., 2023. Rotation-invariant transformer for point cloud matching. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 5384-5393).

Yuan, W., Eckart, B., Kim, K., Jampani, V., Fox, D., Kautz, J., 2020. Deepgmr: Learning latent gaussian mixture models for registration. In *European conference on computer vision* (pp. 733-750). Cham: Springer International Publishing.

Yuan, Y., Wu, Y., Fan, X., Gong, M., Ma, W., Miao, Q., 2023. EGST: Enhanced geometric structure transformer for point cloud registration. *IEEE transactions on visualization and computer graphics, 30*(9), 6222-6234.